AQUACULTURE GENOME TECHNOLOGIES

ZHANJIANG (JOHN) LIU

Blackwell Publishing

Aquaculture Genome Technologies

Aquaculture Genome Technologies

Zhanjiang (John) Liu
Auburn University

Zhanjiang (John) Liu is a Distinguished Alumni Professor, Department of Fisheries and Allied Aquacultures and Program of Cell and Molecular Biosciences, and Director, Aquatic Genomics Unit at Auburn University.

Blackwell Publishing Professional
2121 State Avenue, Ames, Iowa 50014, USA

Orders: 1-800-862-6657
Office: 1-515-292-0140
Fax: 1-515-292-3348
Web site: www.blackwellprofessional.com

Blackwell Publishing Ltd
9600 Garsington Road, Oxford OX4 2DQ, UK
Tel.: +44 (0)1865 776868

Blackwell Publishing Asia
550 Swanston Street, Carlton, Victoria 3053, Australia
Tel.: +61 (0)3 8359 1011

First edition, 2007

Library of Congress Cataloging-in-Publication Data
Liu, Zhanjiang.
Aquaculture genome technologies / Zhanjiang (John) Liu. — 1st ed.
p. cm.
ISBN-13: 978-0-8138-0203-9 (alk. paper)
ISBN-10: 0-8138-0203-2 (alk. paper)
1. Genomics. 2. Aquatic genetic resources. 3. Aquaculture.
I. Title.
QH447.L58 2007
572.8′6—dc22

2006038408

The last digit is the print number: 9 8 7 6 5 4 3 2 1

Contents

Foreword

The birth of livestock genomics 15 years ago was inspired by the human genome initiative and the potential for capturing both its technologies and massive comparative data sets for application to livestock species, most of which are mammals. We are currently reaping the benefits of these efforts, with sequencing projects completed or ongoing in chickens, cattle, pigs, and horses and valuable mapping resources developed for others such as sheep and turkeys. Traits of economic and physiological significance are being mapped, and underlying genes are being discovered. The biological diversity of species used in aquaculture, however, presents a unique set of problems to genomic studies, both in technology development and in the application of genomic information to food production. A book that captures the status of genomic technologies as applied to aquaculture species and the rapid state of advancement of genomics of some of the principal species is a welcome addition to the animal genomics literature.

Species used in aquaculture span both the vertebrate and invertebrate arms of the animal kingdom and incorporate a range of issues related to genome size, genome redundancy, and a variety of reproductive strategies. With the exception of the bony fishes that should benefit from the advanced genomics of zebra fish and puffer fish, aquaculture genomics will not have the direct benefit of extensive comparative genomic data sets provided by the human genome project to mammalian genetics and surprisingly, also to the chicken genome. Nonetheless, technologies for developing DNA markers, linkage and physical maps, and transcription profiling tools are universal and DNA sequencing is now being discussed in terms of a few thousand dollars per Gb in the not too distant future. Efforts to develop tools, make maps, and ultimately sequence genomes of aquaculture species will not only be rewarded by improved health and productivity of important food sources but in defining the biology underlying the genomic and physiological diversity that make these species daunting targets for genetic studies in the first place.

A surprising wealth of tools has already been generated for genome mapping and functional studies in many of the species used in aquaculture. With the potential for sequencing on the horizon, the future is bright for aquaculture genomics. As a mammalian geneticist who thoroughly enjoys a day of sport fishing or a seafood platter, I am delighted with the progress and prospects reported in this book.

James E. Womack, Distinguished Professor
Texas A&M University
College Station, TX

Preface

The completion of the Human Genome Project inspired the entire world and triggered the start of a genomics revolution. Accompanying this revolution was a complete change in the way science was conducted in the field of life sciences. Without exception, the waves produced by the genome revolution are now having a tremendous impact on aquaculture genomics and aquaculture genetics in general. As recently as 10 years ago, there were no large-scale aquaculture genome projects in the entire world! The first Aquaculture Genome Workshop held in Dartmouth, Massachusetts in the fall of 1997 could be regarded as the official start of aquaculture genomics. Today it is a reality that the entire genomes of several important aquaculture species are on the verge of being sequenced. This raises new challenges for aquaculture geneticists, breeders, and fisheries managers regarding how to best use the huge amount of genomic information now available, and how to master and apply continuously changing genome technologies to aquaculture and fisheries.

The purpose of this book is to provide a snapshot of genome technologies from the perspectives of aquaculture and fisheries scientists, and to provide a textbook suitable for students majoring in agricultural sciences. I feel that there are several compelling reasons for producing such a book. First, while it is easy to find genomics books these days, it is rare to find books providing enough background information of the basic principles and concepts underpinning genome technologies. My background was in agriculture, but I have spent most of my recent career on basic genome research. My own experience plus that gained through teaching a graduate course on agricultural genomics suggested that in order to effectively grasp the key issues of genomics, an understanding of genome technologies is essential. Such an understanding can be gained much more effectively if the basic principles behind these technologies are clearly explained, because many students may have not systematically taken courses in molecular biology, genetics, biochemistry, bioinformatics, etc. Second, most genomics books take a pure genomics approach using classical model species examples without consideration of potential applications of genome technologies in practical settings. There is a great gap to be bridged in the understanding of how basic genomics is to be used beyond the area of human health. This book provides a thorough overview of genome technologies and their applications in aquaculture and fisheries. Third, aquaculture and fisheries species have unique biological characteristics that demand modification or adaptation of existing genome technologies. Although no chapters of this book describe novel genome technologies that have originated from or are unique to aquaculture or fisheries species, almost every chapter deals with how genome technologies can be used for aquaculture and fisheries, or for agricultural sciences in general.

This book contains 29 chapters written by well-known scientists from all over the world. Their enriched experience in both genomics and aquaculture and fisheries allowed them to provide discussions of genome technologies with unique angles that will prove to be most helpful for academic professionals, research scientists, and graduate and college students in agriculture, as well as for students of aquaculture and fisheries. In spite of its focus on aquaculture and fisheries, this book should be suitable as well for students in animal sciences, poultry science, agronomy, horticulture,

entomology, and plant pathology. I completely share the sentiments of one contributor, Dr. Eric Hallerman from Virginia Tech, as he wrote in one of his e-mails to me: "This chapter ended up being more demanding, but more rewarding to produce than I had anticipated. I ended up learning a lot, which is in part why I agreed to do the work. (Yes, teaching students was the major motivator)." Teaching students more effectively was similarly my major motivation and passion through the long process of assembling this book.

This book is divided into five parts. In Part 1, Marking Genomes, concepts, principles, and applications of various DNA marker technologies are presented. In Part 2, Mapping Genomes, various genome-mapping techniques are presented including genetic linkage mapping, QTL mapping, physical mapping, radiation hybrid mapping, and comparative mapping. In addition, the principles and applications of marker-assisted selection are presented. Topics in Part 3, Analysis of Genome Expression and Function, include EST analysis, microarrays, environmental genomics, and functional genomics. Part 4 should have been entitled Sequencing the Aquaculture Genomes, but because no genomes of aquaculture species have been sequenced, it is entitled Preparing for Genome Sequencing. This part discusses existing sequencing technologies that brought us to where we are, and the emerging sequencing technologies that will lead us into the future. Nonetheless, strategies for sequencing the genomes of aquaculture species are also discussed in this part. In the last part, Part 5, Dealing with the Daunting Genomes of Aquaculture Species, the unique biology and characteristics of aquaculture genomes are illustrated through a few examples such as the duplicated fish genomes, complexities involved in functional studies of paralogous genes, the enormously high fecundity and segregation distortion of oysters, and extremely high polymorphism in oysters as well as other bivalve species. Not only are such unique features presented in relation to genome technologies, but potential solutions are also provided, supplying researchers with potential shortcuts to avoid having to struggle through these problems again.

I would like to thank all of the chapter contributors who are truly experts in aquaculture genomics. Their willingness to share their knowledge and expertise made this book possible. I am honored to have one of the most prestigious genome scientists in the world working in the area of livestock genomics, Dr. James Womack, a member of the National Academy of Sciences USA from Texas A&M University, to write the Foreword for this book. I am grateful to my students Eric Peatman, Peng Xu, Shaolin Wang, and Jason Abernathy, and my colleague Dr. Huseyin Kucuktas who helped in proofreading some of the chapters. I have had a year of pleasant experience interacting with Erica Judisch, Editorial Assistant for Blackwell Publishing Professional, and Justin Jeffryes, Commissioning Editor for Plant Science, Agriculture, and Aquaculture with Blackwell Publishing Professional. Finally, I must thank the two most important women in my life, my mother Youzhen Wang and my wife Dongya Gao; the former inspires me to succeed, while the latter makes sure I do succeed.

Zhanjiang (John) Liu

List of Contributors

Jonas S. Almeida
Department of Biostatistics and
Applied Mathematics
University of Texas
MD Anderson Cancer Center, Unit 447
1515 Holcombe Boulevard
Houston, TX 77030 USA

Louis Bernatchez
Réseau Aquaculture Québec (RAQ)
Pavillon C-H Marchand
Université Laval
Québec, QC
Canada G1K 7P4

Robert W. Chapman
Marine Resources Research Institute
South Carolina Department of
Natural Resources
Charleston, SC 29412 USA

Alan Christoffels
Computational Biology Group
Temasek Life Sciences Laboratory
1 Research Link
National University of Singapore
Singapore

Karl J. Clark
Department of Animal Sciences
University of Minnesota
St. Paul, MN 55108 USA

Avner Cnaani
Hubbard Center for Genome Studies
University of New Hampshire
Suite 400, Gregg Hall
35 Colovos Road
Durham, NH 03824 USA

Glenn A. Cooper
Centre for Biomedical Research
University of Victoria
Victoria, British Columbia
Canada V8W 3N5

Jason P. Curole
Department of Biological Sciences
University of Southern California
3616 Trousdale Parkway, AHF 107
Los Angeles, CA 90089-0371 USA

Roy G. Danzmann
Department of Integrative Biology
University of Guelph
Guelph, Ontario
Canada N1G 2W1

William S. Davidson
Department of Molecular Biology
and Biochemistry
Simon Fraser University
8888 University Drive
Burnaby, British Columbia
Canada V5A 1S6

Chunguang Du
Department of Biology and
Molecular Biology
Montclair State University
Montclair, NJ 07043 USA

Pierre Duchesne
Réseau Aquaculture Québec (RAQ)
Pavillon C-H Marchand
Université Laval
Québec, QC
Canada G1K 7P4

Karim Gharbi
Division of Environmental and
Evolutionary Biology Institute of
Biomedical and Life Sciences
University of Glasgow
Glasgow, Scotland UK G12 8QQ

Paul J. Grobler
Faculty of Natural and
Agricultural Sciences
University of the Free State
P.O. Box 339, Bloemfontein 9300
South Africa

Ximing Guo
Haskin Shellfish Research Laboratory
Institute of Marine and Coastal Studies
Rutgers University
6959 Miller Avenue
Port Norris, NJ 08349 USA

Perry B. Hackett
Department of Genetics, Cell Biology
and Development
Arnold and Mabel Beckman Center for
Transposon Research
6-106 Jackson Hall
University of Minnesota
Minneapolis, MN 55455 USA

Eric M. Hallerman
Department of Fisheries and
Wildlife Sciences
Virginia Polytechnic Institute and State
University Blacksburg, VA 20461-0321
USA

Limei He
Department of Soil and Crop Sciences
Texas A&M University
College Station, TX 77843 USA

Dennis Hedgecock
Department of Biological Sciences
University of Southern California
3616 Trousdale Pkwy, AHF 107
Los Angeles, CA 90089-0371 USA

Jess W. Jones
U.S. Fish and Wildlife Service
Blacksburg, VA 24061-0321 USA

Thomas D. Kocher
Hubbard Center for Genome Studies
University of New Hampshire
Suite 400, Gregg Hall
35 Colovos Road
Durham, NH 03824 USA

Ben F. Koop
Centre for Biomedical Research
University of Victoria
Victoria, British Columbia
Canada V8W 3N5

Abraham Korol
Institute for Evolution
Haifa University
Haifa, Israel 31905

Huseyin Kucuktas
Department of Fisheries and
Allied Aquacultures
Auburn University
Auburn, AL 36849 USA

Bo-Young Lee
Hubbard Center for Genome Studies
University of New Hampshire
Suite 400, Gregg Hall
35 Colovos Road
Durham, NH 03824 USA

Yaning Li
Department of Plant Pathology
Agricultural University of Hebei
Biological Control Center of Plant
Disease and Plant Pests of Hebei
Province
Baoding, China 071001

Lei Liu
W.M. Keck Center for Comparative
and Functional Genomics
University of Illinois at
Urbana-Champaign
330 Edward R. Madigan Laboratory
1201 W. Gregory Dr.
Urbana, IL 61801 USA

Zhanjiang Liu
Department of Fisheries and
Allied Aquacultures
Auburn University
Auburn, AL 36849 USA

Eric Peatman
Department of Fisheries and
Allied Aquacultures
Auburn University
Auburn, AL 36849 USA

Caird E. Rexroad III
USDA/ARS National Center for Cool
and Cold Water Aquaculture
11861 Leetown Road
Kearneysville, WV 25430 USA

Matthew L. Rise
The Ocean Sciences Centre
Memorial University of Newfoundland,
1 Marine Lab Road
St. John's, NL
Canada A1C 5S7

Max F. Rothschild
Department of Animal Science and the
Center for Integrated Animal Genomics
2255 Kildee Hall
Iowa State University
Ames, IA 50011 USA

Anatoly Ruvinsky
The Institute for Genetics and
Bioinformatics
University of New England
Armidale, Australia NSW 2351

Chantel Scheuring
Department of Soil and
Crop Sciences
Texas A&M University
College Station, TX 77843 USA

Andrey Shirak
Agricultural Research Organization
Institute of Animal Science
Bet Dagan, Israel 50250

Kristian R. von Schalburg
Centre for Biomedical Research
University of Victoria
Victoria, British Columbia
Canada V8W 3N5

Shaolin Wang
Department of Fisheries and
Allied Aquacultures
Auburn University
Auburn, AL 36849 USA

Yongping Wang
Haskin Shellfish Research Laboratory
Institute of Marine and Coastal Studies
Rutgers University
6959 Miller Avenue
Port Norris, NJ 08349 USA

Gregory W. Warr
Department of Biochemistry and
Molecular Biology
Marine Biomedicine and Environmental Sciences Center
Hollings Marine Laboratory
Medical University of South Carolina
Charleston, SC 29412 USA

Peng Xu
Department of Fisheries and
Allied Aquacultures
Auburn University
Auburn, AL 36849 USA

Zhe Xu
Haskin Shellfish Research Laboratory
Institute of Marine and
Coastal Studies
Rutgers University
6959 Miller Avenue
Port Norris, NJ 08349 USA

Hong-Bin Zhang
Department of Soil and Crop Sciences
Texas A&M University
College Station, TX 77843 USA

Aquaculture Genome Technologies

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 1
Concept of Genomes and Genomics

Zhanjiang Liu

When searching for the basic concept of genomics, one may find numerous definitions such as:

- The study of genes and their functions
- The study of the genome
- The molecular characterization of all the genes in a species
- The comprehensive study of the genetic information of a cell or organism
- The study of the structure and function of large numbers of genes simultaneously
- etc., etc.

In order to have a good concept of genomics, let us first explore the concept of genome, and its relationship to genome expression and genome functions.

The Concept of Genome and Genomics

The term genome is used to refer to the complete genetic material of an organism. Strictly speaking, the genetic material of an organism includes the nuclear and mitochondrial genomes for plants and animals, and also chloroplast genomes for plants. Since the mitochondrial and chloroplast genomes are small and contain only a limited number of genes, the focus of genome research is on the nuclear genome. Hence, I will limit this chapter largely to the nuclear genome.

Let us define genomics in its narrowest sense using the genetic central dogma (Figure 1.1) where in most cases, deoxyribonucleic acid (DNA) is transcribed into ribonucleic acid (RNA), and RNA is translated into proteins. Although genetic information is stored in DNA, it cannot be realized without being transcribed into the intermediate molecules RNA, which with a few exceptions, must be translated into proteins in order to have biological functions. Thus, the entire DNA content of an organism is called the genome; the entire RNA world of an organism is called its transcriptome, and the entire protein content of the organism is called its proteome. The science of studying the genome is called genomics; the science of studying the transcriptome is called transcriptomics; and the science of studying the proteome is called proteomics. In spite of such divisions, the term genomics often is used to cover not only this narrow sense of genomics, but also transcriptomics, and in some cases proteomics as well.

Genomics can be divided into structural genomics, which studies the structures, organization, and evolution of genomes, and functional genomics, which studies expression and functions of the genomes. Since genome functions are reflected in the transcripts and proteins that the transcripts encode, genomics must also study the transcriptome and the proteome.

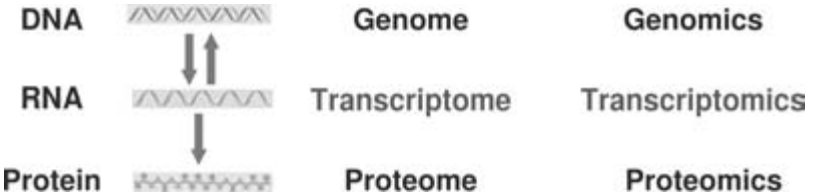


Figure 1.1. The concept of genome and genomics in relation to the genetic central dogma. The entire DNA content of an organism (the genome) is transcribed into RNA (the entire RNA content of the organism is called the transcriptome), and the RNA is translated into proteins (the proteome). Genomics, transcriptomics, and proteomics are sciences that study the genome, transcriptome, and proteome, respectively.

It must be pointed out that while the genome is relatively stable in an organism in most cell types with the exception of gene rearrangements in immune-related cell types, the transcriptome is highly dynamic. The types of transcripts and their relative levels of expression are highly regulated by tissue specificity, developmental stage, physiological state, and the environment. For instance, if an organism has 25,000 genes, not all genes are expressed in every type of cell. Those genes required for the basic cell structure and functions are probably expressed in all tissues, organs, and cell types; whereas each cell type expresses a subset of the genes specific for that cell type. Many genes are expressed throughout development, but certain genes are expressed only at a specific developmental stage. Physiological state can affect gene expression in a fundamental and dramatic way. For instance, gonadotropin genes are expressed only in the pituitary and gonad, and expressed highly during spawning seasons of the reproductive cycle in fish. The environment can insert its effect on gene expression in multiple dimensions. Temperature, pH, water quality, stress, dissolved oxygen, and many other environmental factors can induce or suppress expression of a large number of genes.

In addition to the dynamic nature of the transcriptome, variation of the transcriptome can also be brought about by production of alternative transcripts by the same set of genes. It is now widely believed that the complexity of the transcriptome is much larger than the genome because of alternative transcripts. The largest proportion of alternative transcripts is produced by alternative splicing where a single gene is transcribed into heterogeneous nuclear mRNA (hnRNA); through splicing, more than one mRNA molecule is produced, leading to the phenomenon that introns of one transcript may be exons of another. The second mechanism for the generation of alternative transcripts is through the use of alternative promoters. In a single gene, more than one promoter can be functional leading to the generation of different, but related transcripts. In addition, use of differential polyadenylation sites can also lead to the generation of alternative transcripts. Therefore, it is widely believed that the information stored in the genome is amplified and diversified at the transcriptome level. The genetic information is further amplified and diversified at the protein level. Though each transcript may only encode one protein, the primary protein may be differentially processed to produce more than one active polypeptide; posttranslational glycosylation, acetylation, phosphorylation, and other modifications can result in a much larger complexity leading to drastically different biological functions. Even highly related gene products may encode proteins leading to absolutely opposite biological functions. For instance, an interleukin-1 Type II receptor is a decoy target for

IL-1, whose binding to interleukin 1 intercepts the function of interleukin 1. Therefore, the genetic central dogma is correct in terms of the basic flow of genetic information, and the capacities of the primary functions of transcription and translation, while much larger complexities result from amplification and diversification of the same set of genetic material, lead to the generation of biologically different molecules. Such differences in biological molecules, when considered for the various combinations of many genes, can result in numerous biological outcomes.

As a new branch of science, genomics has its own defined scope of study, its own box of tool kits, and its own unique set of approaches. It is different from traditional molecular genetics which looks at single genes, one or a few genes at a time. Genomics is trying to look at all of the genes as a dynamic system, over time, to determine how they interact and influence biological pathways, networks, physiology, and systems in a global sense. Genome technologies, the focus of this book, have been developed to cope with the global scope of tens of thousands of genes as a snapshot. Much like dealing with a globe, landmarks (or as we have called them molecular markers) are needed to mark the position within the huge genome. Genetic and physical maps have been developed to understand the structure and organization of genomes, and to understand genomic environs and genome evolution in relation to genome expression and function. Specific approaches have been developed to cope with the large number of genes, regardless if it is for gene discovery, cloning and characterization, or for analysis of gene expression. Thus large-scale analysis of expressed sequence tags using highly normalized complementary DNA (cDNA) libraries allows rapid gene discovery and cloning in the scale of tens of thousand of genes. Such operations have also been supported by other genome technologies such as powerful automated sequencing to allow gene discovery and identification in a streamlined industrial fashion. Expression of genes is determined in an entire genome scale, or sometimes referred to as genome expression, to relate complex regulation of genes to their functions in terms of systems biology. Expression of tens of thousands of genes can be monitored simultaneously and continuously, allowing their interactions and networking to be detected. Signal transduction is no longer "behind the scene" molecular events, but can be observed with clustering of co-regulated gene expression under specific development, physiology, or environmental conditions. Genes and their functions are studied much in terms of their sociology, networking, and interactions, rather than looking at one or a few genes at a time, as conducted by traditional molecular biology. Such operations demand the development of very powerful gene expression analysis such as microarray technologies. Such technological advances allow the generation of tremendously large data sets that have been beyond the comprehension capacities of biologists. Assistance is needed from all areas of biology, and more so from disciplines outside biology that can handle large amounts of information. Computer sciences and mathematics are among the first disciplines genomics has demanded cooperation from. While handling large data sets from the genome, genome expression, and genome function, much confusion has emerged regarding whether the observed phenomenon is real or if it is just a fluctuation of the systems biology. As such, statisticians are also called upon to join computer scientists, mathematicians, and the biologists. Because these scientists speak different languages (e.g., English for one group, French for the second, Chinese for the next, and so on), understanding all of the languages and being able to function among these different disciplines is becoming the goal of a large group of scientists who define themselves as bioinformaticians working in the new area of bioinformatics. It is clear that genomics cannot be a science without

bioinformatics. Clearly, the definition of genomics is becoming more complex with this discussion. Now, you can certainly come up with your own definitions.

The excitement and success of genomics has brought the emergence of numerous '–omics' sciences (http://genomicglossaries.com/content/genomics_glossary.asp). Subbranches of genomics are emerging in large numbers. The following list includes some of those subbranches:

- agricultural genomics
- applied genomics
- behavior genomics
- biochemical genomics
- chemogenomics
- clinical genomics
- combinatorial genomics
- comparative genomics
- computational genomics
- deductive genomics
- ecotoxicogenomics
- environmental genomics
- evolutionary genomics
- forward genomics
- functional genomics
- immunogenomics
- industrial genomics
- intergenomics
- inverse genomics
- lateral genomics
- nanogenomics
- network genomics
- oncogenomics
- pharmacogenomics
- phylogenomics
- physiological genomics
- population genomics
- predictive genomics
- reverse genomics
- structural genomics
- toxicogenomics
- translational genomics
- and so on

Cells, Nucleus, Chromosomes, Genomes, and Genomic DNA

Genomes can exist in various forms. A genome can be either RNA or DNA, single-stranded or double-stranded. For example, the human immunodeficiency virus (HIV) is a retrovirus whose genome contains a single-stranded RNA molecule. However, such unusual genomes are mostly found within viruses and bacteriophages. In prokaryotes such as bacteria, by definition they do not have a nucleus; the genomes are made up with double-stranded DNA in either circular or linear forms. For instance, the *Escherichia coli* genome is made of a single circular DNA molecule, whereas the genome of *Borrelia burgdorferi* is composed of a linear chromosome approximately one megabase (million base) in size. Eukaryotic genomes contain two or more linear molecules of double-stranded DNA in the form of chromosomes.

Within each eukaryotic cell, there is a nucleus in which chromosomes are located. Individual species harbor a fixed number of chromosome pairs (2n) with fixed shapes, sizes, and centromere location. These chromosome morphologies are commonly known as the karyotypes. All somatic cells in a diploid organism harbor identical chromosome pairs that are randomly shared into a single chromosome set during meiosis to produce eggs and sperms. Upon fertilization of an egg (n) by a sperm (n), the embryo recovers the diploid state with two sets of chromosomes.

Chromosomes are threadlike structures containing genes and other DNA in the nucleus of a cell. Different kinds of organisms have different numbers of chromosomes.

Humans have 46 chromosomes—44 autosomes and 2 sex chromosomes. Each parent contributes one chromosome to each pair, so children get half of their chromosomes from their mothers and half from their fathers. This is important in sexual reproduction where the gametes (i.e., sperms and eggs) are haploid cells, and upon fertilization of an egg by a sperm, the embryo recovers the diploid state. The number of chromosomes is usually constant for each organism, but may vary greatly from species to species. For instance, the fruit fly *Drosophila melanogaster* has four chromosomes whereas *Ophioglossum reticulatum*, a species of fern, has the largest number of chromosomes with more than 1,260 (630 pairs). The minimum number of chromosomes found in a species occurs in a species of ant, *Myrmecia pilosula*, in which females have one pair of chromosomes and males have just a single chromosome. This species reproduces through a process called haplodiploidy, in which fertilized eggs (diploid) become females, while unfertilized eggs (haploid) develop into males.

Each chromosome is a portion of the genome and all the chromosomes compose the entire genome. Although all chromosomes maintain their own integrity, they each can be viewed as a segment of the genome. The total length of genomic DNA thus is equal to the sum of all chromosomal DNA. In their natural existence, the physical pieces of DNA in each cell are equal to the number of chromosomes. It must be emphasized that such entire chromosomal DNA is essentially impossible to obtain for routine molecular analysis. Chromosomal DNA is randomly broken during genomic DNA extraction even under the most sophisticated preparation by the most skilled researchers. Most often, millions of cells are used in a single DNA extraction. Therefore, genomic DNA used in molecular analysis represents multiple copies of the genome with multiple overlapping segments, simply because the breakage points are random and different in each cell genome.

Genome Sizes

Genome sizes of organisms vary greatly, spanning a range of almost 100,000 fold. The bacterial genomes are commonly at the range of a million base pairs (Mbp), while the largest animal genome reported to date is 133 picograms (pg) (or about 1.3×10^{11} base pairs; 1 pg DNA = 1×10^{-12} g = 978 Mbps) for a species of lungfish, *Protopterus aethiopicus*, which is some 40 times larger than the human genome, followed by a number of amphibians, *Necturus lewisi* and *N. punctatus* at 120 pg, *Necturus maculosus* and *Amphiuma means* and the lungfish *Lepidosiren paradoxa*, all at roughly 80 pg. In general, the genome size is correlated with biological complexities, but many exceptions exist. For instance, some plant species and amphibians can have very large genomes, dozens of times larger than the human genome.

The largest teleost genome size is 4.4 pg in the masked *Corydoras metae*, and the smallest teleost genome size is approximately 0.4 pg in several puffer fish of the family *Tetraodontidae*. Fish as a whole have the largest ranges for genome sizes.

Crustaceans also have a wide range of genome sizes from 0.16 pg to 38 pg with an average of 3.15 pg. The smallest crustacean genome size (0.16 pg) is in a water flea, *Scapholeberis kingii*, and the largest crustacean genome size (38 pg) is in *Hymenodora* sp., a deep-sea shrimp. The most important crustacean species for aquaculture involves several major species of the shrimps. Their genome sizes are approximately 2.5 pg.

The molluscan genome sizes are more uniform ranging from the smallest molluscan genome size of 0.4 pg in the owl limpet *Lottia gigantean*, to the largest molluscan genome size of 5.9 pg in the Antarctic whelk *Neobuccinum eatoni*. Many aquacultured shellfish belong to the molluscans. The most important of these species in aquaculture include the oysters, such as the Pacific oyster (genome size 0.91 pg), the eastern oyster (genome size 0.69 pg), and the scallops (genome size between 0.95 to 2.1 pg).

The size of the genome of an organism is a constant. However, the ploidy of organisms varies. For instance, channel catfish are believed to be a diploid organism, whereas most salmonid fish used in aquaculture are believed to be tetraploid. In cultivated wheat plants, various ploidies exist including diploid, tetraploid, and hexaploid. In order to standardize the genome size so that they can be compared, genome sizes are presented in C-values, which is the haploid genome size in picograms.

Several excellent databases exist for genome sizes. The Animal Genome Database (http://genomesize.com/) is a comprehensive catalogue of animal genome size data. It includes haploid genome sizes for more than 4,000 species including approximately 2,750 vertebrates and 1,315 invertebrates compiled from 5,400 records from more than 425 published sources (Gregory 2005; Animal Genome Size Database, http://www.genomesize.com/). The Database Of Genome Sizes (DOGS) (http://www.cbs.dtu.dk/databases/DOGS/) is also a very useful database that includes a number of links to genome size and genome research resources such as the following:

- the Plant DNA C-Value Database (http://www.rbgkew.org.uk/cval/)
- the Genome Atlases for Sequenced Genomes (http://www.cbs.dtu.dk/services/GenomeAtlas/)
- the DBA mammalian genome size database (http://www.unipv.it/webbio/dbagsdb.htm)
- several other useful databases and resources.

Knowledge of genome size is not only important for genome studies in relation to genome structure, organization, and evolution, but also for a number of practical reasons such as genome mapping, physical mapping, and genome sequencing. As listed in Table 1.1, the primary methods for the determination of genome sizes are Feulgen densitometry (Hardie et al. 2002), flow cytometry, and Feulgen image analysis densitometry (Lamatsch et al. 2000). These three methods account for over 81% of all methods used for the estimation of genome sizes (Table 1.1). Readers with an interest in methodologies for the determination of genome size are referred to the literature list of the Animal Genome Size Database (http://www.genomesize.com/).

Number of Genes

The number of genes in a given organism is fixed, but discovering it is a daunting task. For the best characterized human genome, the number of genes now is believed to be approximately 25,000. In the 1980s, the number of human genes was believed to be 100,000 to 125,000. In the early 1990s, the human genome was believed to include 80,000 genes. Although the final completion of the Human Genome Project was celebrated in April 2003 and sequencing of the human chromosomes is essentially "finished," the exact number of genes encoded by the genome is still unknown. In

Table 1.1. Methods and their frequencies used for the determination of genome sizes. The table was adapted from the Animal Genome Size Database (http://www.genomesize.com/).

Methods	Abbreviation	Number of genomes	Percentage of methods used for genome size determination
Feulgen Densitometry	(FD)	2,480	45.93%
Flow Cytometry	(FCM)	1,075	19.91%
Feulgen Image Analysis Densitometry	(FIA)	839	15.54%
Bulk Fluorometric Assay	(BFA)	471	8.72%
Static Cell Fluorometry	(SCF)	303	5.61%
Biochemical Analysis	(BCA)	142	2.63%
Not Specified	(NS)	63	1.17%
Ultraviolet Microscopy	(UVM)	13	0.24%
Gallocyanin Chrom Alum Densitometry	(GCD)	11	0.20%
Complete Genome Sequencing	(CS)	2	0.04%
Methyl Green Densitometry	(MGD)	1	0.02%

2000 when the human genome project was originally declared as being completed, the human genome was believed to contain 35,000 to 40,000 genes. Now in 2006, the total number of human genes is believed to be around 25,000. Clearly, many of the "gene-like" reading frames were proved not to be genes.

It could still take years before a truly reliable gene count can be assessed. The uncertainty is derived from different methods used for the assessment of genes. Some prediction programs detect genes by looking for distinct patterns that define where a gene begins and ends. Other programs look for genes by comparing segments of sequence with those of known genes and proteins. The first tends to overestimate, while the second tends to underestimate, the gene count. No matter which programs are used, the bottom line is that evidence to support a gene model has to come from expression information. In spite of some 7 million Expressed Sequence Tags (EST) obtained from humans, they cannot support all of the gene models yet because many gene products have not been found. Although the ballpark range of the number of human genes should not change dramatically, finer tuning for the total number of genes is still expected.

The number of genes an organism has is correlated with the biological complexity of the organism. With this belief, the number of human genes came as a shock to many scientists because even the *E. coli* has 4,377 genes with 4,290 protein encoding genes. Saying that we are only six times more complex than a bacteria is truly a humiliation to many, but it is probably worse to say that the human gene count is only one-third greater than that of the simple roundworm *C. elegans* which has about 20,000 genes (Claverie 2001). Nonetheless, the unique number of gene products (proteins) is likely correlated with biological complexities, though the absolute number of genes may vary depending on the level of gene duplications. With such assumptions, it is reasonable to believe that many fish genomes will have a similar number of unique

genes as the human genome, but their total number of genes could even be slightly larger, considering high levels of gene duplications in teleosts.

A basic understanding of the genome, genome size, the number of chromosomes, and the number of genes is important before the start of a genome project. Not only the efforts required to characterize the genome are affected by the genome size and complexity, but also proper methodologies should be taken according to the circumstances as well.

References

Claverie JM. 2001. Gene number. What if there are only 30,000 human genes? Science, 291, pp. 1255–1257.

Gregory TR. 2005. Genome size evolution in animals. In: *The Evolution of the Genome*, edited by TR Gregory. Elsevier, San Diego, CA, pp. 3–87.

Hardie DC, R Gregory, and PDN Hebert. 2002. From pixels to picograms: a beginners' guide to genome quantification by feulgen image analysis densitometry. J Histochem Cytochem, 50, pp. 735–749.

Lamatsch DK, C Steinlein, M Schmid, and M Schartl. 2000. Noninvasive determination of genome size and ploidy level in fishes by flow cytometry: detection of triploid *Poecilia formosa*. Cytometry, 39, pp. 91–95.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Part 1
Marking Genomes

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 2
Restriction Fragment Length Polymorphism (RFLP)

Zhanjiang Liu

Restriction fragment length polymorphism (RFLP) markers (Botstein et al. 1980) were regarded as the first shot in the genome revolution (Dodgson et al. 1997), marking the start of an entirely different era in the biological sciences. RFLP was the most popular approach for analysis of genetic variation during the entire 1980s. As indicated by its name, RFLP is based on deoxyribonucleic acid (DNA) fragment length differences after digesting genomic DNA with one or more restriction enzymes. Most typically, genomic DNA is digested by one or more restriction enzymes and separated on an agarose gel. To adapt to further handling, the DNA in the gel must be transferred to a solid support such as nitrocellulose or nylon membranes. The specific DNA locus with a potential fragment length difference is characterized by hybridization to a probe, a radioactively labeled DNA or ribonucleic acid (RNA) molecule with sequence similarities to the locus of interest. After hybridization, the nonspecific probes must be washed away leaving only hybridized probes to the specific locus. The membrane is then exposed to a piece of X-ray film for autoradiography to visualize the DNA bands. In spite of its popularity, RFLP is able to detect only large shifts in DNA fragment sizes. Therefore, it can detect only insertions and deletions of large sizes, and the gain or loss of restriction sites. It is unable to detect the vast majority of point mutations and deletions or insertions involving small-sized segments because of its low resolution using agarose gel electrophoresis. As a result, polymorphic rates are low at most loci. The efforts involved in RFLP marker development have been enormous. RFLP attempts to detect genetic variation one locus at a time. The low polymorphic rates, when coupled with expensive and laborious processes, have made application of RFLP limited. It should be particularly noted that RFLP requires previous genetic information, such as the availability of probes or sequence information; information often not available for many fish or other aquaculture species.

In this chapter, technology advances leading to the development of RFLP, the principles and molecular basis of RFLP, inheritance of RFLP, power of RFLP, strengths and weaknesses of RFLP, and applications of RFLP for aquaculture genomics research will be summerized.

Technology Advances Leading to the Development of RFLP

Two specific technological advances—the discovery and application of restriction enzymes and the development of DNA hybridization—set the foundation for RFLP. To comprehend and appreciate the principles of RFLP, it is necessary to have a good

understanding of restriction enzymes and their applications, as well as that of DNA hybridizations.

Restriction enzymes are also called restriction endonucleases, proteins produced by bacteria that cleave DNA at specific sites along the molecule. In the bacterial cell, restriction enzymes cleave foreign DNA, thus eliminating infecting organisms. Restriction enzymes can be isolated from bacterial cells and used in the laboratory to manipulate fragments of DNA. They are indispensable tools of recombinant DNA technology or genetic engineering, as well as genomics.

It is generally believed that the biological function of restriction enzymes is to protect cells from foreign DNA. A bacterium uses a restriction enzyme to defend against bacterial viruses called bacteriophages, or phages. When a phage infects a bacterium, it inserts its DNA into the bacterial cell so that it might be replicated. The restriction enzyme prevents replication of the phage DNA by cutting it into many pieces. Restriction enzymes were named for their ability to restrict, or limit, the number of strains of bacteriophage that can infect a bacterium. An obvious question that often arises is why the restriction enzymes do not digest bacterial DNA. The answer is that the bacteria also harbor a set of defense weaponry containing so-called restriction enzyme modification systems. Usually, organisms that make restriction enzymes also make a companion modification enzyme (DNA methyltransferase) that protects their own DNA from cleavage. These enzymes recognize the same DNA sequence as the restriction enzyme they accompany, but instead of cleaving the sequence, they disguise it by methylating one of the bases in each DNA strand.

To date, more than 10,000 bacteria species have been screened for the existence of restriction enzymes; more than 2,500 restriction enzymes have been found with more than 250 distinct specificities. Occasionally enzymes with novel DNA sequence specificities are still found, but most now prove to be duplicates (isoschizomers) of already discovered specificities.

There are three classes of restriction enzymes, designated Type I, II, and III (Table 2.1). Type I and III enzymes are similar in that both restriction and methylase activities are carried out by one large enzyme complex, in contrast to the Type II system, in which the restriction enzyme is independent of its methylase. Type II restriction enzymes also differ from the other types in that they cleave DNA at specific sites within the recognition site; the others cleave DNA randomly, sometimes hundreds of bases from the recognition sequence. Type II restriction enzymes are endonucleases that cut DNA at specific sites, and are most useful for molecular biology research.

Table 2.1. Classification of restriction enzymes and their characteristics.

Type I	Type II (93%)	Type III
Restriction-methylase on the same subunit	Homo-dimers, methylase on a separate subunit	Restriction-methylase on the same subunit
ATP-dependent	Mg^{++} dependent	ATP-dependent
Binds to DNA recognition site and cuts DNA randomly—any DNA as long as it comes in contact	Recognize symmetric DNA sequences and cleave within the sequences	Cut the DNA at the recognition site and then dissociate from the DNA

Each restriction enzyme recognizes a short, specific sequence of nucleotide bases. These regions are called recognition sequences and are randomly distributed throughout the DNA. Different bacterial species make restriction enzymes that recognize different nucleotide sequences. Generally speaking, Type II restriction enzymes recognition sites are palindromes. A palindrome read from both sides yields the same sequence of characters (e.g., 121, IFFI, ABA). However, for a DNA sequence, a palindrome refers to reading the sequence from both strands from 5′-3′. For instance, the *Eco*R1 site is 5′-GAATTC-3′; and its complementary strand should also read 5′-GAATTC-3′. Thus, most 4–8 base pair palindromes are likely restriction sites. There are numerous commercial suppliers of restriction enzymes, such as New England Biolabs, Amersham Pharmacia Biotech, Qiagen, Promega, Invitrogen, and Stratagene, to name a few.

Restriction enzymes are named by using the first letter of the genus name and the first two letters of the species name from which they were isolated. Often, additional letters are used to designate the strains from which they were derived, or the chronological order in which the enzyme was isolated from the species. For example, the enzyme *Eco*RI is produced by *Escherichia coli* strain RY13; *Pst* I was isolated from *Providencia stuartii*; *Hind* III was isolated from *Haemophilus influenza*, and *Not* I was isolated from *Norcardia otitidis-caviarum*.

The odds or frequency of restriction enzymes digesting DNA depends on their recognition sequences. The shorter the recognition sequences, the higher the cutting frequency. Restriction enzymes have recognition sequences of 4, 6, or 8 base pairs. Examples of 4-base pair (bp) cutters are *Taq* 1, *Hpa* II, *Msp* I; examples of 6-bp cutters are *Eco*R1, *Hin*d III, *Bam* H1, *Pst* I, *Sal* I; and examples of 8-bp cutters are *Not* I and *Sfi* I. To date, many 4-bp cutters and 6-bp cutters are available, but the number of 8-bp cutters is limited. In addition to these 4-, 6-, and 8-bp cutters, some restriction enzymes have interrupted or ambiguous recognition sequences. For instance, *Acc* I has a recognition sequence of GT(at/gc)AC; *Bgl* I has a recognition sequence of GCCNNNNNGGC; and *Afl* III has a recognition sequence of ACPuPyGT. Restriction enzymes with 4-bp recognition sequences digest DNA at a frequency of one per 4^4 = 256 bp; restriction enzymes with 6-bp recognition sequences digest DNA at a frequency of one per 4^6 = 4,096 bp; restriction enzymes with 8-bp recognition sequences digest DNA at a frequency of one per 4^8 = 65,536 bp. When genomic DNA is digested with 4-, 6-, or 8-bp cutters, a smear should result except that the average size of the 8-bp cutter is the largest centered at approximately 65 kb; the average of the 4-bp cutter is the smallest centered at approximately 256 bp.

Three types of ends can be produced by Type II restriction enzymes including 3′-overhang (protruding), 5′-overhang, and blunt-ended molecules. These are important for the selection of restriction enzymes for cloning, filling-in labeling, or other operations. Proper planning should be made for the most efficient use of restriction endonucleases. In addition, some restriction enzymes do not digest DNA efficiently when the recognition sites are located close to the end of DNA. This is particularly important when incorporating restriction sites into PCR primers for cloning. For more information concerning this, readers are referred to an excellent list from New England Biolabs (http://www.neb.com/nebecomm/default.asp). With more than 250 commercially available and more than 2,000 total, considerations have to be made based on cutting frequency, what types of end they produce, ease of use, and economic considerations. Sources with patent rights and cloned products can be much cheaper than other sources.

In the early 1970s, the discovery of restriction enzymes offered biologists a great tool to cleave huge DNA into smaller pieces for analysis. At the same time, another line of technological advances, the establishment of principles of molecular hybridization using molecular probes, set the foundation for RFLP. The revolution brought about by molecular biology depended heavily on nucleic acid hybridization procedures. These techniques are used extensively in the research laboratory for detecting specific nucleotide sequences in DNA and RNA and are increasingly being applied in medicine for diagnosing diseases. All of the hybridization techniques started with a simple hybridization technique called Southern blot (Southern 1975). A Southern blot is a method in molecular biology of enhancing the result of an agarose gel electrophoresis by marking specific DNA sequences. The method is named after its inventor, the British biologist Edwin Southern. This caused other blot methods to be named as plays on Southern's name (for example, western blot, northern blot, southwestern blot, etc.). All of these blotting techniques require the use of molecular probes.

A probe refers to the agent that is used to detect the presence of a molecule in the sample. For Southern blot, the probe is a DNA sequence that is used to detect the presence of a complementary sequence by hybridization with a DNA sample. Probes are needed to screen for a gene of interest, to determine genomic structure and gene copy numbers, to analyze gene expression, or to validate allelic amplification in PCR.

Probes can consist of DNA, RNA, or antibodies. For DNA, the probes can be double-stranded or single-stranded. The probes can be continuously labeled to make very hot probes or can be end-labeled to trace the segments. Two methods are most frequently used to make continuously labeled probes: (1) Nick translation and (2) Random primer labeling (Sambrook et al. 1989). In the nick translation procedure, double-stranded DNA is nicked with a limited concentration of DNase I. The nicked ds-DNA is a perfect substrate for DNA polymerase I. DNA polymerase I has two major activities: 5′ to 3′ exonuclease activity and 5′ to 3′ polymerase activity. DNA polymerase I makes the new strand DNA with labeled dNTP while degrading the old strand of the DNA. In the random primer labeling procedure, DNA templates are heat-denatured and annealed to short random primers (hexomers), creating a perfect template for Klenow polymerase that makes the new strand with labeled dNTP. DNA synthesis continues until it reaches the next primer.

End-labeled probes can be made by labeling at the 3′ by filling-in reactions using a polymerase or by labeling at the 5′ by using polynucleotide kinase (Sambrook et al. 1989). Probes can be labeled in various other ways. For additional reading, readers are referred to Sambrook and others (1989), or *Current Protocols in Molecular Biology* edited by Fred M. Ausubel, Roger Brent, Robert E. Kingston, David D. Moore, J.G. Seidman, John A. Smith, and Kevin Struhl (2003).

Principles and Molecular Basis of RFLP

The molecular basis of RFLP is summarized in Figure 2.1.

Restriction endonucleases cut DNA wherever their recognition sequences are encountered so that changes in the DNA sequence due to indels, base substitutions, or rearrangements involving the restriction sites can result in the gain, loss, or relocation of a restriction site (Figure 2.1). Digestion of DNA with restriction enzymes

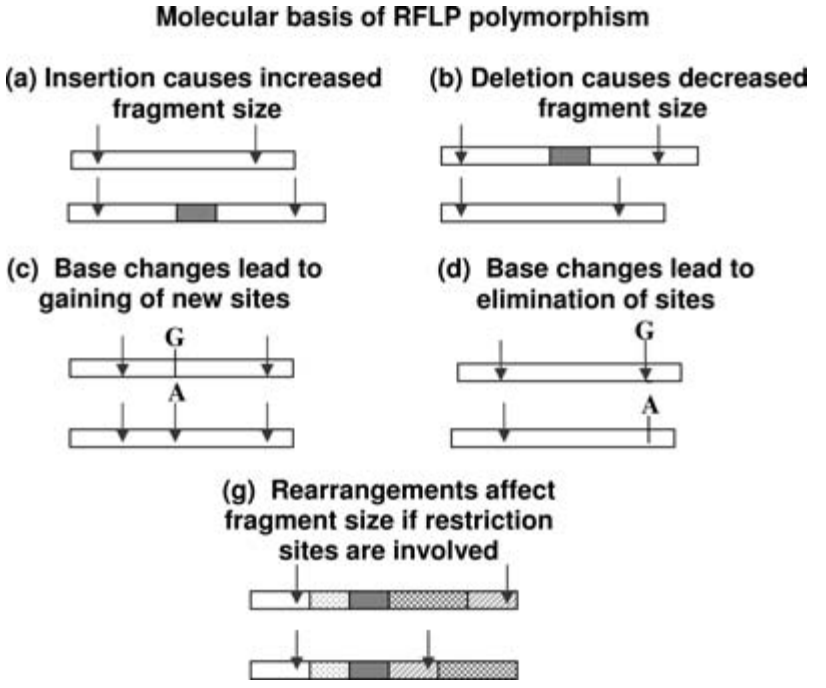


Figure 2.1. Molecular basis of RFLP polymorphism.

results in fragments whose number and size can vary among individuals, populations, and species. As RFLP analysis usually uses agarose gels, only large-size variations can be resolved. In terms of molecular basis, deletion and insertion between restriction sites within the locus of interest generates RFLP; base substitutions at restriction sites within the locus of interest leads to the loss of restriction sites and thus generating larger restriction fragments. Alternatively, base substitutions may lead to the generation of new restriction sites. For instance, the first base of AAATTC (not a restriction site) can mutate to G leading to GAATTC (now a site for *Eco*R1). In cases of rearrangements, the rearranged segments must involve the restriction enzyme sites under consideration to generate RFLP (Figure 2.1).

Two approaches are widely used for RFLP analysis. The first involves the use of hybridization, and the second involves the use of PCR. Traditionally, fragments were separated using Southern blot analysis (Southern 1975), in which genomic DNA is digested, subjected to electrophoresis through an agarose gel, transferred to a membrane, and visualized by hybridization to specific probes. Most recent analysis replaces the tedious Southern blot analysis with techniques based on the polymerase chain reaction (PCR). If flanking sequences are known for a locus, the segment containing the RFLP region is amplified via PCR. If the length polymorphism is caused by a relatively large (>approximately 100 bps depending on the size of the undigested PCR product) deletion or insertion, gel electrophoresis of the PCR products should reveal the size difference. However, if the length polymorphism is caused by base substitution at a restriction site, PCR products must be digested with a restriction enzyme to reveal the RFLP. With the increasing number of 'universal' primers available in the literature,

a researcher can target DNA regions that are either relatively conserved or rapidly evolving, depending on the amount of variation observed and the taxonomic level under examination. Also, PCR products can be digested with restriction enzymes and visualized by simple staining with ethidium bromide due to the increased amount of DNA produced by the PCR method. If the size shift is small, polyacrylamide gels or sequencing gels should be considered rather than agarose gels.

Inheritance of RFLP Markers

RFLP markers are inherited in a Mendelian fashion as codominant markers (Figure 2.2). Both alleles are expressed in molecular phenotypes (here, bands on gels). In the case of an individual heterozygous for two allelic RFLP patterns on alternative chromosomes, the phenotype includes both of the patterns (Figure 2.2). The codominance mode of inheritance is a strength of RFLP markers. In the mapping population, polymorphic RFLP bands segregate in a Mendelian fashion (Figure 2.3).

Differentiating Power of RFLP and Its Strengths and Weaknesses

The potential power of RFLP markers in revealing genetic variation is relatively low compared to more recently developed markers and techniques such as amplified fragment length polymorphism (AFLP) or microsatellites. Indels and rearrangements of regions containing restriction sites are perhaps widespread in the genomes of most species, but the chances of such an event happening within any given locus under study should be rare. Similarly, in a given genome of 10^9 base pairs, approximately 250,000 restriction sites should exist for any restriction enzyme with a 6-bp recognition sequence (that accounts for 1.5×10^6 bp or 0.15% of the entire genome). Base

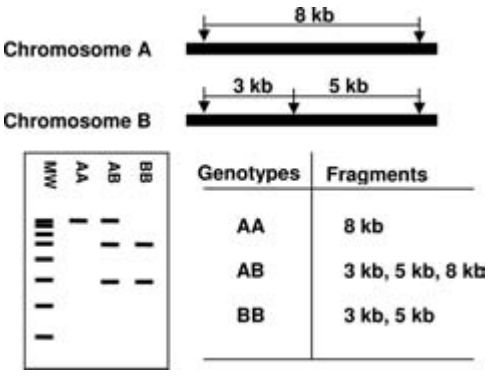


Figure 2.2. Codominant mode of inheritance of RFLP markers. In the example, a base substitution within the 8 kb fragment leads to the gaining of a new restriction site. For homozygous AA, one band of 8 kb should be generated; for homozygous BB, two bands of 3 kb and 5 kb should be generated; for heterozygous AB, three bands of 8 kb (from allele A), 3 kb and 5 kb (both from allele B) should be generated.

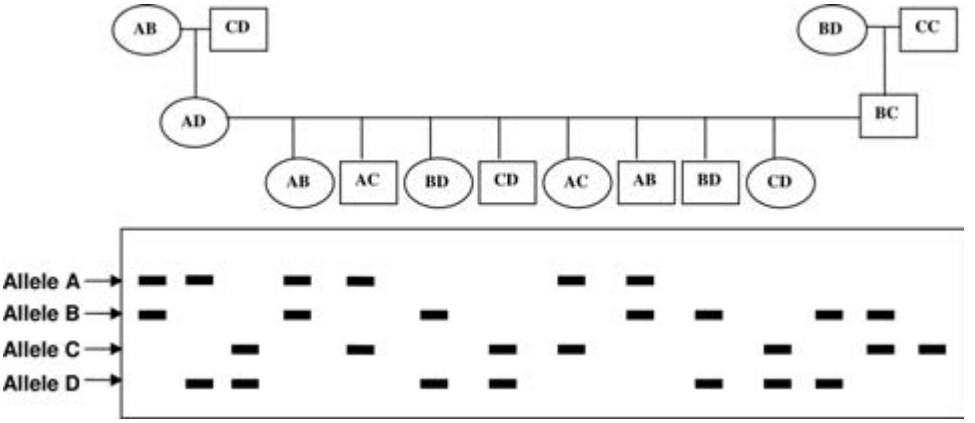


Figure 2.3. Segregation of RFLP markers highlighting codominant mode of inheritance. The first pair of grandparents is both heterozygous (AB and CD) and the second pair of grandparents is homozygous (BB and CC). When the daughter of the first grandparents (AD) mates the son from the second grandparents (BC), four types of segregation are possible: AB, AC, BD, and CD.

substitutions within these restriction sites must be widespread as well, but again, the chances that such base substitutions occur within the locus under study would be relatively small.

The major strength of RFLP markers is that they are codominant markers (i.e., both alleles in an individual are observed in the analysis). Because the size difference is often large, scoring is relatively easy. The major disadvantage of RFLP is the relatively low level of polymorphism. In addition, either sequence information (for PCR analysis) or probes (for Southern blot analysis) are required, making it difficult and time-consuming to develop markers in species lacking known molecular information.

Applications of RFLP in Aquaculture Genomics

RFLP markers are one of the most popular markers used in genetic studies. A search of the PUBMED database using RFLP as a key word led to the generation of 30,000 citations in early 2006. However, much of the popularity of RFLP markers was during earlier decades. Its popularity is reduced now due to availability of other more efficient marker systems. In spite of the popularity of RFLP markers, even in the earlier decades, its application in aquaculture genetics research was limited. (For a recent review, see Liu and Cordes 2004.) In most cases, RFLP markers have been used to differentiate species (Chow et al. 2006, Klinbunga et al. 2005), strains, or populations (Docker et al. 2003, Ohara et al. 2003, Zhang et al. 2005, Aranishi 2005, Sellos et al. 2003, Papakostas et al. 2006, Lehoczky et al. 2005, Apte et al. 2003). Of these studies using RFLP markers, many of them involved the use of mitochondrial DNA or the 16S rDNA (see Chapter 7 as well as Mamuris et al. 2001, Klinbunga et al. 2001, Lopez-Pinon et al. 2002, de los Angeles et al. 2005), which are not highly useful for genomic studies. In consideration of the

availability of several other more efficient marker systems and the relative difficulties involved in the development of RFLP from nuclear genes, the anticipated use of RFLP markers and their significance for aquaculture genome research is limited.

Acknowledgments

Research in my laboratory is supported by grants from the United States Department of Agriculture (USDA) National Research Initiative (NRI) Animal Genome and Genetic Mechanisms Program, the USDA NRI Basic Genome Reagents and Tools Program, the Mississippi-Alabama Sea Grant Consortium, the Alabama Department of Conservation, the United States Agency for International Development (USAID), and the Binational Agricultural Research and Development Fund (BARD).

References

Apte S, B Star, and JPA Gardner. 2003. A comparison of genetic diversity between cultured and wild populations, and a test for genetic introgression in the New Zealand greenshell mussel *Perna canaliculus* (Gmelin 1791). Aquaculture, 219, pp. 193–220.

Aranishi F. 2005. Rapid PCR-RFLP method for discrimination of imported and domestic mackerel. Mar Biotechnol, 7, pp. 571–575.

Ausubel FM, R Brent, RE Kingston, DD Moore, JG Seidman, JA Smith, and K Struhl. 2003. Current Protocols in Molecular Biology, Wisley InterScience, New York.

Bossier P, X Wang, F Catania, S Dooms, G Van Stappen, E Naessens, and P Sorgeloos. 2004. An RFLP database for authentication of commercial cyst samples of the brine shrimp *Artemia* spp. (International Study on *Artemia* LXX). Aquaculture, 231, pp. 93–112.

Botstein D, RL White, M Skolnick, and RW Davis. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet, 32, pp. 314–331.

Bouchon D, C Souty-Grosset, and R Raimond. 1994. Mitochondrial DNA variation and markers of species identity in two penaeid shrimp species: *Penaeus monodon* Fabricius and *P. japonicus* Bate. Aquaculture, 127, pp. 131–144.

Chow S, N Suzuki, H Imai, and T Yoshimura. 2006. Molecular Species Identification of Spiny Lobster *Phyllosoma* Larvae of the Genus *Panulirus* from the Northwestern Pacific. Mar Biotechnol, Mar 29 (Epub ahead of print).

Cunningham CO. 2002. Molecular diagnosis of fish and shellfish diseases: present status and potential use in disease control. Aquaculture, 206, pp. 19–55.

de los Angeles Barriga-Sosa I, MY Pérez-Ramírez, F Soto-Aguirre, M Castillo-Rivera, and JL Arredondo-Figueroa. 2005. Inter-specific variation of the mitochondrial r16S gene among silversides, "Peces Blancos" (*Atherinopsidae: Menidiinae*) and its utilization for species identification. Aquaculture, 250, pp. 637–651.

Docker MF, A Dale, and DD Heath. 2003. Erosion of interspecific reproductive barriers resulting from hatchery supplementation of rainbow trout sympatric with cutthroat trout. Mol Ecol, 12, pp. 3515–3521.

Dodgson JB, HH Cheng, and R Okimoto. 1997. DNA marker technology: a revolution in animal genetics. Poult Sci, 76, pp. 1108–1114.

Gross R, K Kohlmann, and P Kersten. 2002. PCR-RFLP analysis of the mitochondrial ND-3/4 and ND-5/6 gene polymorphisms in the European and East Asian subspecies of common carp (*Cyprinus carpio* L.). Aquaculture, 204, pp. 507–516.

Gross R and J Nilsson. 1999. Restriction fragment length polymorphism at the growth hormone 1 gene in Atlantic salmon (*Salmo salar* L.) and its association with weight among the offspring of a hatchery stock. Aquaculture, 173, pp. 73–80.

Hansen MM, K-LD Mensberg, G Rasmussen, and V Simonsen. 1997. Genetic variation within and among Danish brown trout (*Salmo trutta* L.) hatchery strains, assessed by PCR-RFLP analysis of mitochondrial DNA segments. Aquaculture, 153, pp. 15–29.

Klinbunga S, B Khamnamtong, N Puanglarp, P Jarayabhand, W Yoosukh, and P Menasveta. 2005. Molecular taxonomy of cupped oysters (*Crassostrea, Saccostrea*, and *Striostrea*) in Thailand based on COI, 16S, and 18S rDNA polymorphism. Mar Biotechnol, 7, pp. 306–317.

Klinbunga S, D Siludjai, W Wudthijinda, A Tassanakajon, P Jarayabhand, and P Menasveta. 2001. Genetic heterogeneity of the giant tiger shrimp (*Penaeus monodon*) in Thailand revealed by RAPD and mitochondrial DNA RFLP analyses. Mar Biotechnol, 3, pp. 428–438.

Lehoczky I, Z Jeney, I Magyary, C Hancz, and K Kohlmann. 2005. Preliminary data on genetic variability and purity of common carp (*Cyprinus carpio* L.) strains kept at the live gene bank at Research Institute for Fisheries, Aquaculture and Irrigation (HAKI) Szarvas, Hungary. Aquaculture, 247, pp. 45–49.

Liu ZJ and JF Cordes. 2004. DNA marker technologies and their applications in aquaculture genetics. Aquaculture, 238, pp. 1–37.

Lopez-Pinon MJ, A Insua, and J Mendez. 2002. Identification of four scallop species using PCR and restriction analysis of the ribosomal DNA internal transcribed spacer region. Mar Biotechnol, 4, pp. 495–502.

Mamuris Z, C Stamatis, KA Moutou, AP Apostolidis, and C Triantaphyllidis. 2001. RFLP analysis of mitochondrial DNA to evaluate genetic variation in striped red mullet (*Mullus surmuletus* L.) and red mullet (*Mullus barbatus* L.) populations. Mar Biotechnol, 3, pp. 264–274.

Morán P, AM Pendás, and E García-Vázquez. 1996. Mitochondrial DNA variation in wild and hatchery brown trout (*Salmo trutta* L.) populations from Spain. Aquaculture, 141, pp. 59–65.

Naish KA, M Warren, F Bardakci, DOF Skibinski, and GR Carvalho. 1995. Use of DNA fingerprinting, RAPD and RAPD/RFLP markers for estimating variation between aquacultural strains of Tilapia (*Oreochromis niloticus*). Aquaculture, 137, pp. 48–49.

Ohara K, T Ariyoshi, E Sumida, and N Taniguchi. 2003. Clonal diversity in the Japanese silver crucian carp, *Carassius langsdorfii* inferred from genetic markers. Zool Sci, 20, pp. 797–804.

Palti Y, JE Parsons, and GH Thorgaard. 1999. Identification of candidate DNA markers associated with IHN virus resistance in backcrosses of rainbow (*Oncorhynchus mykiss*) and cutthroat trout (*O. clarki*). Aquaculture, 173, pp. 81–94.

Papakostas S, S Dooms, A Triantafyllidis, D Deloof, I Kappas, K Dierckens, T De Wolf, P Bossier, O Vadstein, S Kui, et al. 2006. Evaluation of DNA methodologies in identifying Brachionus species used in European hatcheries. Aquaculture, In Press.

Sambrook J, EF Frisch, and T Maniatis. 1989. Molecular Cloning, A Laboratory Manual, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Sellos D, J Moal, L Degremont, A Huvet, JY Daniel, S Nicoulaud, P Boudry, JF Samain, and A Van Wormhoudt. 2003. Structure of amylase genes in populations of Pacific Cupped oyster (*Crassostrea gigas*): tissue expression and allelic polymorphism. Mar Biotechnol, 5, pp. 360–372.

Southern E. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol, 98, pp. 503–517.

Zhang Q, SK Allen Jr, and KS Reece. 2005. Genetic variation in wild and hatchery stocks of Suminoe Oyster (*Crassostrea ariakensis*) assessed by PCR-RFLP and microsatellite markers. Mar Biotechnol, 7, pp. 588–599.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 3
Randomly Amplified Polymorphic DNA (RAPD)

Zhanjiang Liu

Random amplification of polymorphic DNA (RAPD) is a polymerase chain reaction (PCR)-based multilocus DNA fingerprinting technique. The RAPD procedure was first developed in 1990 (Welsh and McClelland 1990, Williams et al. 1990) using PCR to randomly amplify anonymous segments of nuclear DNA with a single short PCR primer (8–10 base pairs [bp] in length). Because the primers are short and relatively low annealing temperatures (often 36–40°C) are used, the likelihood of amplifying multiple products is pretty good, with each product presumably representing a different locus. Once different bands are amplified from related species, population, or individuals, RAPD markers are produced. RAPD markers thus are differentially amplified bands using a short PCR primer from random genome sites. Because most of the nuclear genome in vertebrates is noncoding, it is presumed that most of the amplified loci will be selectively neutral. Genetic variation and divergence within and between the taxa of interest are assessed by the presence or absence of each product, which is dictated by changes in the DNA sequence at each locus. RAPD polymorphisms can occur due to base substitutions at the primer binding sites or to indels in the regions between the sites. The potential power for detection of polymorphism is relatively high; typically, 5–20 bands can be produced using a given primer, and multiple sets of random primers can be used to scan the entire genome for differential RAPD bands. Because each band is considered a bi-allelic locus (presence or absence of an amplified product), polymorphic information content (PIC) values for RAPDs fall below those for microsatellites and single nucleotide polymorphisms (SNP), and RAPDs may not be as informative as amplified fragment length polymorphisms (AFLP) because fewer loci are generated simultaneously. However, because of its relatively high level of polymorphic rates, its simple procedure, and a minimal requirement for both equipment and technical skills, RAPD has been widely used in genetic analysis, including that of aquaculture species.

In this chapter, technology advances leading to the development of RAPD, the principles and molecular basis of RAPD, inheritance of RAPD markers, the power of RAPD analysis, strengths and weakness of RAPD, and applications of RAPD in aquaculture genomics research will be summerized.

Technology Advances Leading to the Development of RAPD

RAPD is a PCR-based fingerprinting technique. The invention of PCR in the mid-1980s revolutionized the entire life sciences, earning a Nobel Prize in 1993 for its inventor, Dr. Kary B. Mullis. Understanding how PCR works is fundamentally

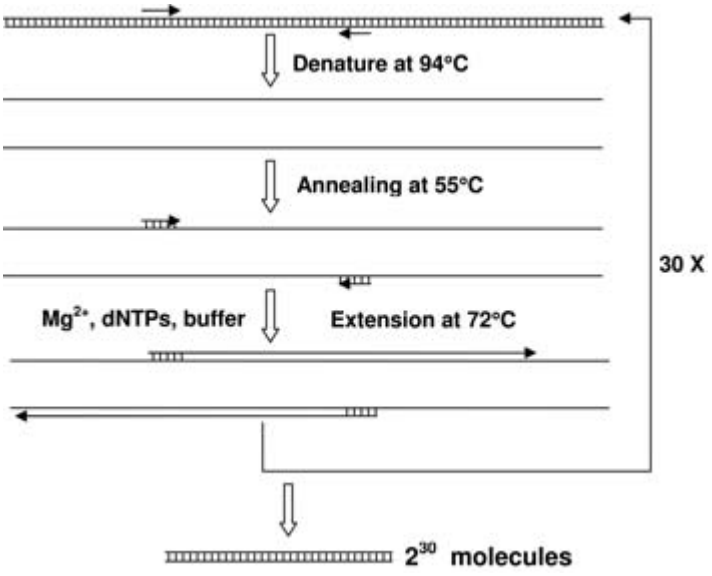


Figure 3.1. Principles and procedures of PCR.

important for appreciation of the principles of RAPD. Hence, I will briefly review the principles of PCR technology. PCR reactions start with double-stranded DNA templates. The first step is the denaturation of template DNA by heating it at 94°C; the second step is to anneal PCR primers to the templates. This step requires optimized temperatures according to the primers. Two factors significantly influence the fidelity of PCR—the length of the PCR primers and the annealing temperature. Generally speaking, the longer the PCR primer and the higher the annealing temperature, the higher the fidelity of the PCR reactions. Most often, however, PCR primers longer than 17 bp, and with an annealing temperature above 55°C, are sufficient to produce reasonably high fidelity for PCR. The third step of PCR is the extension of the annealed primers to synthesize new DNA. Once the extension is complete to the end of the template, PCR finishes its first cycle, and the original single molecule has been copied into two molecules. Let the process repeat 30 or more times, and one DNA molecule can be amplified into 2^{30} or more molecules (Figure 3.1).

Principles of RAPD

It is clear that in order to have exponential amplification, PCR requires two primers. Now we have genomic DNA, for which no sequence information is available. How can we conduct PCR reactions to produce genomic fingerprints revealing polymorphism? RAPD procedures are based on a fundamental understanding of the annealing process. At a given base position, any DNA has four possibilities of bases: A, C, G, or T. Therefore, if the primers are short enough, there would be numerous binding sites for them in genomic DNA. The odds for a perfect binding site to exist for a 10-base primer are once every 4^{10} base pairs (i.e., approximately once every million base pairs).

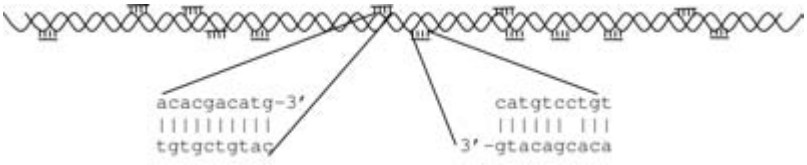


Figure 3.2. Schematic presentation of RAPD primer binding to genomic DNA. Short RAPD primers find their perfect and/or subperfect sites, anneal to genomic DNA, and amplify segments of genomic DNA when they are annealed close enough (generally <2,000 bp) on opposite strands of DNA.

In most eukaryotic organisms, such as fish, their genomes are at the billion base pairs range. There should be 1,000 perfect binding sites on each strand of the genome. However, the binding sites do not have to be perfect to initiate PCR if the annealing temperature is low enough. For instance, unless the last base at the 3′ is mismatched, when 9 out of 10 bases of the PCR primer have perfect matches to the template, PCR is likely to proceed if the annealing temperature is low. The possibility of subperfect binding greatly increases the number of binding sites in the genome from which a PCR reaction may proceed. The only exception is when the mismatches occur at the 3′ end of the primer. Therefore, there should be a large number of binding sites in a large genome for a short primer. However, PCR reactions are often limited to a certain length. Therefore, the short primers must bind to both strands of DNA close enough (within several kilobases [kb]) to produce a RAPD band. Using this principle, Welsh and McClelland (1990) and Williams and others (1990) used a single short PCR primer of 10 bases and conducted the special PCR reaction at 36°C, leading to the generation of PCR products using a single random short primer (Figure 3.2).

Based on the fundamental principles of RAPD, the technique can be regarded as a method of creating genomic fingerprints from species for which little is known about the target sequence to be amplified using arbitrarily primed PCR (AP-PCR). The creators of RAPD solved the dilemma of how to create a PCR primer without sequence information by using arbitrary short primers that increase the odds of finding suitable binding sites. The short primers require low temperatures for annealing.

Molecular Basis of RAPD Polymorphism

All mechanisms that led to the differential amplification of RAPD bands account for the molecular basis of RAPD polymorphism. First, RAPD depends on primer binding at adjacent sites on opposite strands of DNA. Any base substitutions at the primer binding sites may knock out primer binding and PCR amplification, thus leading to the loss of a RAPD band. Inversely, any base substitutions at a site originally with a sequence similar to the primer binding sites can lead to the generation of new primer binding sites. Once newly generated primer binding sites are close enough to another primer binding site on opposite strands of DNA, a RAPD band can be generated,

leading to polymorphism. Obviously, deletions and insertions within the RAPD bands would lead to either shorter or longer RAPD bands, producing polymorphism.

Inheritance of RAPD Markers

RAPD markers are inherited as Mendelian markers in a dominant fashion. As dominant markers, RAPD are scored as present/absent. Dominance means that one dose is enough, and therefore, a RAPD band is produced by dominant homozygotes as well as heterozygotes, though band intensity may differ. In spite of the theoretical validity of differentiating the dominant homozygotes from heterozygotes, variations in PCR efficiency make scoring of band intensities difficult. As a result, attempting to distinguish homozygous dominant from heterozygous individuals is not generally recommended. Also, it is difficult to determine whether bands represent different loci or alternative alleles of a single locus so that the number of loci under study can be erroneously assessed. This is especially true if the RAPD is caused by deletion or insertion within the locus rather than at the primer binding sites. As a result, the number of loci of RAPD markers can be inflated up to twofold.

As dominant markers, the alternative allele of a RAPD band is the absence of the band. Even though sometimes it is possible to determine alternative alleles by examination of the presence of alternative phases of RAPD bands, the exact nature of alternative RAPD bands must be verified by hybridization or by sequencing before calling them alternative alleles. As dominant markers, the number of RAPD bands seen in the F1 generation should equal the sum of the bands seen in the parents, assuming parental homozygosity at each locus; polymorphic RAPD then segregate in a 3:1 ratio in F2 populations, as shown in Figure 3.3 (Liu et al. 1998, 1999). If the RAPD bands are heterozygous in the parents, they segregate in a 1:1 fashion in F1 populations.

The Differentiating Power of RAPD

The RAPD approach is based on the fact that short oligonucleotide primers can bind to DNA with predicted odds. For instance, every 1 million (4^{10}) bp should contain one

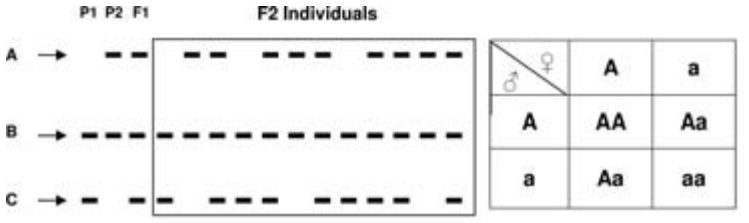


Figure 3.3. Inheritance of dominant markers. Here RAPD band A is inherited from parent 2 and band C is inherited from parent 1; both band A and C are heterozygous in F1. RAPD band B is inherited from both parents and thus are homozygous in F1. Heterozygous bands of F1 segregate in a 3:1 ratio among F2 individuals. Figure was modified from Liu and Cordes (2004).

sequence that matches with a primer of 10 nucleotides long. Therefore, a genome of 1 billion base pairs should contain 1,000 perfect binding sites for the 10-bp primer on each of its two strands of DNA. The 2,000 perfect binding sites plus many more subperfect binding sites (with 8–9 out of 10 nucleotides) would make it possible to amplify DNA using a single arbitrary short primer. The conditions for this special PCR reaction follow:

- The annealing temperature must be low because of the short primer.
- The short primer must bind to the opposite strands of DNA with its 3′ ends facing each other.
- The two binding sites must be close enough to allow a successful PCR reaction using *Taq* DNA polymerase, which often travels only several kilobases.

Generally, all these conditions can be met and often multiple bands can be amplified. Any deletion/insertion existing between the two successful primers would produce a polymorphic band. Additionally, base substitutions at primer binding sites can also cause gain or loss of amplified bands. Because about a dozen bands can be analyzed simultaneously and genome sequence information is not required, RAPD rapidly gained popularity for analysis of genetic variation in the 1990s.

Most often 5–20 bands can be amplified by using a single RAPD primer. Theoretically, primers with equal length should be equally efficient for generating RAPD bands, but G/C-rich RAPD primers were reported to produce more bands than A/T-rich primers, presumably due to stronger annealing of G/C-rich primers (Kubelik and Szabo 1995). Closely related species from which hybrids can be made often exhibit high levels of RAPD polymorphism; reproductively isolated populations often exhibit a reasonable level of RAPD polymorphism so that RAPD can be used to differentiate various strains, lines, and populations. RAPD tends to exhibit low levels of polymorphism among individuals of the same population, and thus are not ideal markers for parentage analysis, for which microsatellite markers are optimal.

Strengths and Weaknesses of RAPD

RAPDs have all of the advantages of a PCR-based marker, with the added benefit that primers are commercially available and do not require prior knowledge of the target DNA sequence or gene organization. Multilocus amplifications can be separated electrophoretically on agarose gels and stained with ethidium bromide, although higher resolution of bands has been achieved with discontinuous polyacrylamide gel electrophoresis (dPAGE) and silver staining (Dinesh et al. 1995), a somewhat costlier and more labor-intensive method. Other advantages of RAPDs are the ease with which a large number of loci and individuals can be screened.

The major weakness of RAPD is its low reproducibility due to the use of low annealing temperatures. However, if one stays conservative and scores only highly reproducible strong bands, this problem can be minimized. In our own experience, we have not encountered too much trouble with reproducibility. However, if one pushes to maximize the number of RAPD bands, then many very weak bands may not be reproduced, leading to a lack of reliability. Because of this reproducibility problem, there are reports that many RAPD bands do not follow Mendelian inheritance, though homozygous status was incorrectly assumed in cases. The second major

weakness of RAPD is its dominant mode of inheritance. Because of the dominant nature of inheritance, RAPD lack the ability to distinguish between dominant homozygotes and heterozygotes. In addition, the presence of paralogous PCR products (amplified from different DNA regions that have the same lengths and thus appear to be a single locus), limit the use of this marker system. These difficulties have limited the application of this marker in fisheries and aquaculture sciences (Wirgin and Waldman 1994).

Applications of RAPD Markers in Aquaculture Genome Research

RAPD markers have been widely used for species and strain identification in fish (Partis and Wells 1996, Liu et al. 1998, 1999) and mollusks (Klinbunga et al. 2000, Crossland et al. 1993), analysis of population structure in black tiger shrimp (Tassanakajon et al. 1998) and marine algae (van Oppen et al. 1996), analysis of genetic impact of environmental stressors (Bagley et al. 2001), and analysis of genetic diversity (Wolfus et al. 1997, Hirschfeld et al. 1999, Yue et al. 2002).

In addition to identification of species, strains, lines, and populations, RAPD markers have been extensively used in the model fish species such as zebrafish (Johnson et al. 1994). RAPD markers have also been used in many linkage-mapping studies in fish species (Table 3.1). However, as more efficient and reliable marker systems such as AFLP emerged, the use of RAPD markers in genome research declined rapidly. Due to the intrinsic problems as discussed under its weaknesses, the use of RAPD for future genome characterization of aquaculture species should be limited. Its coupled usage with codominant markers, such as microsatellites, may provide more reliable information. In closed aquaculture systems where the number of founders of the broodstock population is limited, RAPD may provide some rapid ways for association analysis of traits with markers. After the initial identification of the RAPD markers, it is highly recommended that the marker be converted into SCAR markers (sequence characterized amplified region) for further analysis. In spite of very limited uses of RAPD for long-term genome research, it is a useful marker system for rapid hybrid identification and strain identification commonly encountered in aquaculture breeding operations.

Table 3.1. Some examples of the use of RAPD markers for the construction of linkage maps in aquaculture or fish species. Note that most of these efforts were made earlier, and linkage mapping using RAPD markers is not highly recommended.

Species	Common name	References
Cyprinus carpio	Common carp	Sun and Liang 2004
Danio rerio	Zebrafish	Postlethwait et al. 1994, Mohideen et al. 2000
Oryzias latipes	Medaka	Ohtsuka et al. 1999
Oncorhynchus mykiss	Rainbow trout	Sakamoto et al. 2000
Astyanax mexicanus	Cave fish	Borowsky and Wilkens 2002
Xiphophorus sp.		Kazianis et al. 1996
Poecilia reticulata	Guppy	Khoo et al. 2003

Acknowledgments

Research in my laboratory is supported by grants from the USDA NRI Animal Genome and Genetic Mechanisms Program, the USDA NRI Basic Genome Reagents and Tools Program, the Mississippi-Alabama Sea Grant Consortium, the Alabama Department of Conservation, the USAID, and the BARD.

References

Bagley MJ, SL Anderson, and B May. 2001. Choice of methodology for assessing genetic impacts of environmental stressors: polymorphism and reproducibility of RAPD and AFLP fingerprints. Ecotoxicology, 10, pp. 239–244.

Borowsky R and H Wilkens. 2002. Mapping a cave fish genome: polygenic systems and regressive evolution. J Hered, 93, pp. 19–21.

Crossland S, D Coates, J Grahame, and PJ Mill. 1993. Use of random amplified polymorphic DNAs (RAPDs) in separating two sibling species of *Littorina*. Mar Ecol Prog Ser, 96, pp. 301–305.

Dinesh KR, WK Chan, TM Lim, and VPE Phang. 1995. RAPD markers in fishes: an evaluation of resolution and reproducibility. Asia-Pac J Mol Biol Biotechnol, 3, pp. 112–118.

Hirschfeld D, AK Dhar, K Rask, and A Alcivar-Warren. 1999. Genetic diversity in the eastern oyster (*Crassostrea virginica*) from Massachusetts using RAPD technique. J Shellfish Res, 18, pp. 121–125.

Johnson SL, CN Midson, EW Ballinger, and JH Postlethwait. 1994. Identification of RAPD primers that reveal extensive polymorphisms between laboratory strains of zebrafish. Genomics, 19, pp. 152–156.

Kazianis S, DC Morizot, BB McEntire, RS Nairn, and RL Borowsky. 1996. Genetic mapping in *Xiphophorus* hybrid fish: assignment of 43 AP-PCR/RAPD and isozyme markers to multipoint linkage groups. Genome Res, 6, pp. 280–289.

Khoo G, MH Lim, H Suresh, DK Gan, KF Lim, F Chen, WK Chan, TM Lim, and VP Phang. 2003. Genetic linkage maps of the guppy (*Poecilia reticulata*): assignment of RAPD markers to multipoint linkage groups. Mar Biotechnol, 5, pp. 279–293.

Klinbunga S, P Ampayup, A Tassanakajon, P Jarayabhand, and W Yoosukh. 2000. Development of species-specific markers of the tropical oyster (*Crassostrea belcheri*) in Thailand. Mar Biotechnol, 2, pp. 476–484.

Kubelik AR and LJ Szabo. 1995. High-GC primers are useful in RAPD analysis of fungi. Curr Genet, 28, pp. 384–389.

Liu ZJ and J Cordes. 2004. DNA marker technologies and their applications in aquaculture genetics. Aquaculture, 238, pp. 1–37.

Liu ZJ, P Li, B Argue, and R Dunham. 1998. Inheritance of RAPD markers in channel catfish (*Ictalurus punctatus*), blue catfish (*I. furcatus*) and their F1, F2 and backcross hybrids. Anim Genet, 29, pp. 58–62.

Liu ZJ, P Li, BJ Argue, and RA Dunham. 1999. Random amplified polymorphic DNA markers: usefulness for gene mapping and analysis of genetic variation of catfish. Aquaculture, 174, pp. 59–68.

Mohideen MA, JL Moore, and KC Cheng. 2000. Centromere-linked microsatellite markers for linkage groups 3, 4, 6, 7, 13, and 20 of zebrafish (*Danio rerio*). Genomics, 67, pp. 102–106.

Ohtsuka M, S Makino, K Yoda, H Wada, K Naruse, H Mitani, A Shima, K Ozato, M Kimura, and H Inoko. 1999. Construction of a linkage map of the medaka (*Oryzias latipes*) and mapping of the *Da* mutant locus defective in dorsoventral patterning. Genome Res, 9, pp. 1277–1287.

Partis L and RJ Wells. 1996. Identification of fish species using random amplified polymorphic DNA (RAPD). Mol Cell Probes, 10, pp. 435–441.

Postlethwait JH, SL Johnson, CN Midson, WS Talbot, M Gates, EW Ballinger, D Africa, R Andrews, T Carl, JS Eisen, et al. 1994. A genetic linkage map for the zebrafish. Science, 264, pp. 699–703.

Sakamoto T, RG Danzmann, K Gharbi, P Howard, A Ozaki, SK Khoo, RA Woram, N Okamoto, MM Ferguson, LE Holm, R Guyomard, and B Hoyheim. 2000. A microsatellite linkage map of rainbow trout (*Oncorhynchus mykiss*) characterized by large sex-specific differences in recombination rates. Genetics, 155, pp. 1331–1345.

Sun X and L Liang. 2004. A genetic linkage map of common carp (*Cyprinus carpio* L.) and mapping of a locus associated with cold tolerance. Aquaculture, 238, pp. 165–172.

Tassanakajon A, S Pongsomboon, P Jarayabhand, S Klinbunga, and VV Boonsaeng. 1998. Genetic structure in wild populations of black tiger shrimp (*Penaeus monodon*) using randomly amplified polymorphic DNA analysis. J Mar Biotechnol, 6, pp. 249–254.

van Oppen MJH, H Klerk, JL Olsen, and WT Stam. 1996. Hidden diversity in marine algae: some examples of genetic variation below the species level. J Mar Biol Assoc UK, 76, pp. 239–242.

Welsh J and M McClelland. 1990. Fingerprinting genomes using PCR with arbitrary primers. Nucl Acids Res, 18, pp. 7213–7218.

Williams JGK, AR Kubelik, KJ Livak, JA Rafalski, and SV Tingey. 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucl Acids Res, 18, pp. 6531–6535.

Wirgin II and JR Waldman. 1994. What DNA can do for you. Fisheries, 19, pp. 16–27.

Wolfus GM, DK Garcia, and A Alcivar-Warren. 1997. Application of the microsatellite techniques for analyzing genetic diversity in shrimp breeding programs. Aquaculture, 152, pp. 35–47.

Yue G, Y Li, F Chen, S Cho, LC Lim, and L Orban. 2002. Comparison of three DNA marker systems for assessing genetic diversity in Asian arowana (*Scleropages formosus*). Electrophoresis, 23, pp. 1025–1032.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 4
Amplified Fragment Length Polymorphism (AFLP)

Zhanjiang Liu

Multilocus DNA fingerprinting technologies based on polymerase chain reaction (PCR) are of enormous value for the study of genetic variations. These fingerprinting technologies, such as random amplified polymorphic DNA (RAPD) (Chapter 3 of this book, as well as Williams et al. 1990, Welsh and McClelland 1990) and amplified fragment length polymorphism (AFLP) (Vos et al. 1995), allow rapid generation of large amounts of genetic data. Genetic fingerprinting using these technologies does not require prior knowledge, making them "ready to be used" technologies for any species without previous genetic information. The fingerprints may be used as a tool to identity a specific DNA sample or to assess the relatedness between samples. Conserved common bands define relatedness, while polymorphic bands define differentiation in phylogenetic and population genetic analyses.

AFLP technology combines the advantages of restriction enzyme fingerprinting using restriction fragment length polymorphism (RFLP) and those of PCR-based fingerprinting using RAPD. It is based on the selective amplification of a subset of genomic restriction fragments using PCR. DNA is digested with restriction enzymes, and double-stranded DNA adaptors are ligated to the ends of the DNA fragments to generate primer-binding sites for PCR amplification. The sequence of the adaptors and the adjacent restriction site serve as primer binding sites for subsequent amplification of the restriction fragments by PCR. Selective nucleotides extending into the restriction sites are added to the 3′ ends of the PCR primers in such a way that only a subset of the restriction fragments is recognized. Only restriction fragments in which the nucleotides flanking the restriction site match the selective nucleotides will be amplified. The subset of amplified fragments is then analyzed by denaturing polyacrylamide gel electrophoresis to generate the fingerprints.

To fully appreciate the advantages and applications of AFLP, this chapter is written to describe the course of the technology development in relation to several other existing technologies, the procedures and principles, the molecular basis of polymorphism, and the potential power for genetic analysis using AFLP. As detailed below, AFLP is a technology that provides robustness, reliability, and efficiency. Its simultaneous analysis of hundreds of loci using only a single primer combination offers a robust power of differentiation. AFLP is also advantageous because markers are inherited in Mendelian fashion; it does not require prior genetic information and is therefore adaptable to genetic analysis of any species. AFLP truly provides the multilocus coverage and potential for genome-wide coverage for analysis of genetic variations. For comparisons of AFLP with other marker systems, readers are referred to other chapters of this book and a review on applications of DNA markers in fisheries and aquaculture (Liu and Cordes 2004).

Background of AFLP Technology

Genomic DNA must first be cut into small pieces for molecular analysis. Geneticists have a limited capacity for making direct analysis of large segments of DNA. Although chromosomes or chromosome segments can be directly analyzed through a special gel electrophoresis known as pulse field electrophoresis, little genetic information can be obtained from such analysis concerning genetic variation. In contrast, resolution of differentiation can be drastically increased when DNA is cut into small segments.

Restriction enzymes are site-specific "molecular scissors" for DNA (for details, see Chapter 2). They recognize specific sequences 4–8 base pairs (bp) long. In a restriction digest reaction, a restriction enzyme is mixed with genomic DNA and incubated under specific buffer and temperature conditions as required by the restriction enzyme. Within usually 1 hour of incubation, a restriction enzyme "searches" through the entire DNA lengths for its specific 4–8 bp recognition sequences. The genomic DNA is then "cut" by the enzyme whenever the proper recognition sequences are found. The cutting frequency of any restriction enzyme is directly related to the length of its recognition sequences. On average, a restriction enzyme with 4-bp recognition sequences should cut DNA once every 256 bp ($1/4^4$); a restriction enzyme with 6-bp recognition sequences should cut DNA once every 4,096 bp ($1/4^6$); and a restriction enzyme with 8-bp recognition sequences should cut DNA once approximately every 65,000 bp ($1/4^8$). These cutting frequencies should be considered when choosing restriction enzymes. This is particularly true for AFLP, which will be discussed later in the chapter. Assuming a fish genome of one billion base pairs (10^9 bp), a 4-bp cutter will digest the genome into approximately 4 million segments; a 6-bp cutter will digest the genome into about a quarter-million segments; and a 8-bp cutter will digest the genome into just 15,000 segments. In addition to the length of recognition sequences, the genomic content also affects the cutting frequency of a restriction enzyme. For instance, AT-rich genomes are in favor of restriction enzymes with AT-rich recognition sequences, while GC-rich genomes are in favor of restriction enzymes with GC-rich recognition sequences.

Molecular Basis of Genetic Variation Detected in AFLP

In the long history of evolution, genomes have evolved in each species to have a fixed number of chromosomes whose shape and sizes are constant. The number of genes and gene locations on each chromosome are also relatively constant so that genetic linkage maps can be constructed. Such structural and organizational order is maintained by accurate inheritance of genes from generation to generation. However, just as constant as the inheritance of genes and traits from parents to progenies, mutations are also constant events. Mutations can happen spontaneously or under induction of adverse environmental cues such as radiation, UV light, or chemical mutagens. Spontaneous mutations occur at a very low rate of 1×10^{-5}–2×10^{-6} per gene per generation. Assuming an average gene size of 2,000 bp, this low spontaneous mutation rate translates into only 1 to 5 base mutations throughout the entire genome of one billion base pairs per generation. However, through the long process of evolution, many mutations have accumulated. The basic idea behind genetic analysis lies in using

the accumulation of different mutations in reproductive isolated populations and individuals.

Mutations are random events and can happen in any part of the genome, although mutation hot spots are often reported. As a result, mutations are accumulated in evolution more often in noncoding regions. First, because the nonprotein coding regions of the genome account for the vast majority of the entire genome, most mutations occur by chance in these regions; second, nature has placed great selection pressure for advantageous mutations and neutral mutations, but against deteriorating mutations inside the protein coding sequences. In the coding regions, silent mutations (single base substitutions that do not change the amino acid sequence) are the most predominant. Mutations can be categorized at the molecular level as having been caused by deletions, insertions, inversions, base substitutions, and rearrangements. Deletions are losses of, while insertions are additions of, DNA bases of variable sizes ranging from a single base to long stretches of DNA. Base substitutions are changes of a specific base to any other of the three bases. For instance, base A can be mutated to any of C, G, or T. Mutations from purines (A and G) to purines or from pyrimidines (C and T) to pyrimidines are called transitions; mutations from purines to pyrimidines or vice versa are called transversions. Transitions are the most frequent mutations because the chemical reactions involved in such mutations are more likely to occur. In relation to molecular analysis, deletions and insertions are expected to cause changes of fragment lengths of at least one base pair while base substitutions generally do not affect fragment sizes unless the base substitutions cause the gain or loss of restriction sites. A base substitution within the restriction enzyme recognition site causes loss of the restriction site, and therefore, leads to loss of the restriction fragment. In contrast, a single base change may lead to the generation of a new cutting site for the restriction enzyme. For instance, the recognition sequences for restriction enzyme *Eco* RI are GAATTC. If the original sequence was GgATTC, a single base change of the second G into A would generate a new restriction site for *Eco* RI and the production of an additional restriction fragment. Rearrangements do not affect fragment lengths unless the rearranged fragments contain restriction enzyme sites.

Other Genetic Variations at the Molecular Level Affecting AFLP Profiles

In addition to the mechanisms of mutations mentioned above, several other highly mutable sequences should also be noted because they may account for a significant portion of polymorphism as revealed by AFLP analysis. The first is microsatellite sequences. As detailed in Chapter 5, microsatellites are simple sequence repeats of 1–6 bp. High levels of mutation rates can happen at microsatellite loci. In some cases, mutation rates can be as high as 0.2% per locus per generation (Crawford and Cuthbertson 1996, Levinson and Gutman 1987). Such a high mutation rate is believed to be caused by slippage of DNA polymerase with the repeated microsatellite sequences, leading to microsatellite expansion or contraction. The differences in repeat numbers of microsatellite sequences cause changes in fragment lengths. In a sense, this type of mutation is a special form of insertions or deletions. Due to large numbers of microsatellite loci existing in fish and their high mutation rates, their contribution to the overall polymorphism of genomes should

not be neglected regardless of the approach used for genetic variation analysis. Secondly, unequal crossing over of minisatellite and satellite sequences may also contribute to a significant level of genetic variations among genomes.

Molecular Analysis Related to Development of AFLP Technology

AFLP methodology was developed by using and combining several of the technological advances that ushered in the genomics era; it was based on RFLP and PCR reactions resolved by sequencing gel electrophoresis.

The need for a sensitive, efficient approach to analyzing genetic variation on a genomic scale was evident early on during the genomics revolution. PCR appeared to offer the power necessary for such an approach and was used in the development of RAPD in 1990 (Williams et al. 1990, Welsh and McClelland 1990). However, RAPD's usefulness is limited by its low reproducibility because of low annealing temperatures necessitated by using short, arbitrary primers during PCR.

AFLP combines the strengths of RFLP and RAPD. It is a PCR-based approach requiring only a small amount of starting DNA. It does not require any prior genetic information or probes, and it overcomes the problem of low reproducibility inherent to RAPD. AFLP is capable of producing far greater numbers of polymorphic bands than RAPD in a single analysis, significantly reducing costs and making possible the genetic analysis of closely related populations. The use of AFLP markers in genetic linkage mapping (Meksem et al. 1995, Cho et al. 1996, Mackill et al. 1996) and analyses of genetic resource pools (Folkertsma et al. 1996, Keim et al. 1997) has facilitated progress that would otherwise take a much longer time using other technologies. It is particularly well adapted for stock identification because of the robust nature of its analysis. The other advantage of AFLP is its ability to reveal genetic conservation as well as genetic variation. In this regard, it is superior to microsatellites for applications in stock identification. Microsatellites often possess large numbers of alleles, too many to obtain a clear picture with small samples. Identification of stocks using microsatellites, therefore, would require large sample sizes. For instance, if 10 fish are analyzed, each of the 10 fish may exhibit distinct genotypes at a few microsatellite loci, making it difficult to determine relatedness without any commonly conserved genotypes. In closely related populations, AFLP can readily reveal commonly shared bands that define the common roots in a phylogenetic tree and polymorphic bands that define branches in the phylogenetic tree.

The Procedures and Principles of AFLP Analysis

Genetic variations are widely spread among genomes of even very closely related individuals. The problem is how to reveal the very minor differences among genomes. In principle, AFLP can be viewed as a multilocus or genome-wide RFLP analysis (Figure 4.1). The technique starts with restriction digestion of genomic DNA using two restriction enzymes, most often, *Eco* RI and *Mse* I. *Eco* RI recognizes a 6-bp sequence of GAATTC, and *Mse* I recognizes a 4-bp sequence of AATT. For a genome of one billion base pairs, *Eco* RI digestion should produce about 250,000 fragments,

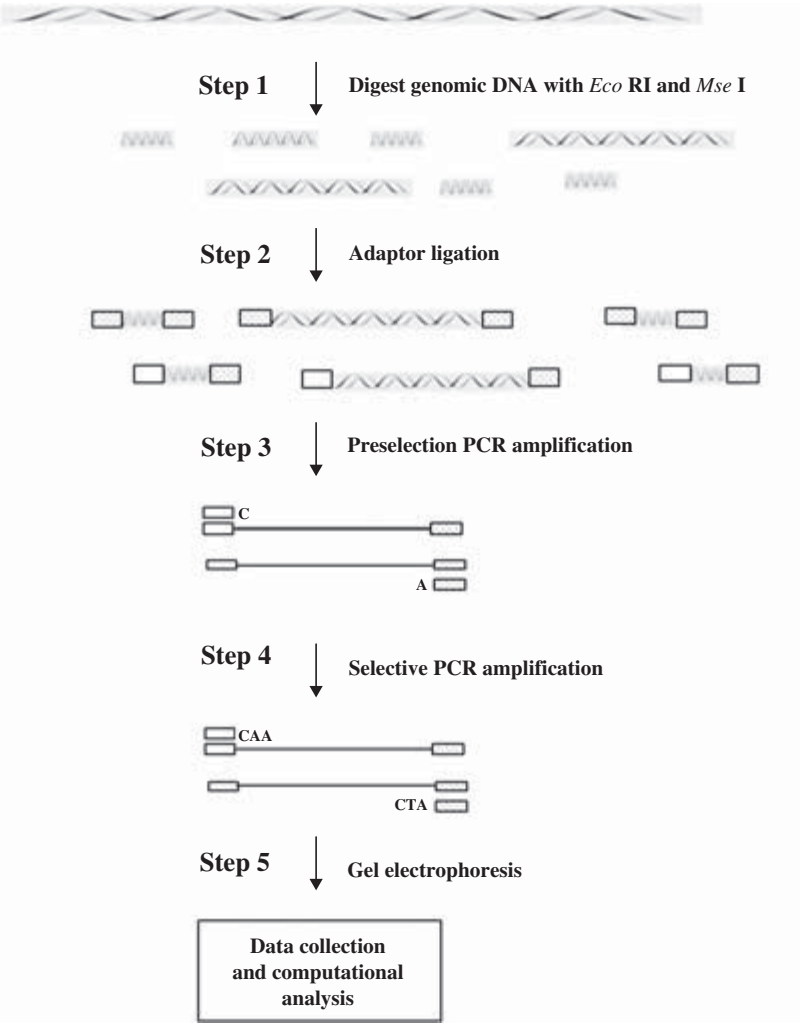


Figure 4.1. Schematic presentation of AFLP analysis. Step 1, genomic DNA is digested by *Eco*RI and *Mse*I into many segments of various sizes. For a genome of 10^9 bp, you expect ~250,000 *Eco*RI fragments (10^9/4,000) and 4 million *Mse*I fragments (10^9/256); step 2, adaptors are ligated to the ends of the DNA fragments. The majority of fragments should be *Mse*I-*Mse*I fragments, some *Eco*RI-*Mse*I fragments, few, if any, *Eco*RI-*Eco*RI fragments. This step is to add adaptors with known sequences to create PCR primer binding sites. For a genome of 10^9 bp, you expect 2× 250,000 *Eco*RI-*Mse*I fragments. Note that the sequence of *Eco*RI adaptor (open box) is different from that of the *Mse*I adaptor (dotted box); step 3, preselection amplification of a subset of the restriction fragments by adding an extra arbitrary base at the 3′ end of the PCR primers, which leads to 1/16 fragments to be amplified; step 4, selective amplification of a subset of the restriction fragments by adding three extra arbitrary bases at the 3′ end of the PCR primers, which leads to a subset of *Eco*RI-*Mse*I fragments (1/4,096) to be amplified; step 5, PCR products are resolved on a sequencing gel.

and *Mse* I digestion should produce 4 million fragments. Because the 4-bp cutter *Mse* I cuts DNA at a frequency 16 times greater than *Eco* RI, essentially all *Eco* RI fragments should be further digested by *Mse* I. The double enzyme digest would produce approximately 500,000 *Eco* RI-*Mse* I fragments (one original *Eco* RI fragment is now cleaved by many *Mse* I sites leaving both ends as *Eco* RI-*Mse* I fragments), and about 4 million *Mse* I-*Mse* I fragments.

The second step of AFLP analysis is to add adaptors on both ends of digested DNA fragments. The *Eco* RI-*Mse* I fragments must be amplified by PCR to be detected because they represent a small amount of DNA. However, there is no sequence information about these fragments. The first challenge is to "create" two stretches of known sequences in each of these fragments for PCR. This can be achieved simply by connecting a short segment of DNA with known sequences on the *Eco* RI end and a short segment of DNA with different known sequences on the *Mse* I end. These short segments of DNA with known sequences are called *Eco* RI adaptors and *Mse* I adaptors. They are called adaptors because they harbor specific end sequences allowing them each to be perfectly paired and ligated to the double digested *Eco* RI-*Mse* I fragments. After ligation, each *Eco* RI-*Mse* I fragment now harbors known sequences on both ends allowing PCR amplification of these segments by using primers with the same sequences as the adaptors.

The third step of AFLP is the preselection PCR amplification. In the 500,000 *Eco* RI-*Mse* I fragment pool, one can imagine that many of these *Eco* RI-*Mse* I fragments must exhibit size difference or length polymorphism even between two highly related individuals. However, 500,000 fragments are too many to be resolved in any kind of gel electrophoresis. This demands that somehow the 500,000 bands must be reduced approximately 2,000 times to reach the resolvable goal of a couple hundred bands. Vos and others (1995) intelligently met this challenge by adding additional arbitrary bases at the 3′ end of the PCR primers. As each extra arbitrary base is added, the PCR primer can match to only 1/4 subset of the fragments because at each base of DNA, and there are four possibilities: A, C, G, or T. When a given base is added to the 3′ end of the PCR primer, only 1/4 of the total fragments are amplifiable. When a single base is added to the 3′ end of both PCR primers, only a subset of 1/16 of the total fragments will be amplifiable. When two additional bases are added to each of the PCR primers, the reduction power is now 256 (16×16). When three additional bases are added to each 3′ end of the two PCR primers, the reduction power now is 4,096 (64×64). Now with a reduction power of 4,096, the original 500,000 fragments should become about 100 bands. These bands can then be visualized after electrophoresis. The preselective PCR first reduces the *Eco* RI-*Mse* I fragments to a subset containing 1/16 of the original fragments. The selective PCR further reduces the number of bands by amplifying only a subset of the preselective PCR products. AFLP chooses to analyze only the *Eco* RI-*Mse* I fragments. This is achieved by labeling only *Eco* RI primers. Since the *Mse* I primer is not labeled, none of the amplified *Mse* I-*Mse* I fragments are visible during electrophoresis.

The Power of AFLP Analysis

It is possible to scan the entire genome for examination of all 500,000 *Eco* RI-*Mse* I fragments by use of all possible combinations of the selective bases. That would take 64 *Eco* RI primers and 64 *Mse* I primers or 4,096 primer combinations. However, it is

probably never necessary to perform such an exhaustive analysis. Since more than 100 loci can be analyzed by a single primer combination, a few primer combinations should display thousands of fingerprints. For genetic resource analysis, the number of primer combinations required for construction of phylogenetic trees/dendrograms depends on the level of polymorphism in the populations, but probably will take no more than 10 primer combinations. However, dense genetic maps can be constructed by using a large number of primer combinations.

The potential power of AFLP in the study of genetic variation is enormous. In principle, any combination of a 6-bp cutter with a 4-bp cutter in the first step can be used to determine potential fragment length polymorphism. In the above tour through the procedures, *Eco* RI and *Mse* I were used as restriction enzymes to examine the 500,000 *Eco* RI-*Mse* I fragments. Theoretically, 4,096 primer combinations compose a complete genome-wide scan of the fragment length polymorphism using the two restriction enzymes. Because hundreds of restriction endonucleases are commercially available, the total power of AFLP for analysis of genetic variation can never be exhausted.

Molecular Basis of AFLP Polymorphism

AFLP analysis is an advanced form of RFLP. Therefore, the molecular basis for RFLP and AFLP are the same. First, any deletions and/or insertions between the two restriction enzymes (e.g., between *Eco* RI and *Mse* I in the above example) will cause shifts in fragment sizes. Second, base substitution at the restriction sites will lead to loss of restriction sites and thus a size change. However, only base substitutions in all *Eco* RI sites and 1/8 of *Mse* I sites are detected by AFLP since only the *Eco* RI primer is labeled and AFLP is designed to analyze only the *Eco* RI-*Mse* I fragments. Third, base substitutions leading to new restriction sites may also produce AFLP. Once again, gaining *Eco* RI sites always leads to production of AFLP, gaining *Mse* I sites must be within the *Eco* RI-*Mse* I fragments to produce new AFLP. In addition to the common mechanisms involved in polymorphism of RFLP and AFLP, AFLP also scans for any base substitutions at the first three bases immediately after the two restriction sites. Considering large numbers of restriction sites for the two enzymes (250,000 *Eco* RI sites and 500,000 *Mse* I sites immediately next to *Eco* RI sites), a complete AFLP scan would also examine more than 2 million bases immediately adjacent to the restriction sites.

Inheritance of AFLP Markers

AFLP markers are inherited in a Mendelian fashion as dominant markers. Similar to the traditional meaning of dominance in genetics, one dose is enough to determine phenotype (hence the band patterns). Dominant markers provide relatively less genetic information since homozygous and heterozygous individuals cannot be differentiated; they each produce a band at the locus though band intensities may vary depending on allele numbers. Although double alleles often produce double the amount of PCR products, homozygous alleles and heterozygous alleles cannot be distinguished

with certainty. As detailed below, caution must be exercised when scoring AFLP markers as codominant markers.

Major Strengths and Weaknesses of AFLP Markers

Several major strengths make AFLP the markers of choice in certain situations. The first strength is the procedure's requirement of no prior molecular information for application to the species of interest. This is particularly useful for aquaculture species where there is often no molecular information available. Second, AFLP is highly robust allowing the generation of a large number of polymorphic markers with limited efforts and resources. Third, when the robustness is coupled to the availability of many primer combinations, AFLP is a very powerful marker system for genomic differentiation. With such a technique, very minor genomic differences can be readily revealed. Fourth, as long as PCR primers are used, stringent annealing temperatures can be used for high reproducibility. This is definitely a significant improvement over the less reliable RAPD procedure (see Chapter 3). Finally, AFLP markers are relatively economical because each primer combination can often produce many polymorphic markers. Even though AFLP kits are relatively expensive, the cost on a per polymorphic marker basis is low. The major weakness of AFLP markers is their dominant nature of inheritance. Genetic information is limited with dominant markers because essentially only one allele is scored; at the same time, since the true alternative allele is scored as a different locus, AFLP also inflates the number of loci under study. As dominant markers, information transfer across laboratories is difficult. In addition, AFLP is more technically demanding, requiring special equipment such as automated DNA sequencers for optimal operations.

Genotyping AFLP Gels

AFLP markers are inherited as dominant markers. Because of the dominance nature of AFLP, they are scored as presence/absence type of markers in genotyping. Each band is treated as a locus (not an allele). Although the true alternative allele must be somewhere in the gel with a different fragment size, there is practically no way to know the exact location. In some cases, complementary phases of bands are observed, indicating they may be the alternative alleles of the locus, but in the absence of molecular evidence, each band is still scored as a separate locus. Therefore, the total number of AFLP loci under analysis is inflated about twofold because all of the alleles are treated as loci. Under this treatment, the presence of one band is treated as one allele at the locus, and the absence of the band is treated as the alternative allele.

In strictly controlled mating systems, it is possible to score AFLP markers as codominant markers. In such cases, the scoring is based not only on length polymorphism, but also on intensity polymorphism. The rationale is that two alleles in homozygotes should produce twice the amount of PCR products as that produced from a single allele in heterozygotes. As a matter of fact, computer software is available for quantitative scoring of intensity polymorphism. AFLP-Quantar™ Pro marketed by Keygene Products B.V. in the Netherlands is an example. In spite of its success, I would like to urge caution in

the use of intensity polymorphism, simply because of the nonlinear nature of PCR at high rounds of cycles. For identification of stocks and population analysis, use of intensity polymorphism should be discouraged because scoring may be extremely difficult with samples from random mating populations.

The term "informative AFLP" is used to indicate only polymorphic AFLP bands in genetic linkage mapping analysis. In the case of linkage mapping, only polymorphic bands are expected to segregate and thus provide genetic linkage information. Therefore, commonly shared nonpolymorphic bands are not scored. For population studies, all of the bands are actually informative. In fact, the commonly shared bands are extremely important since they define the common ancestor or roots for dendrogram grouping. The shared bands are used to calculate Nei's similarity F values (Nei and Li 1979). Of course, the polymorphic bands provide information about differentiation or branches for dendrogram grouping. Therefore, all AFLP bands need to be scored for population genetic analysis.

Application of AFLP for Aquaculture Genome Research

AFLP is well adapted for many types of genetic analysis such as:

- molecular systematics
- analysis of population structures
- migration
- hybrid identification
- strain identification
- parentage identification
- genetic resource analysis
- genetic diversity
- reproduction contribution
- endangered species protection
- molecular ecology
- marker-assisted selection
- genome mapping

Different authors discuss the applications of AFLP later in this book concerning analysis of data and choice of models and software for population genetic analysis, therefore, here I will only briefly discuss its application in fish population studies with an emphasis on genetic linkage analysis using AFLP.

Despite the advantages of AFLP, published literature on its application for the analysis of genetic variation of fish populations is still limited (Seki et al. 1999, Jorde et al. 1999, Sun et al. 1999, Cardoso et al. 2000, Chong et al. 2000, Kai et al. 2002, Mickett et al. 2003, Whitehead et al. 2003, Mock et al. 2004, Campbell and Bernatchez 2004, Simmons et al. 2006). Many AFLP analyses in fish so far have been limited to genetic linkage analysis (Liu et al. 1998, 1999; Kocher et al. 1998; Griffiths and Orr 1999; Agresti et al. 2000; Robison et al. 2001; Rogers et al. 2001; Liu et al. 2003; Li et al. 2003; William et al. 2005), and analysis of parental genetic contribution involving interspecific hybridization (Young et al. 2001) and meiogynogenesis (Felip et al. 2000). In a recent study of the black rockfish (*Sebastes inermis*), Kai and others (2002) used AFLP to distinguish three-color morphotypes, in which diagnostic AFLP loci were identified as well as loci with significant frequency differences. In such reproductive isolated populations, it is likely that "fixed markers" of AFLP can be identified to serve as diagnostic markers. Fixed markers are associated most often with relatively less migratory, reproductive

Table 4.1. Some examples of the use of AFLP markers for the construction of linkage maps in aquaculture or fish species.

Species	Common name	Reference
Oncorhynchus mykiss	Rainbow trout	Young et al. 1998
Salmo salar	Atlantic salmon	Moen et al. 2004
Salvelinus alpinus	Arctic char	Woram et al. 2004
Oreochromis sp.	Tilapia	Kocher et al. 1998 Agresti et al. 2000
Ictalurus punctatus	Channel catfish	Liu et al. 2003
Clarias macrocephalus	Walking catfish	Poompuang and Na-Nakorn 2005
Paralichthys olivaceus	Japanese flounder	Coimbra et al. 2003
Plecoglossus altivelis	Ayu	Watanabe et al. 2004
Penaeus monodon	Black tiger shrimp	Wilson et al. 2002
Penaeus japonicus	Kuruma prawn	Li et al. 2003
Penaeus vannamei	White shrimp	Pérez et al. 2004
Penaeus chinensis	Chinese shrimp	Li et al. in press
Crassostrea virginica	Eastern oyster	Yu and Guo 2003
Chlamys farreri	Zhikong scallop	Li et al. 2005
Haliotis discus hannae	Pacific abalone	Liu et al. 2006

isolated populations (Kucuktas et al. 2002). With highly migratory fish species, fixed markers may not be available. However, distinct populations are readily differentiated by difference in allele frequencies. For instance, Chong and others (2000) used AFLP for the analysis of five geographical populations of Malaysian river catfish (*Mystus nemurus*) and found that AFLP was more efficient for the differentiation of subpopulations and for the identification of genotypes within the populations than RAPD, although similar clusters of the populations were concluded with either analysis.

AFLP can be used effectively for genetic linkage mapping. As a matter of fact, many genetic linkage maps have been constructed using AFLP markers among aquaculture species, as summarized in Table 4.1. However, AFLP as a dominant marker, lacks the ability to be transferred across species borders, and it is difficult to transfer data among laboratories. As a result of its high efficiency, it is well suited for association analysis of traits with markers. However, after initial identification of associated AFLPs, it is highly recommended that such AFLPs be converted to sequence characterized amplified region (SCAR) markers.

Conclusion

AFLP analysis is a robust, multilocus PCR-based DNA fingerprinting technique that provides the most efficient, reliable, and economical analysis of population genetics. AFLPs are nuclear DNA markers inherited in Mendelian fashion, in contrast to environmental markers and mitochondrial DNA markers. As compared to other nuclear markers such as RFLP and RAPD, AFLPs provide a much greater level of polymorphism and a much wider genomic coverage. AFLP is probably also superior to microsatellites for population genetic studies because of its ability to display hundreds

of loci simultaneously. However, AFLP markers are inherited as dominant markers. Caution should be exercised for transfer of information across laboratories. The need for special equipment such as sequencers may limit its wide application. These disadvantages can be compensated for by the robustness of the multilocus AFLP analysis, which not only provides high levels of polymorphism, but also provides a great level of band sharing, which is required to establish relatedness among populations. Most importantly, AFLP (and also RAPD) analysis does not require any previous knowledge and thus is suitable to population genetic analysis of any species. Because of these advantages, the application of AFLP in fish population genetic studies is increasing. As time goes on, its application in the studies of fish population genetics is likely to widen. For genome research, use of AFLP markers may provide a rapid shortcut for the assessment of markers linked to certain traits, but its coupled uses with codominant markers such as microsatellites should be beneficial. In a well-defined closed mating system involving limited number of founders, genetic mapping using AFLP can add much greater resolution to framework linkage maps made with microsatellites.

Acknowledgments

Research in my laboratory is supported by grants from the USDA NRI Animal Genome and Genetic Mechanisms Program, the USDA NRI Basic Genome Reagents and Tools Program, the Mississippi-Alabama Sea Grant Consortium, the Alabama Department of Conservation, the USAID, and the BARD.

References

Agresti JJ, S Seki, A Cnaani, S Poompuang, EM Hallerman, N Umiel, G Hulata, GAE Gall, and B May. 2000. Breeding new strains of tilapia: development of an artificial center of origin and linkage map based on AFLP and microsatellite loci. Aquaculture, 185, pp. 43–56.

Bagley MJ, SL Anderson, and B May. 2001. Choice of methodology for assessing genetic impacts of environmental stressors: polymorphism and reproducibility of RAPD and AFLP fingerprints. Ecotoxicology, 10, pp. 239–244.

Campbell D and L Bernatchez. 2004. Generic scan using AFLP markers as a means to assess the role of directional selection in the divergence of sympatric whitefish ecotypes. Mol Biol Evol, 21, pp. 945–956.

Cardoso SRS, NB Eloy, J Provan, MA Cardoso, and PCG Ferreira. 2000. Genetic differentiation of *Euterpe edulis* Mart. Populations estimated by AFLP analysis. Mol Ecol, 9, pp. 1753–1760.

Chalhoub BA, S Thibault, V Laucou, C Rameau, H Hofte, and R Cousin. 1997. Silver staining and recovery of AFLP amplification products on large denaturing polyacrylamide gels. Biotechniques, 22, pp. 216–218.

Cho YG, MW Blair, O Panaud, and SR McCouch. 1996. Cloning and mapping of variety-specific rice genomic DNA sequences: amplified fragment length polymorphisms (AFLP) from silver-stained polyacrylamide gels. Genome, 39, pp. 373–378.

Chong LK, SG Tan, K Yusoff, and SS Siraj. 2000. Identification and characterization of Malaysian river catfish, *Mystus nemurus* (C&V): RAPD and AFLP analysis. Biochem Genet, 38, pp. 63–76.

Coimbra MRM, K Kobayashi, S Koretsugu, O Hasegawa, E Ohara, and A Ozaki. 2003. A genetic linkage map of the Japanese flounder, *Paralichtys olivaceus*. Aquaculture, 220, pp. 203–218.

Crawford AM and RP Cuthbertson. 1996. Mutations in sheep microsatellites. Genome Res, 6, pp. 876–879.

Dresler-Nurmi A, Z Terefework, S Kaijalainenl, K Lindstrom, and A Hatakka. 2000. Silver stained polyacrylamide gels and fluorescence-based automated capillary electrophoresis for detection of amplified fragment length polymorphism patterns obtained from white-rot fungi in the genus *Trametes*. J Microbiol Methods, 41, pp. 161–172.

Felip A, G Martinez-Rodriguez, F Piferrer, M Carrillo, and S Zanuy. 2000. AFLP Analysis Confirms Exclusive Maternal Genomic Contribution of Meiogynogenetic Sea Bass (*Dicentrarchus labrax* L.). Mar Biotechnol, 2, pp. 301–306.

Folkertsma RT, JNR van der Voort, KE de Groot, PM van Zandvoort, A Schots, FJ Gommers, J Helder, and J Backer. 1996. Gene pool similarities of potato cyst nematode populations assessed by AFLP analysis. Mol Plant-Microbe Interactions, 9, pp. 47–54.

Griffiths R and K Orr. 1999. The use of amplified fragment length polymorphism (AFLP) in the isolation of sex-specific markers. Mol Ecol, 8, pp. 671–674.

Jorde PE, S Palm, and N Ryman. 1999. Estimating genetic drift and effective population size from temporal shifts in dominant gene marker frequencies. Mol Ecol, 8, pp. 1171–1178.

Kai Y, K Nakayama, and T Nakabo. 2002. Genetic differences among three colour morphotypes of the black rockfish, *Sebastes inermis*, inferred from mtDNA and AFLP analyses. Mol Ecol, 11, pp. 2591–2598.

Keim P, A Kalif, J Schupp, K Hill, SE Travis, K Richmond, DM Adair, M Hugh-Jones, CR Kuske, and P Jackson. 1997. Molecular evolution and diversity in *Bacillus anthracis* as detected by amplified fragment length polymorphism markers, J Bacteriol, 179, pp. 818–824.

Kocher T, W Lee, H Sobolewska, D Penman, and B Andrew. 1998. A genetic linkage map of a cichlid fish the tilapia (*Oreochromis niloticus*). Genetics, 148, pp. 1225–1232.

Kucuktas H, BK Wagner, R Shopen, M Gibson, RA Dunham, and ZJ Liu. 2002. Genetic analysis of Ozark Hellbenders (*Cryptobranchus alleganiensis bishopi*) utilizing RAPD markers. Proc Ann Conf SEAFWA, 55, pp. 126–137.

Levinson G and GA Gutman. 1987. High frequency of short frameshifts in poly-CA/GT tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. Nucleic Acid Res, 15, pp. 5323–5338.

Li L and X Guo. 2004. AFLP-based genetic linkage maps of the pacific oyster Crassostrea gigas Thunberg. Mar Biotechnol, 6, pp. 26–36.

Li L, J Xiang, X Liu, Y Zhang, B Dong, and X Zhang. 2005. Construction of AFLP-based genetic linkage map for Zhikong scallop, *Chlamys farreri* Jones et Preston and mapping of sex-linked markers. Aquaculture, 245, pp. 63–73.

Li Y, K Byrne, E Miggiano, V Whan, S Moore, S Keys, P Crocos, N Preston, and S Lehnert. 2003. Genetic mapping of the kuruma prawn *Penaeus japonicus* using AFLP markers. Aquaculture, 219, pp. 143–156.

Li Z, J Li, Q Wang, Y He, and P Liu. (in press) AFLP-based genetic linkage map of marine shrimp *Penaeus* (*Fenneropenaeus*) *chinensis*. Aquaculture.

Liu X, X Guo, Q Gao, H Zhao, and G Zhang. 2006. A preliminary genetic linkage map of the Pacific abalone *Haliotis discus hannae* Ino. Mar. Biotechnol. 8, pp. 386–397.

Liu Z, P Li, H Kucuktas, A Nichols, G Tan, X Zheng, BJ Argue, and RA Dunham. 1999. Development of amplified fragment length polymorphism (AFLP) markers suitable for genetic linkage mapping of catfish. Trans Am Fish Soc, 128, pp. 317–327.

Liu Z, A Nichols, P Li, and RA Dunham. 1998. Inheritance and usefulness of AFLP markers in channel catfish (*Ictalurus punctatus*) blue catfish (*I furcatus*) and their F1, F2 and backcross hybrids. Mol Gen Genet, 258, pp. 260–268.

Liu ZJ and J Cordes. 2004. DNA marker technologies and their applications in aquaculture genetics. Aquaculture, 238, pp. 1–37.

Liu ZJ, A Karsi, P Li, D Cao, Z Ju, and R Dunham. 2003. An AFLP-based genetic linkage map of channel catfish (*Ictalurus punctatus*) constructed by using an interspecific hybrid resource family. Genetics, 165, pp. 687–694.

Mackill DJ, Z Zhang, ED Redona, and PM Colowit. 1996. Level of polymorphism and genetic mapping of AFLP markers in rice. Genome, 39, pp. 969–977.

Maria R, M Coimbra, K Kobayashi, S Koretsugu, O Hasegawa, E Ohara, A Ozaki, T Sakamoto, K Naruse, and N Okamoto. 2003. A genetic linkage map of the Japanese flounder, *Paralichthys olivaceus*. Aquaculture, 220, pp. 203–218.

Meksem K, D Leiser, J Peleman, M Zabeau, F Salamini, and C Gebhardt. 1995. A high-resolution map of the vicinity of the R1 locus on chromosome V of potato based on RFLP and AFLP markers, Mol Gen Genet, 249, pp. 74–81.

Mickett K, C Morton, J Feng, P Li, M Simmons, D Cao, R Dunham, and ZJ Liu. 2003. Assessing genetic diversity of domestic populations of channel catfish (*Ictalurus punctatus*) in Alabama using AFLP markers. Aquaculture, 228, pp. 91–105.

Mock KE, JC Brim-Box, MP Miller, ME Downing, and WR Hoeh. 2004. Genetic diversity and divergence among freshwater mussel (*Anodonta*) populations in the Bonneville Basin of Utah. Mol Ecol, 13, pp. 1085–1098.

Moen T, B Hoyheim, H Munck, and L Gomez-Raya. 2004. A linkage map of Atlantic salmon (*Salmo salar*) reveals an uncommonly large difference in recombination rate between the sexes. Anim Genet, 35, pp. 81–92.

Nei M and WH Li. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA, 76, pp. 5269–5273.

Nichols KM, WP Young, RG Danzmann, BD Robison, C Rexroad, M Noakes, RB Phillips, P Bentzen, I Spies, K Knudsen, FW Allendorf, BM Cunningham, J Brunelli, H Zhang, S Ristow, R Drew, KH Brown, PA Wheeler, and GH Thorgaard. 2003. A consolidated linkage map for rainbow trout (*Oncorhynchus mykiss*). Anim Genet, 34, pp. 102–115.

Pérez F, C Erazo, M Zhinaula, F Volckaert, and J Calderón. 2004. A sex-specific linkage map of the white shrimp *Penaeus (Litopenaeus) vannamei* based on AFLP markers. Aquaculture, 242, pp. 105–118.

Poompuang S and U Na-Nakorn. 2005. A preliminary genetic map of walking catfish (*Clarias macrocephalus*). Aquaculture, 232, pp. 195–203.

Reineke A and P Karlovsky. 2000. Simplified AFLP protocol: replacement of primer labeling by the incorporation of alpha-labeled nucleotides during PCR. Biotechniques, 28, pp. 622–623.

Robison BD, PA Wheeler, K Sundin, P Sikka, and GH Thorgaard. 2001. Composite interval mapping reveals a major locus influencing embryonic development rate in rainbow trout (*Oncorhynchus mykiss*). J Hered, 92, pp. 16–22.

Rogers SM, D Campbell, SJ Baird, RG Danzmann, and L Bernatchez. 2001. Combining the analyses of introgressive hybridisation and linkage mapping to investigate the genetic architecture of population divergence in the lake whitefish (*Coregonus clupeaformis*, Mitchill). Genetica, 111, pp. 25–41.

Seki S, JJ Agresti, GAE Gall, N Taniguchi, and B May. 1999. AFLP analysis of genetic diversity in three populations of ayu *Plecoglossus altivelis*. Fish Sci, 65, pp. 888–892.

Simmons M, K Mickett, H Kucuktas, P Li, R Dunham, and ZJ Liu. 2006. Comparison of domestic and wild catfish populations provide no evidence for genetic impact. Aquaculture, 252, pp. 133–146.

Sun Y, W Song, Y Zhong, R Zhang, TJ Abatzopoulos, and R Chen. 1999. Diversity and genetic differentiation in *Artemia* species and populations detected by AFLP markers. Int J Salt Lake Res, 8, pp. 341–350.

Vos P, R Hogers, M Bleeker, M Reijans, T van de Lee, M Hornes, A Frijters, J Pot, J Peleman, M Kuiper, and M Zabeay. 1995. AFLP: a new technique for DNA fingerprinting. Nucl Acids Res, 23, pp. 4407–4414.

Watanabe T, H Fujita, K Yamasaki, S Seki, and N Taniguchi. 2004. Preliminary study on linkage mapping based on microsatellite DNA and AFLP markers using homozygous clonal fish in ayu (*Plecoglossus altivelis*). Mar. Biotechnol, 6, pp. 327–334.

Welsh J and M McClelland. 1990. Fingerprinting genomes using PCR with arbitrary primers. Nucl Acids Res, 18, pp. 7213–7218.

Whitehead A, SL Anderson, KM Kuivila, JL Roach, and B May. 2003. Genetic variation among interconnected populations of *Catostomus occidentalis*: implications for distinguishing impacts of contaminants from biogeographical structuring. Mol Ecol, 12, pp. 2817–2833.

William AF, P Young, PA Wheeler, and GH Thorgaard. 2005. An AFLP-based approach for the identification of sex-linked markers in rainbow trout (*Oncorhynchus mykiss*). Aquaculture, 247, pp. 35–43.

Williams JGK, AR Kubelik, KJ Livak, JA Rafalski, and SV Tingey. 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucl Acids Res, 18, pp. 6531–6535.

Wilson K, Y Li, V Whan, S Lehnert, K Byrne, S Moore, S Pongsomboon, A Tassanakajon, G Rosenberg, E Ballment, et al. 2002. Genetic mapping of the black tiger shrimp *Penaeus monodon* with amplified fragment length polymorphism. Aquaculture, 204, pp. 297–309.

Woram RA, C McGowan, JA Stout, K Gharbi, MM Ferguson, B Hoyheim, EA Davidson, WS Davidson, C Rexroad, and RG Danzmann. 2004. A genetic linkage map for Arctic char (*Salvelinus alpinus*): evidence for higher recombination rates and segregation distortion in hybrid versus pure strain mapping parents. Genome, 47, pp. 304–315.

Young WP, CO Ostberg, P Keim, and GH Thorgaard. 2001. Genetic characterization of hybridization and introgression between anadromous rainbow trout (*Oncorhynchus mykiss irideus*) and coastal cutthroat trout (*O. clarki clarki*). Mol Ecol, 10, pp. 921–930.

Young WP, PA Wheeler, VH Coryell, P Keim, and GH Thorgaard. 1998. A detailed linkage map of rainbow trout produced using doubled haploids. Genetics, 148, pp. 839–850.

Yu ZN and XM Guo. 2003. Genetic linkage map of the eastern oyster *Crassostrea virginica*. Gmelin Biol Bull, 204, pp. 327–338.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 5
Microsatellite Markers and Assessment of Marker Utility

Zhanjiang Liu

Microsatellites consist of multiple copies of tandemly arranged simple sequence repeats (SSR) that range in size from 1–6 base pairs (bp) (e.g., AC, CCA, or GATA) (Tautz 1989). Based on the repeat composition, microsatellites are often classified into mononucleotide microsatellites, dinucleotide microsatellites, trinucleotide microsatellites, tetranucleotide microsatellites, and so on. Microsatellites containing only one type of repeats are called simple microsatellites; microsatellites containing more than one type of repeats are called composite microsatellites. For instance, (CA)15 is a simple microsatellite, but (CA)8(CG)12 is a composite microsatellite. The advantages of microsatellites as molecular markers include its abundance in genomes, even distribution, small locus size facilitating polymerase chain reaction (PCR)-based genotyping, codominant nature of Mendelian inheritance, and high levels of polymorphism.

Abundance of Microsatellites

Microsatellites are highly abundant in various eukaryotic genomes including all aquaculture species studied to date. In most of the vertebrate genomes, microsatellites make up a few percent of the genomes in terms of the involved base pairs, depending on the compactness of the genomes. Generally speaking, more compact genomes tend to contain a smaller proportion of repeats including SSRs, but this generality is not always true. For example, the highly compact genome of the Japanese puffer fish contains 1.29% of microsatellites, but its closely related *Tetraodon nigroviridis* genome contains 3.21% of microsatellites (Grollius et al. 2000). During a genomic sequencing survey of channel catfish, microsatellites were found to represent 2.58% of the catfish genome based on the 11.4 million bps sequenced (Xu et al. 2006). A total of 4,262 microsatellites were found within 11.4 million bps (i.e., one microsatellite exists per 2.67 kilobases [kb] of channel catfish sequences). In fugu, one microsatellite was found every 1.87 kb of DNA. For comparison, in the human genome, one microsatellite was found every 6 kb of DNA (Beckmann and Weber 1992). It is reasonable to predict that in most aquaculture fish species, one microsatellite should exist every 10 kb or less of the genomic sequences, on average.

Dinucleotide repeats are the most abundant forms of microsatellites. For instance, in channel catfish, 67.9% of all microsatellites are present in the form of dinucleotide repeats; 18.5% present as trinucleotide repeats; and 13.5% as tetranucleotide repeats, excluding mononucleotide repeats, which are not nearly useful enough for molecular markers. Generally speaking, dinucleotide microsatellites are the most abundant, followed by tri- or tetra-nucleotide repeats, but in some cases,

tetranucleotide repeats can be more frequent than the trinucleotide repeats. For instance, in the genome of Japanese puffer fish *Takifugu rubripes*, dinucleotide repeats have the highest relative frequency (34%) followed by tetranucleotides (21%), trinucleotides (19%), mononucleotides (16.5%), hexanucleotides (6%), and pentanucleotides (3%) (Edwards et al. 1998).

Of the dinucleotide repeat types, $(CA)_n$, is the most common dinucleotide repeat type, followed by $(AT)_n$, and then $(CT)_n$ (Toth et al. 2000, Xu et al. 2006). $(CG)_n$ type of repeats are relatively rare in the vertebrate genomes. Partially this is because the vertebrate genomes are often A/T-rich. Of the trinucleotide repeats and tetranucleotide repeats, relatively A/T-rich repeat types are generally more abundant than G/C-rich repeat types. For instance, in channel catfish, the most abundant trinucleotide repeat is AAT, followed by AAC, ATC, and ATG. Microsatellites longer than tetranucleotide repeats (penta- and hexanucluotides) are much less abundant and therefore, are less important as molecular markers (Toth et al. 2000). It is important to point out that the definition of microsatellites limiting to repeats of six bases long are quite arbitrary. Technically speaking, repeats with seven bases or longer sequences are also microsatellites, but because they become rarer as the repeats are longer, they are less relevant as molecular markers.

Genomic Distribution of Microsatellites

Microsatellites tend to be, relatively speaking, evenly distributed in the genome on all chromosomes and all regions of the chromosome. They have been found inside gene coding regions (e.g., Liu et al. 2001), introns, and in the nongene sequences (Toth et al. 2000). The best known examples of microsatellites within coding regions are those causing genetic diseases in humans, such as the CAG repeats that encode the polyglutamine tract, resulting in mental retardation. In spite of their wide distribution in genes, microsatellites are predominantly located in noncoding regions (Metzgar et al. 2000). Only about 10–15% of microsatellites reside within coding regions (Moran 1993, van Lith and van Zutphen 1996, Edwards et al. 1998, Serapion et al. 2004). This distribution should be explained by negative selection against frame shift mutations in the translated sequences (Metzgar et al. 2000, Li et al. 2004). Because the majority of microsatellites exist in the form of dinucleotide repeats, any mutation by expansion or shrinking would cause frame shift of the protein encoding open frames if they reside within the coding region. That also explains why the majority of microsatellites residing within coding regions have been found to be trinucleotide repeats, though the presence of dinucleotide repeats and their mutations within the coding regions do occur.

Locus Size of Microsatellites

Most microsatellite loci are relatively small, ranging from a few to a few hundred repeats. The relatively small size of microsatellite loci is important for PCR-facilitated genotyping. Generally speaking, within a certain range, microsatellites containing a larger number of repeats tend to be more polymorphic, though polymorphism has been observed in microsatellites with as few as five repeats (Karsi et al. 2002).

For practical applications, microsatellite loci must be amplified using PCR. For best separations of related alleles that often differ from one another by as few as one repeat unit, it is desirable to have small PCR amplicons, most often within 200 bp. However, due to the repetitive nature of microsatellites, their flanking sequences can be a quite simple sequence, prohibiting design of PCR primers for the amplification of microsatellite loci within a small size limit. Consideration should be given regarding whether the quality of PCR primers can be sacrificed to a certain extent to reduce the amplicon size.

Polymorphism of Microsatellites

Microsatellites are highly polymorphic as a result of their hypermutability and thereby cause the accumulation of various forms in the population of a given species. Microsatellite polymorphism is based on size differences due to varying numbers of repeat units contained by alleles at a given locus (Figure 5.1). Microsatellite mutation rates have been reported as high as 10^{-2} per generation (Weber and Wong 1993, Crawford and Cuthbertson 1996, Ellegren 2000), which are several orders of magnitude greater than that of nonrepetitive DNA (10^{9}) (Li 1997).

Two mechanisms have been suggested to explain the hypermutability of microsatellites. (For a recent review, see Chistiakov et al. 2006.) The first involves polymerase slippage during DNA replication, resulting in differences in the number of repeat units (Levinson and Gutman 1987, Tautz et al. 1989). Transient dissociation of the replicating DNA strands followed by subsequent reassociation (Schlötterer et al. 1991, Richards and Sutherland 1994) would lead to slippage of the two strands, leading to the change of repeat numbers in the newly replicated DNA. Direct studies of human families have shown that new microsatellite mutations usually differed from the parental allele by only one or two repeats (Weber and Wong 1993), favoring a stepwise mutation model (see review by Estoup and Cornuet 1999) due to polymerase slippage. Microsatellite stability is controlled at multiple steps *in vivo* through the DNA mismatch repair (MMR) system, as shown for *Escherichia coli*, yeast, and humans (Sia et al. 1997). MMR proteins are found in a wide variety of taxa and are responsible for the correction of replication mistakes and suppression of the recombination between diverged sequences (Kolodner and Marsischky 1999). If the MMR system is defective, coding sequences with tandem repeats become subject to mutations, for example in tumor tissues (Sia et al. 1997). High-frequency microsatellite instability, therefore, plays a pivotal role in carcinogenesis (Atkin 2001). Both minor and major MMR genes contain short $(A)_n$ tracts in their coding regions, which are highly vulnerable to spontaneous deletion or insertion mutations, that could result in the inactivation of the MMR gene and hence cause MMR deficiency (Chang et al. 2001).

The second mechanism involves nonreciprocal recombination within the SSRs, leading to production of significantly larger and smaller alleles (Jakupciak and Wells 2000). In a few fish species, we have observed alleles with very large differences in repeat numbers, predictive of a infinite allele model (Balloux and Lugon-Moulin 2002), as well as alleles with differences of just one repeat unit, characteristic of the stepwise mutation model. Regardless of specific mechanisms, changes in numbers of repeat units can result in a large number of alleles at each microsatellite locus in a population. For evolutionary studies in populations, however, most often the stepwise

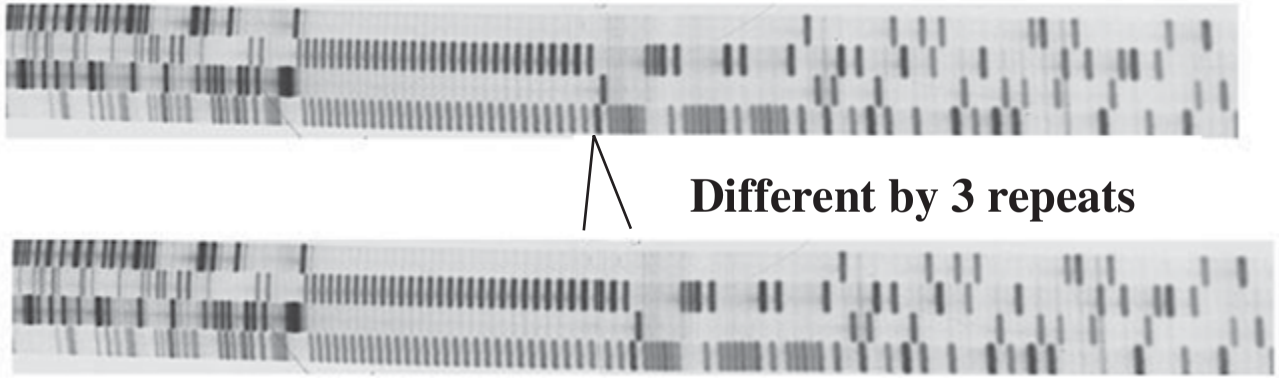


Figure 5.1. Microsatellite polymorphism is caused by the difference in repeat numbers. Note that the sequence of the lower fish harbors three more repeat units leading to a length difference of 6 bp. Figure was adopted from Liu and Cordes (2004b).

mutation model is assumed, right or wrong, as individuals with a similar number of repeats are regarded to be more closely related than those with a large difference in their numbers of repeats.

Inheritance of Microsatellites

Microsatellites are inherited in a Mendelian fashion as codominant markers. This is one of the strengths of microsatellite markers in addition to their abundance, even genomic distribution, small locus size, and high polymorphism.

Genotyping of microsatellite markers is usually straightforward. However, due to the presence of null alleles (alleles that cannot be amplified using the primers designed), complications do exist. As a result, caution should be exercised to assure the patterns of microsatellite genotypes fit the genetic model under application (Figure 5.2).

The disadvantage of microsatellites as markers include the requirement for existing molecular genetic information, the large amount of up front work for microsatellite development, and the tedious and labor intensive nature of microsatellite primer design, testing, and optimization of PCR conditions. Each microsatellite locus has to be identified and its flanking region sequenced for the design of PCR primers. Due to polymerase slippage during replication, small size differences between alleles of a given microsatellite locus (as little as 2 bp in a locus comprised of dinucleotide repeats) are possible. Because of this, PCR-amplified microsatellite DNA was traditionally labeled radioactively, separated on a sequencing gel, and then exposed on X-ray film overnight (Sambrook et al. 1989). Significant increases in the number of samples that can be typed in a day have been achieved by using automated fluorescent sequencers coupled with computer imaging systems (O'Reilly and Wright 1995).

The large number of alleles per locus results in the highest heterozygosity and polymorphic information content (PIC) values of any DNA markers. Microsatellites have recently become an extremely popular marker type in a wide variety of genetic investigations, as evidenced by the recent debut of the journal *Molecular Ecology Notes*, dedicated almost entirely to publishing primer and allele frequency data for newly characterized microsatellite loci in a wide range of species. Over the past decade, microsatellite markers have been used extensively in fishery research including studies of genome mapping, parentage, kinships, and stock structure (see O'Connell and Wright 1997 for a review). A cursory online literature search produced more than 500 entries since 1998 involving the use of microsatellites in such studies.

Development of Microsatellite Markers

Technically, the simplest way to identify and characterize a large number of microsatellites is through the construction of microsatellite-enriched small-insert genomic libraries (Orstrander et al. 1992, Lyall et al. 1993, Kijas et al. 1994, Zane et al. 2002). In spite of the variation in techniques for the construction of microsatellite-enriched libraries, the enrichment techniques usually include selective hybridization of fragmented genomic DNA with a tandem repeat-containing oligonucleotide probe and further PCR amplification of the hybridization products. Libraries highly

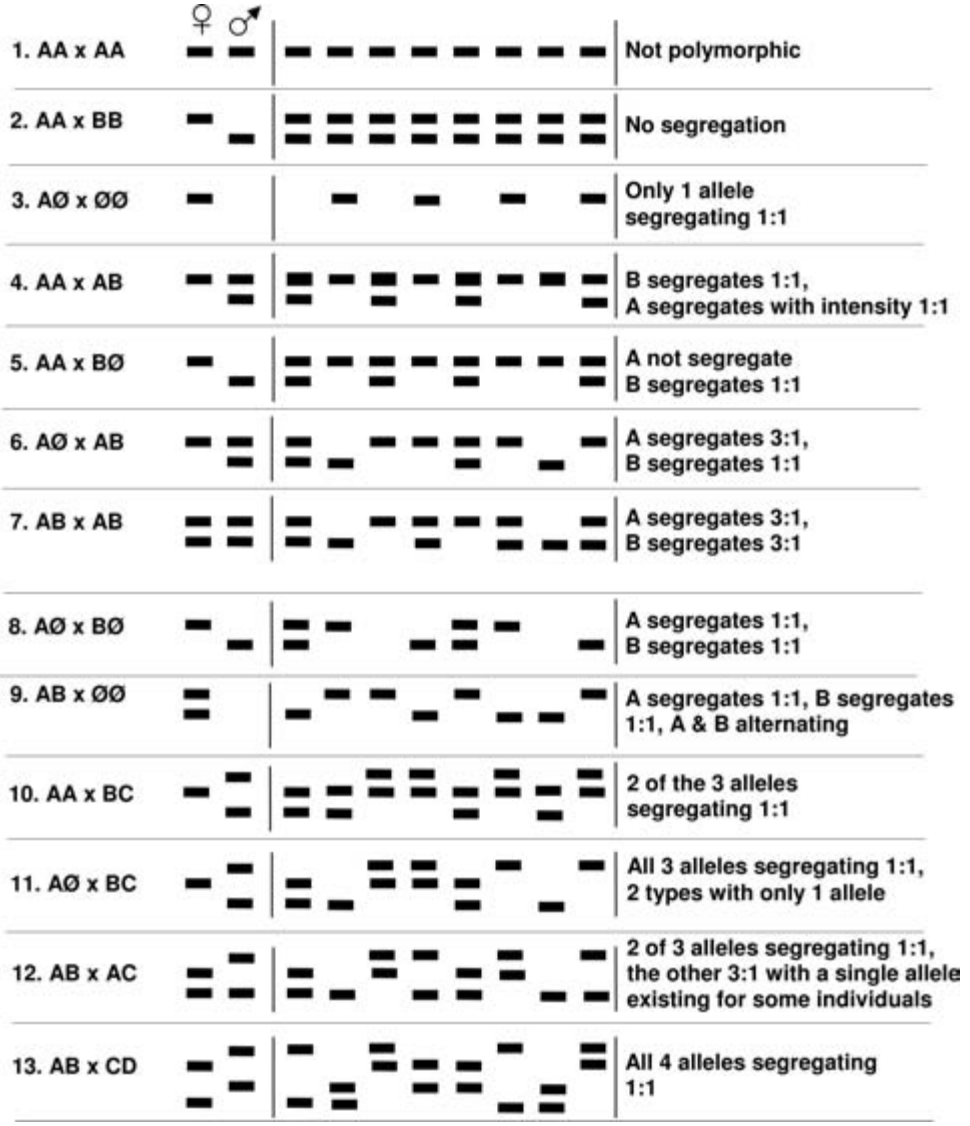


Figure 5.2. Various segregation patterns and the predicted genotypes.

enriched by tandem repeats have been constructed for many organisms, including fish such as *Lophius* sp. (Garoia et al. 2003), gilthead sea bream *Sparus aurata* (Zane et al. 2002), and Nile tilapia (Carleton et al. 2002).

In spite of the simplicity in the construction of microsatellite-enriched libraries and thereby the identification and characterization of microsatellite markers, for large genome projects, the real need of direct microsatellite marker development may not be the wisest approach. This is because microsatellites derived from enriched libraries most often are from anonymous genomic regions and thus are Type II markers. O'Brien (1991) divided molecular markers into Type I markers associated with genes of known functions and Type II markers associated with anonymous genomic

sequences. Microsatellites usually represent Type II markers, since by chance more than 90% of the genome are noncoding regions. Type II markers most often cannot be used for comparative genome studies across a wide spectrum of species. More importantly, microsatellites can be identified and sequenced directly from genome sequence surveys such as BAC end sequencing (see Chapter 15), and from EST analysis from which many microsatellites can be developed into Type I markers (Liu et al. 1999, Serapion et al. 2004).

Type I markers are more difficult to develop (Liu et al. 1999b, Liu and Cordes 2004a). Although nongene sequences are free to mutate, causing higher levels of polymorphism, sequences within protein-coding regions generally show lower levels of polymorphism because of functional restraints. The most effective and rapid way for producing Type I microsatellites is the sequencing of clones from cDNA libraries. Both 5'- and 3'-ends of cDNA clones can be sequenced to produce ESTs (see Chapter 20). Such collections provide a robust sequence resource that can be used for gene discovery and annotation, as well as for the identification of gene-associated microsatellite markers for comparative genetics (Liu 2003, Serapion et al. 2004).

Microsatellites can be searched in EST sequence databases. However, the prerequirement for the development of Type I microsatellites is the access to sufficient sequence information. For instance, in the channel catfish, generation of 45,000 EST sequences allowed the identification of thousands of microsatellites (Serapion et al. 2004). Sequence analysis of 1,909 ESTs from a skin cDNA library of *Ictalurus punctatus* revealed the presence of 89 (4.7% of 1,909) microsatellite-containing genes (Karsi et al. 2002). Screening of 1,201 ESTs from a channel catfish brain cDNA library yielded 88 (7.3%) clones with microsatellites (Ju et al. 2000). It is therefore, reasonable to argue that one does not have to spend resources to intentionally develop microsatellite markers that not only take time and effort but are mostly Type II markers. Instead, identification of microsatellite markers from EST resources should generate mostly Type I markers. Likewise, many microsatellites can be identified from BAC end sequences that fulfill two duties with one action (Xu et al. 2006).

Caution has to be exercised, however, on microsatellites developed from ESTs. First, due to the presence of introns, one has to be careful not to design primers at the exon-intron boundaries. Second, the presence of introns would make allele sizes unpredictable. Finally, many microsatellites exist at the 5'- or 3'-UTR, making flanking sequences not sufficient for the design of PCR primers. Although introns are not a problem for microsatellites derived from BAC end sequencing, sequencing reactions often terminate immediately after the microsatellite repeats, also making flanking sequences not sufficient for the design of PCR primers. Additional sequencing would be required for the development of microsatellite markers from these sequences with SSRs at the ends.

Applications of Microsatellite Markers

Microsatellite markers, one of the many types of molecular markers, are ideal for many types of applications. Microsatellites are the most useful type of molecular markers for genetic linkage mapping, and they are anticipated to remain as the markers of choice for the construction of linkage maps, especially for framework linkage maps. The fundamental reason for this is because of their high polymorphism, high abundance, small

Table 5.1. Some examples of linkage maps constructed with microsatellite markers in aquaculture species.

Species	Common name	References
Salmo trutta	Brown trout	Gharbi et al. 2006
Dicentrarchus labrax	European sea bass	Chistiakov et al. 2005
Oreochromis spp.	Tilapia	Kocher et al. 1998, Agresti et al. 2000, McConnell et al. 2000, Lee et al. 2005
Plecoglossus altivelis	Ayu	Watanabe et al. 2004
Xiphophorus		Walter et al. 2004
Salvelinus alpinus	Arctic char	Woram et al. 2004
Salmo salar	Atlantic salmon	Gilbey et al. 2004, Moen et al. 2004
Oncorhynchus mykiss	Rainbow trout	Sakamoto et al. 2000, Nichols et al. 2003
Ictalurus punctatus	Channel catfish	Waldbieser et al. 2001
Danio rerio	Zebrafish	Knapik et al. 1998, Shimoda et al. 1999
Crassostrea gigas	Pacific oyster	Hubert and Hedgecock 2004
Crassostrea virginica	Eastern oyster	Yu and Guo 2003
Seriola quinqueradiata and *Seriola lalandi*	Yellowtails	Ohara et al. 2005
Cyprinus carpio	Common carp	Sun and Liang 2004
Paralichthys olivaceus	Japanese flounder	Coimbra et al. 2005

locus size, and codominance of inheritance, so all four alleles from a pair of parents can be observed from the gel. In addition, because microsatellite markers are sequence-tagged markers, integration with a physical map is possible.

Some examples of genetic linkage maps constructed using microsatellites are listed in Table 5.1. An excellent review was just published by Chistiakov and others (2005) in which various applications of microsatellites were discussed; interested readers are referred to this review. In Chapter 8, use of microsatellite markers for stock analysis and parentage analysis is described. Chapter 9 describes methodologies involving the use of microsatellites and other markers for population analysis. The use of microsatellite markers for the construction of genetic linkage maps is the subject of Chapter 10. QTL mapping involving the use of microsatellite markers are described in Chapter 11.

Assessing the Utility of Markers

Mostly, the number of alleles and allele frequencies measure informativeness of a genetic marker. Specific parameters have been developed through the course of marker development and application in the last two decades. These parameters include two measures of marker informativeness: heterozygosity and polymorphic information content. In addition, the efficiency of marker systems can be evaluated based on the mean number of polymorphic genetic markers per assay. Genotyping errors and multiplex ratio are also used to measure the utility of molecular markers.

Genotyping errors reflect the reproducibility of the marker assay and clarity of the marker genotypes. Multiplex ratio refers to the number of simultaneously assayed loci. In the case of microsatellites, multiplex ratio indicate the number of loci for which PCR primers are compatible for multiplex PCR, and their products are distinct allowing simultaneous analysis on the same gel run.

Heterozygosity (H)

Heterozygosity is a widely used measure of marker informativeness. The informativeness of a genetic marker increases as *H* increases. The heterozygosity of a genetic marker is estimated by the number of alleles and their relative frequencies. Heterozygosity is a function of the individuals and populations sampled. When individuals are sampled from genetically narrow or genetically isolated populations, fewer alleles and a higher frequency of monomorphic loci can be expected than when individuals are sampled from genetically diverse populations. Heterozygosity is defined with the following formula:

$$H = 1 - \sum_{i=1}^{k} Pi^2 \tag{5.1}$$

where P_i is the frequency of the *i*th allele and *k* is the number of alleles (Nei and Roychoudhury 1974, Ott 1992).

Heterozygosity, as its name suggests, is an estimate of probability that a randomly sampled individual is heterozygous when the individuals are sampled from outbred populations. For instance, if a microsatellite has three alleles in the population with a frequency of 10%, 30%, and 60%, H value for this population would be $H = 1-(0.1^2 + 0.3^2 + 0.6^2) = 0.54$. In other words, assuming random mating, the probability of finding a random individual in this population to be heterozygous is 54%.

From the above formula, it is obvious that when the number of alleles is given, the more equal the distribution of all alleles, the greater the H value is; when the allele frequencies are given, the greater the number of alleles, the greater the H value is. For instance, if a population contains two alleles of a microsatellite locus with 10% and 90% frequency each for the two alleles, H value for this microsatellite locus is 0.18. For this microsatellite locus, if the allele frequency of the two alleles is 50% each, then the H value becomes 0.5. If the population has two alleles, then each of the two alleles has a frequency of 50% and the H value would be 0.5. If the population has 10 alleles with 10% each, then the H value would be 0.9.

Polymorphic Information Content

Another measure of the marker informativeness in outbred species is the polymorphic information content (PIC) (Botstein et al. 1980). PIC is defined by the following formula:

$$PIC = 1 - \sum_{i=1}^{k} pi^2 - \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} 2Pi^2 Pj^2 \tag{5.2}$$

In the above example, the population has three alleles with allele frequency of 10%, 30%, and 60%. The PIC value should be calculated as:

$$PIC = 1 - [(0.1)^2 + (0.3)^2 + (0.6)^2] - [2 \times (0.1)^2 \times (0.3)^2 + 2 \times (0.1)^2 \times (0.6)^2 + 2 \times (0.3)^2 \times (0.6)^2] = 0.4662 \quad (5.3)$$

Because of the complexities in deriving PIC, for practical purposes in most cases, H values are used rather than PIC. PIC is always slightly smaller than H, but in most cases is close enough to H so it would not make a major difference. Interested readers are referred to Botstein and others (1980) for more information on PIC.

Mean Heterozygosity

Mean heterozygosity is a useful measurement when dealing with multiple markers or marker types. The mean heterozygosity of n genetic markers is

$$H_{av} = 1 - \sum_{j=1}^{n} H_j / n \quad (5.4)$$

where H_j is the heterozygosity of the jth genetic marker. H_{av} can be estimated with or without monomorphic markers (depending on the context of the analysis). To distinguish between the two cases, let H_{pav} be the mean heterozygosity estimated from polymorphic markers only and H_{Tav} be the mean heterozygosity estimated from polymorphic and monomorphic markers. Powell and others (1996) compared the informativeness and multiplex ratios of RFLP, AFLP, RAPD, and microsatellite markers in soybeans and proposed estimating the mean heterozygosity of genetic markers by summing over monomorphic and polymorphic genetic markers when comparing markers with different multiplex ratios:

$$H_{Tav} = p \sum_{j=1}^{n_p} H_j / n_p \quad (5.5)$$

where H_j is the heterozygosity of the jth polymorphic genetic marker, n_p is the number of polymorphic genetic markers, and p is the percentage of polymorphic markers (number of polymorphic markers over the sum of the polymorphic and monomorphic markers).

The Mean Number of Polymorphic Genetic Markers per Assay

The mean number of polymorphic genetic markers per assay is the product of the mean number of bands per assay and mean heterozygosity:

$$P_{av} = mH_T \quad (5.6)$$

where m is the mean number of bands per assay, and HT is the average heterozygosity. This measurement takes both the heterozygosity and the efficiency of the markers into consideration. For instance, AFLP is much more efficient in producing polymorphic bands than RFLP and it has a greater mean number of polymorphic genetic markers per assay. For example, 10 AFLP primer combinations produce a total of 1,000 bands (100 per AFLP assay on average) and 10 RFLP marker assays produce a total of 12 bands (8 RFLP probes produce 1 band each, and the other 2 RFLP probes produce 2 bands each). If $H_T = 0.2$ for the AFLP markers and $H_T = 0.52$ for the RFLP markers (clearly in this case even though the heterozygosity is lower with AFLP markers than RFLP markers for individual markers), then P_{av} for the AFLP markers is 20, whereas that for the RFLP markers is only 0.624. Thus, the AFLP markers are $20/0.624 = 32.05$ times more informative than the RFLP markers on a per assay basis. Multilocus fingerprinting techniques such as RAPD and AFLP should have a higher mean number of polymorphic genetic markers per assay than single locus marker systems such as RFLP and microsatellites.

Conclusion

Microsatellite markers have been and likely will remain the marker type of choice for genome research because of their high polymorphism, high abundance, small locus size, even genomic distribution, and codominance of inheritance. Microsatellites have the highest heterozygosity among all marker types because of their high number of alleles. Because most RFLP and SNP markers are regarded as biallelic markers, they have a maximal heterozygosity value of 0.5 (when the two alleles have equal allele frequencies). RAPD and AFLP are both biallelic dominant markers, and they can have a maximal heterozygosity of 0.5 as well. Thus, microsatellites are most informative as genetic markers. This feature makes microsatellites the unique marker system for identification of individuals such as parentage analysis, as shown in Chapter 8, as well as the choice of markers for many other types of applications.

The major application of microsatellite markers is for the construction of genetic linkage and QTL maps. This is also because of the high polymorphic rate of microsatellite markers. When a resource family is produced, the male and female fish parents are likely heterozygous in most microsatellite loci. The high polymorphism of microsatellites makes it possible to map many markers using a minimal number of resource families.

There are other reasons for the popularity of microsatellites. One of these is because microsatellites are sequence-tagged markers that allow them to be used as probes for the integration of different maps including genetic linkage and physical maps. Communication using microsatellite markers across laboratories is easy, and use of microsatellite across species borders is sometimes possible if the flanking sequences are conserved (FitzSimmons et al. 1995, Rico et al. 1996, Leclerc et al. 2000, Cairney et al. 2000). As a result, microsatellites can be used also for comparative genome analysis. If microsatellites can be tagged to gene sequences, their potential for use in comparative mapping is greatly enhanced.

Development of microsatellite markers has traditionally been conducted by the development of microsatellite-enriched DNA libraries. However, this may not be the

most optimal situation for genome research. In most cases, EST and BAC end sequence resources are needed earlier or later. Therefore, direct investment into resource development involving EST and BAC end sequencing may prove to be very effective.

References

Agresti J, S Seki, A Cnaani, S Poompuang, EM Hallerman, N Umiel, G Hulata, GAE Gall, and B May. 2000. Breeding new strains of tilapia: development of an artificial center of origin and linkage map based on AFLP and microsatellite loci. Aquaculture, 185, pp. 43–56.

Atkin NB. 2001. Microsatellite instability. Cytogenet Cell Genet, 92, pp. 177–181.

Balloux F and N Lugon-Moulin. 2002. The estimation of population differentiation with microsatellite markers. Mol Ecol, 11, pp. 155–165.

Beckman JS and JL Weber. 1992. Survey of human and rat microsatellites. Genomics, 12, pp. 627–631.

Botstein D, RL White, M Skolnick, and RW Davis. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet, 32, pp. 314–331.

Cairney M, JB Taggart, and B Hoyheim. 2000. Characterization of microsatellite and minisatellite loci in Atlantic salmon (*Salmo salar* L.) and cross-species amplification in other salmonids. Mol Ecol, 9, pp. 2175–2178.

Carleton KL, JT Streelman, BY Lee, N Garnhart, M Kidd, and TD Kocher. 2002. Rapid isolation of CA microsatellites from the tilapia genome. Anim Genet, 33, pp. 140–144.

Chang DK, D Metzgar, C Wills, and CR Boland. 2001. Microsatellites in the eukaryotic DNA mismatch repair genes as modulators of evolutionary mutation rate. Genome Res, 11, pp. 1145–1146.

Chistiakov DA, B Hellemans, CS Haley, AS Law, CS Tsigenopoulos, G Kotoulas, D Bertotto, A Libertini, and FA Volckaert. 2005. A microsatellite linkage map of the European sea bass *Dicentrarchus labrax* L. Genetics, 170, pp. 1821–1826.

Coimbra MRM, K Kobayashi, S Koretsugu, O Hasegawa, E Ohara, A Ozaki, T Sakamoto, K Naruse, and N Okamoto. 2005. A genetic linkage map of the Japanese flounder, *Paralichthys olivaceus*. Aquaculture, 220, pp. 203–218.

Crawford AM and RP Cuthbertson. 1996. Mutations in sheep microsatellites. Genome Res, 6, pp. 876–879.

Crollius HR, O Jaillon, C Dasilva, C Ozouf-Costaz, C Fizames, C Fischer, L Bouneau, A Billault, F Quetier, W Saurin, A Bernot, and J Weissenbach. 2000. Characterization and repeat analysis of the compact genome of the freshwater pufferfish Tetraodon nigroviridis. Genome Res, 10, pp. 939–949.

Edwards YJ, G Elgar, MS Clark, and MJ Bishop. 1998. The identification and characterization of microsatellites in the compact genome of the Japanese pufferfish, *Fugu rubripes*: perspectives in functional and comparative genomic analyses. J Mol Biol, 278, pp. 843–854.

Ellegren H. 2000. Microsatellite mutations in the germline: implications for evolutionary inference, Trends Genet, 16, pp. 551–558.

Estoup A and J Cornuet. 1999. Microsatellite evolution: inferences from population data, pp. 49–65 In *Microsatellites: Evolution and Applications*, edited by DB Goldstein and C Schlotterer. Oxford University Press, New York.

FitzSimmons NN, C Moritz, and SS Moore. 1995. Conservation and dynamics of microsatellite loci over 300 million years of marine turtle evolution. Mol Biol Evol, 12, pp. 432–440.

Garoia F, I Guarniero, and F Tinti. 2003. Polymorphic microsatellites for the Mediterranean angler species (*Lophiidae*). Mol Ecol Notes, 3, pp. 294–296.

Gharbi K, A Gautier, RG Danzmann, S Gharbi, T Sakamoto, B Hoyheim, JB Taggart, M Cairney, R Powell, F Krieg, N Okamoto, MM Ferguson, LE Holm, and R Guyomard. 2006.

A linkage map for brown trout (Salmo trutta): chromosome homeologies and comparative genome organization with other salmonid fish. Genetics, 172, pp. 2405–2419.
Gilbey J, E Verspoor, A McLay, D Houlihan. 2004. A microsatellite linkage map for Atlantic salmon (*Salmo salar*). Anim Genet, 35, pp. 98–105.
Hubert S and D Hedgecock. 2004. Linkage maps of microsatellite DNA markers for the Pacific oyster *Crassostrea gigas*. Genetics, 168, pp. 351–62.
Jakupciak JP and RD Wells. 2000. Genetic instabilities of triplet repeat sequences by recombination. IUBMB Life, 50, pp. 355–359.
Ju Z, A Karsi, A Kocabas, A Patterson, P Li, D Cao, R Dunham, and ZJ Liu. 2000. Transcriptome analysis of channel catfish (*Ictalurus punctatus*): genes and expression profile from the brain. Gene, 261, pp. 373–382.
Karsi A, D Cao, P Li, A Patterson, A Kocabas, J Feng, Z Ju, K Mickett, and ZJ Liu. 2002. Transcriptome analysis of channel catfish (*Ictalurus punctatus*): Initial analysis of gene expression and microsatellite-containing cDNAs in the skin. Gene, 285, pp. 157–168.
Kijas JM, JC Fowler, CA Garbett, MR Thomas. 1994. Enrichment of microsatellites from the citrus genome using biotinylated oligonucleotide sequences bound to streptavidin-coated magnetic particles. Biotechniques, 16, pp. 656–660.
Knapik EW, A Goodman, M Ekker, M Chevrette, J Delgado, S Neuhauss, N Shimoda, W Driever, MC Fishman, HJ Jacob. 1998. A microsatellite genetic linkage map for zebrafish (*Danio rerio*). Nat Genet, 18, pp. 338–343.
Kocher TD, WJ Lee, H Sobolewska, D Penman, B McAndrew. 1998. A genetic linkage map of a cichlid fish, the tilapia (*Oreochromis niloticus*). Genetics, 148, pp. 1225–1232.
Kolodner RD, GT Marsischky. 1999. Eukaryotic DNA mismatch repair. Curr Opin Genet Dev, 9, pp. 89–96.
Leclerc D, T Wirth, and L Bernatchez. 2000. Isolation and characterization of microsatellite loci in the yellow perch (*Perca flavescens*), and cross-species amplification within the family Percidae. Mol Ecol, 9, pp. 995–997.
Lee BY, WJ Lee, JT Streelman, KL Carleton, AE Howe, G Hulata, A Slettan, JE Stern, Y Terai, and TD Kocher. 2005. A second-generation genetic linkage map of tilapia (*Oreochromis spp.*). Genetics, 170, pp 237–244.
Levinson G and GA Gutman. 1987. High frequency of short frameshifts in poly-CA/GT tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. Nucl Acids Res, 15, pp. 5323–5338.
Li W-H. 1997. *Molecular Evolution.* Sinauer Associates, Sunderland, MA, pp. 177–213.
Liu ZJ. 2003. A review of catfish genomics: progress and perspectives. Comp Funct Genomics, 4, pp. 259–265.
Liu ZJ and JF Cordes. 2004a. DNA marker technologies and their applications in aquaculture genetics. Aquaculture, 238, pp. 1–37.
Liu ZJ and JF Cordes. 2004b. Erratum to DNA marker technologies and their applications in aquaculture genetics. Aquaculture, 242, pp. 735–736.
Liu ZJ, A Karsi, and RA Dunham. 1999. Development of polymorphic EST markers suitable for genetic linkage mapping of catfish. Mar Biotechnol, 1, pp. 437–447.
Liu ZJ, P Li, A Kocabas, Z Ju, A Karsi, D Cao, and A Patterson 2001. Microsatellite-containing genes from the channel catfish brain: evidence of trinucleotide repeat expansion in the coding region of nucleotide excision repair gene RAD23B. Biochem Biophys Res Comm, 289, pp. 317–324.
Lyall JE, GM Brown, RA Furlong, MA Ferguson-Smith, and NA Affara. 1993. A method for creating chromosome-specific plasmid libraries enriched in clones containing [CA]n microsatellite repeat sequences directly from flow-sorted chromosomes. Nucleic Acids Res. 21, pp. 4641–4642.
McConnell SK, C Beynon, J Leamon, and DO Skibinski. 2000. Microsatellite marker based genetic linkage maps of *Oreochromis aureus* and *O. niloticus* (Cichlidae): extensive linkage group segment homologies revealed. Anim Genet, 31, pp. 214–218.

Metzgar D, J Bytof, and C Wills. 2000. Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res, 10, pp. 72–80.

Moen T, B Hoyheim, H Munck, and L Gomez-Raya. 2004. A linkage map of Atlantic salmon (*Salmo salar*) reveals an uncommonly large difference in recombination rate between the sexes. Anim Genet, 35, pp. 81–92.

Moran C. 1993. Microsatellite repeats in pig (*Sus domestica*) and chicken (*Gallus domesticus*) genomes. J Hered, 84, pp. 274–280.

Nei M and AK Roychoudhury. 1974. Sampling variances of heterozygosity and genetic distance. Genetics, 76, pp. 379–390.

Nichols KM, WP Young, RG Danzmann, BD Robison, C Rexroad, M Noakes, RB Phillips, P Bentzen, I Spies, K Knudsen, FW Allendorf, BM Cunningham, J Brunelli, H Zhang, S Ristow, R Drew, KH Brown, PA Wheeler, and GH Thorgaard. 2003. A consolidated linkage map for rainbow trout (*Oncorhynchus mykiss*). Anim Genet, 34, pp. 102–115.

O'Brien SJ. 1991. Molecular genome mapping lessons and prospects. Curr Opin Genet Dev, 1, pp. 105–111.

O'Connell M and JM Wright. 1997. Microsatellite DNA in fishes. Rev Fish Biol Fish, 7, pp. 331–363.

Ohara E, T Nishimura, Y Nagakura, T Sakamoto, K Mushiake, and N Okamoto. 2005. Genetic linkage maps of two yellowtails (*Seriola quinqueradiata* and *Seriola lalandi*). Aquaculture, 244, pp. 41–48.

O'Reilly P and JM Wright. 1995. The evolving technology of DNA fingerprinting and its application to fisheries and aquaculture. J Fish Biol, 47, pp. 29–55.

Ostrander EA, PM Jong, J Rine, and G Duyk. 1992. Construction of small-insert genomic DNA libraries highly enriched for microsatellite repeat sequences. Proc Natl Acad Sci, USA, 89, pp. 3419–3423.

Ott J. 1992. Strategies for characterizing highly polymorphic markers in human gene mapping. Am J Hum Genet, 51, pp. 283–290.

Powell W, M Morgante, C Andre, M Hanafey, J Vogel, S Tingey, and A Rafalski. 1996. The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. Molecular Breeding, 2, pp. 225–238.

Rico C, I Rico, and G Hewitt. 1996. 470 million years of conservation of microsatellite loci among fish species, Proc R Soc Lond B Biol Sci, 263, pp. 549–557.

Sakamoto T, RG Danzmann, K Gharbi, P Howard, A Ozaki, SK Khoo, RA Woram, N Okamoto, MM Ferguson, LE Holm, R Guyomard, and B Hoyheim. 2000. A microsatellite linkage map of rainbow trout (*Oncorhynchus mykiss*) characterized by large sex-specific differences in recombination rates. Genetics, 155, 1331–1345.

Sambrook J, EF Fritsch, and T Maniatis. 1989. Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Schlötterer C, B Amos, and D Tautz. 1991. Conservation of polymorphic simple sequence loci in cetacean species. Nature, 354, pp. 63–65.

Serapion J, H Kucuktas, J Feng, and Z Liu. 2004. Bioinformatic mining of type I microsatellites from expressed sequence tags of channel catfish (*Ictalurus punctatus*). Mar Biotechnol, 6, pp. 364–377.

Shimoda N, EW Knapik, J Ziniti, C Sim, E Yamada, S Kaplan, D Jackson, F de Sauvage, H Jacob, and MC Fishman. 1999. Zebrafish genetic map with 2,000 microsatellite markers. Genomics, 58, pp. 219–232.

Sia EA, RJ Kokoska, M Dominska, P Greenwell, TD Petes. 1997. Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes. Mol Cell Biol, 17, pp. 2851–2858.

Sun X and L Liang. 2004. A genetic linkage map of common carp (*Cyprinus carpio* L.) and mapping of a locus associated with cold tolerance. Aquaculture, 238, pp. 165–172.

Tautz, D. 1989. Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucl Acids Res, 17, pp. 6563–6571.

Toth G, Z Gaspari, and J Jurka. 2000. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res, 10, pp. 967–981.

Van Lith HA and LF Van Zutphen. 1996. Characterization of rabbit DNA microsatellites extracted from the EMBL nucleotide sequence database. Anim Genet, 27, pp. 387–395.

Waldbieser GC, BG Bosworth, DJ Nonneman, and WR Wolters. 2001. A microsatellite-based genetic linkage map for channel catfish, *Ictalurus punctatus*. Genetics, 158, pp. 727–734.

Walter RB, JD Rains, JE Russell, TM Guerra, C Daniels, DA Johnston, J Kumar, A Wheeler, K Kelnar, VA Khanolkar, EL Williams, JL Hornecker, L Hollek, MM Mamerow, A Pedroza, and S Kazianis. 2004. A microsatellite genetic linkage map for *Xiphophorus*. Genetics, 168, pp. 363–372.

Watanabe T, H Fujita, K Yamasaki, S Seki, and N Taniguchi. 2004. Preliminary study on linkage mapping based on microsatellite DNA and AFLP markers using homozygous clonal fish in ayu (*Plecoglossus altivelis*). Mar Biotechnol, 6, pp. 327–334.

Weber JL and C Wong. 1993. Mutation of human short tandem repeats. Hum Mol Genet, 2, pp. 1123–1128.

Woram RA, C McGowan, JA Stout, K Gharbi, MM Ferguson, B Hoyheim, EA Davidson, WS Davidson, C Rexroad, and RG Danzmann. 2004. A genetic linkage map for Arctic char (*Salvelinus alpinus*): evidence for higher recombination rates and segregation distortion in hybrid versus pure strain mapping parents. Genome, 47, pp. 304–315.

Xu P, S Wang, L Liu, E Peatman, B Somridhivej, J Thimmapuram, G Gong, and ZJ Liu. 2006. Generation of channel catfish BAC end sequences for marker development and assessment of syntenic conservation with other fish species. Anim Genet, 37, pp. 321–326.

Yu Z and X Guo. 2003. Genetic linkage map of the eastern oyster *Crassostrea virginica* Gmelin. Biol Bull, 204, pp. 327–338.

Zane L, L Bargelloni, and T Patarnello. 2002. Strategies for microsatellite isolation: a review. Mol Ecol, 11, pp. 1–16.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 6
Single Nucleotide Polymorphism (SNP)

Zhanjiang Liu

Single nucleotide polymorphism (SNP) describes polymorphisms caused by point mutations that give rise to different alleles containing alternative bases at a given nucleotide position within a locus. Such sequence differences due to base substitutions have been well characterized since the beginning of DNA sequencing in 1977, but the ability to genotype SNPs rapidly in large numbers of samples was not possible until several major technological advances in the late 1990s. With the development of the TaqMan technology, gene chip technology, pyrosequencing, and MALDI-TOF, which is matrix-assisted laser desorption ionization-time of flight mass spectrometry (Haff et al. 1997, Tost et al. 2005), SNPs are again becoming a focal point in molecular marker development because they are the most abundant polymorphism in any organism (as shown in Table 6.1), adaptable to automation, and reveal hidden polymorphism not detected with other markers and methods. SNP markers are regarded by many as the projected markers of choice for the future. In this chapter, I will summarize methods for SNP discovery, review the traditional approaches for SNP genotyping and their principles, present several major platforms for SNP genotyping using recently developed technologies, and discuss the pros and cons of SNP markers for aquaculture genome research.

What Are SNP Markers and Why Are They the Future Markers of Choice?

SNP can be defined as base variation at any site of the genome among individuals, or an alternative base at a given DNA site (Figure 6.1). Single nucleotide polymorphisms are by no means new. They were noticed ever since DNA was sequenced back in 1977. They were not used as markers for two decades because of the lack of efficient genotyping technologies. The differences between alleles of SNPs are very small; the length is the same, the only difference is one base substitution, most often from A to G, or from C to T. Separation of such subtle differences requires special technologies. After two decades of genetic analysis, there are several types of very efficient DNA markers. As pointed out by Vignal and others (2002) in their review on SNP markers, in terms of genetic information provided, as simple biallelic codominant markers, SNPs can be considered to be a step backward when compared to the highly informative multiallelic microsatellites (Middleton et al. 2004, John et al. 2004, Lin et al. 2005, Ma et al. 2005, Thalamuthu et al. 2005). Why are SNPs regarded as the choice marker system of the future? The major reasons include the recent need for very high densities of genetic markers for the studies of multifactorial diseases (Schaid et al. 2004, Kim et al. 2005, Wilcox et al. 2005, Xiang et al. 2005) and the recent progress in polymorphism detection and genotyping techniques. Because

Table 6.1. Estimation of SNP rates in various organisms.

Organism	Genome or genes studied	π Value	One SNP per DNA	References
Human	Genome	0.0008	1,250 bp	Sachidanandam et al. 2001
Mouse	Genome	0.0096	104 bp	Abe et al. 2004
Bos taurus and B. indicus	Amyloid	0.019	52.6 bp	Konfortov et al. 1999
Bos taurus and B. indicus	Leptin	0.0026	384 bp	Konfortov et al. 1999
Bos taurus	Amyloid	0.0096	104 bp	Konfortov et al. 1999
Bos taurus	Leptin	0.0023	434 bp	Konfortov et al. 1999
Bos taurus	Cytokine	0.0022	443 bp	Heaton et al. 2001
Chickens	31 kb survey	0.0044	225 bp	Schmid et al. 2000
Chickens	ESTs	0.00047	2,119 bp	Kim et al. 2003
Channel and blue catfish	161 ESTs	0.013	76 bp	He et al. 2003

Figure 6.1. SNPs are alternative bases at a given site of DNA.

SNPs represent the ultimate reason for differences among individuals, and their analyses are most adaptable to automation, they are once again becoming the most popular molecular markers (Lai et al. 2001, Rafalski et al. 2002).

SNP Discovery

In spite of its increasing popularity as the choice of markers for the future, SNP discovery is a daunting task because as stated in its definition, single nucleotide polymorphism discovery depends on sequencing. Several approaches have been used for the discovery of SNPs in humans and animals. Earlier efforts used approaches such as single-strand conformation polymorphism (SSCP) analysis (Gonen et al. 1999), heteroduplex analysis (Sorrentino et al. 1992), and direct DNA sequencing. However, several recently developed approaches provide greater efficiencies.

The first option and the simplest is to conduct direct sequencing of genomic polymerase chain reaction (PCR) products obtained from different individuals. However, two factors really limit the use of this strategy. First, this approach requires the use of locus-specific PCR primers, and when large numbers of loci are involved, this approach is costly. Second, accurate sequencing of PCR products for the discovery of SNPs can be a great challenge. Dealing with sequencing ambiguities while attempting to identify

SNPs is not an easy job. Sequencing artifact with double peaks cannot be distinguished from true heterozygotes. In addition, for many aquaculture species, sequence information is limited.

The second strategy involves data mining from Expressed Sequence Tag (EST) projects, if EST libraries were constructed using multiple individuals. This approach is realistic because EST resources already exist, or are to be developed for the majority of important aquaculture species. This approach is advantageous in that the SNPs are coming from genes and many of them can come from coding regions and, therefore, the SNPs discovered are Type I markers. Also the coding region SNPs would allow analysis of association of SNPs with traits for the discovery of the "causing SNPs" for the traits (Bader 2001, Marnellos et al. 2003, Halldorsson et al. 2004, Stram 2004). However, this approach has major limitations. Because of evolutionary restraint on mutations in coding regions, SNP rates are generally much lower in coding regions than in noncoding regions. In addition, in some rare cases, sequence variation in ESTs may not represent SNPs in the genome due to RNA editing.

The third approach involves data mining from genome sequencing projects. Sequence comparisons in overlapped bacterial artificial chromosome (BAC) clone regions can be used for the discovery of SNPs. This approach depends heavily on genetic background of the DNA used for the construction of BAC libraries. Obviously, only SNPs that reside within the overlapping BAC regions can be discovered. More importantly, this approach is not applicable for species without a whole-genome sequencing project, which is the case currently for almost all aquaculture species.

The fourth approach is called reduced representation shotgun sequencing (RRS) (Altshuler et al. 2000). This approach is based on the fact that genomic segments of the same origin with the same size will migrate to the same position in gel electrophoresis. In this approach, DNA from multiple individuals (and in the case of humans, many individuals from all ethnic groups) is mixed together, cleaved with a restriction enzyme, and separated on agarose gels. A subset of the genomic digest contained within a slice of the gel is cloned and subjected to sequencing. A 2–5 fold shotgun sequencing is conducted to generate overlapping sequences, allowing sequence alignment and SNP discovery.

Traditional Approaches for SNP Genotyping

Unlike microsatellites, for which genotyping is standard with PCR amplification and sizing, many approaches have been considered for SNP genotyping. Traditional methods available for SNP genotyping include direct sequencing, single base sequencing (reviewed by Cotton 1993), allele-specific oligonucleotide (ASO) (Malmgren et al. 1996), heteroduplex analysis, denaturing gradient gel electrophoresis (DGGE) (Cariello et al. 1988), SSCP assays (Suzuki et al. 1990), and ligation chain reaction (LCR) (Kalin et al. 1992). Each approach has its advantages and limitations, but all are useful for SNP genotyping, especially in small laboratories limited by budget and labor constraints. Large-scale analysis of SNP markers, however, depends on the availability of expensive, cutting-edge equipment. Obviously, direct sequencing is the most accurate way of SNP genotyping, but the cost, efforts, and time requirement made it impractical.

Single-base Sequencing (Primer Extension Typing, PET)

Sequence only one base or analyze the primer extension product under conditions only one base is allowed to be extended. In this procedure, the sequencing primer is designed with its last base ends one base ahead of the SNP sites. The primer is extended only one base by using dideoxynucleotide triphosphates (ddNTP). The primer extension product is then analyzed on a sequencing gel.

Allele-specific Oligo Hybridization (ASO)

This approach uses the principle of reverse Southern blot hybridization. Oligo primers with SNPs are synthesized and immobilized on a solid support. Genomic DNA is then amplified and used to hybridize to the oligos. Hybridization is conducted in strictly controlled temperature regimes so that the perfectly matched oligo hybridizes, but the oligo with a single mismatch does not.

Single-strand Conformation Polymorphism (SSCP)

SSCP relies on the fact that within a short DNA segment (usually no more than 300 base pairs [bp]), a single base change in the sequence can cause major changes in single-stranded conformation that is a reflection of the secondary structure of single-stranded DNA upon hairpin formation and minor base pairings (Figure 6.2). In the procedure, double-stranded DNA is first generated by PCR, followed by denaturation and formation of single-stranded structures by self-annealing under relative diluted concentrations that favors formation of single-stranded conformation over annealing between the two strands. The SSCP is then analyzed on nondenaturing gels.

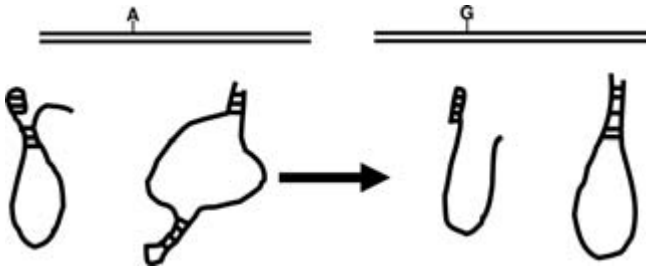


Figure 6.2. Schematic presentation of conformational changes due to single base substitutions. Note the conformations drawn are by no means a reflection of a real situation, but just an illustration of major conformation change. A single nucleotide change by base substitutions can lead to major changes in conformation of the single-stranded DNA, allowing allelic variations of SNPs to be differentiated by nondenaturing gel electrophoresis.

Heteroduplex Analysis

This approach relies on the fact that heteroduplexes run slower in gels than the homoduplexes. Upon amplification of a specific locus, the PCR products are subjected to heteroduplex analysis. In this procedure, the PCR products are first heat-denatured, and then allowed to re-anneal among strands. Three annealing products will be formed: two homoduplex representing the two genotypes with the SNP and one heteroduplex annealed between the molecules with SNP (Figure 6.3). Two types of heteroduplexes are common. Mutations involving deletions often lead to the formation of a bulge type of heteroduplex that can be readily differentiated on agarose gels, while base substitutions often lead to the formation of a bulb type of heteroduplex that requires a special gel mix to differentiate the heteroduplex from homoduplex (Bhattacharyya and Lilley 1989).

Denaturing Gradient Gel Electrophoresis (DGGE)

The principle of this technique is to separate DNA strands, based on their actual base composition, or the ratio of GC to AT base pairs that make up a particular segment of DNA. This is accomplished by exposing the DNA to a gradient of denaturant at

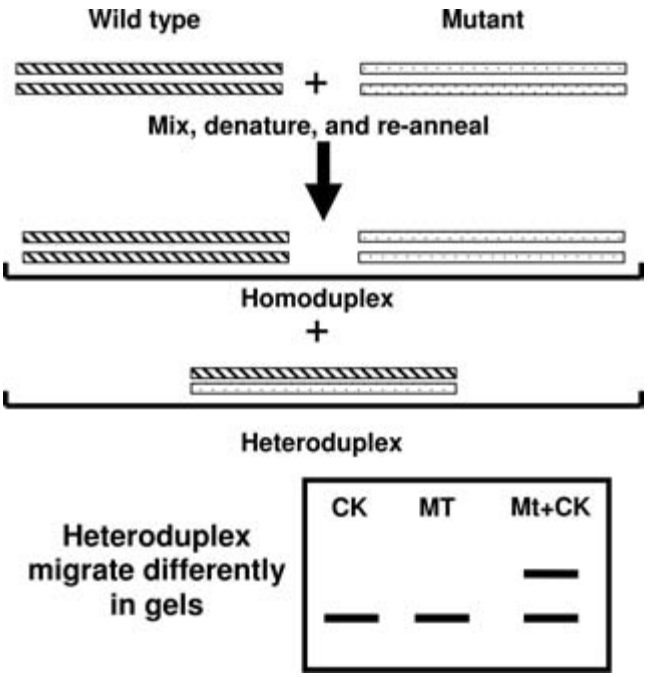


Figure 6.3. The principles of heteroduplex analysis.

elevated temperatures within a polyacrylamide gel. As the DNA sample progresses through the gel, from a low denaturant concentration to a higher one, it starts to melt at varying points. This is akin to the DNA "unzipping." The higher the GC content of the sample, the harder it is to melt. Thus, the DNA sample is able to progress further into the gel before stopping. Samples with lower GC content melt more rapidly in comparison. Therefore, they progress more slowly within the gel, thus becoming separated from the other faster-moving strands of DNA.

Ligation-mediated PCR

DNA ligases catalyze covalent joining of two DNA strands on the DNA template at a nick junction. The strict requirement of a base pair complementarity at the nick junction has been exploited for development of ligase-based technologies aimed at detection of sequence variations. After discovery of thermostable ligases, methods employing amplification of the diagnostic signal through repeated cycles of denaturation, annealing, and ligation have been developed analogous to PCR. The test is usually performed by designing two oligonucleotides specific for each allele and labeled differently on one side of SNP, and one common oligonucleotide on the other. Detection of the alleles can be performed directly in the microtiter plate wells by colorimetric approaches (Tobe et al. 1996), or by gel separations.

Recently Developed SNP Genotyping Approaches and Platforms

Several options are available for efficient genotyping using state-of-the-art equipment. Particularly popular are methods involving MALDI-TOF mass spectrometry (Ross et al. 1998, Storm et al. 2003), pyrosequencing (Ahmadian et al. 2000, Alderborn et al. 2000, He et al. 2003), *TaqMan* allelic discrimination (Li et al. 2004), real-time quantitative PCR (Nurmi et al. 2001), and the use of microarray or gene chips (Hacia et al. 1999). Mass spectrometry and microarray technologies require a large investment in equipment. The equipment for pyrosequencing and quantitative PCR is generally less than $100,000, and should be more affordable for laboratories working in the area of aquaculture genetics. Another consideration is the expense of genotyping in relation to sample sizes. Microarray (gene chip) technology and quantitative PCR are particularly useful in medical and clinical settings where large numbers of samples (thousands of individuals per locus) are involved and that can justify the cost involved in the development of the gene chips and hybridization probes. Mass spectrometry and pyrosequencing are relatively cost-effective (after acquisition of the equipment) when working with relatively small sample sizes (e.g., hundreds of individuals per locus), as is most likely the case in aquaculture research.

In addition to these major platforms, several recently developed SNP genotyping systems are particularly adaptable for situations involved in aquaculture genomics. The SnapShot SNP detection system and the invader assay are especially attractive because of their ease to automation. The principles of several SNP genotyping systems are described below.

TaqMan Technology

TaqMan technology integrates a PCR-based assay with laser scanning technology to excite fluorescent dyes present in the specially designed *TaqMan* probes. Its use in SNP genotyping is based on allele-specific hybridization. Briefly, the method is based on the 5′-3′ exonuclease activity of the *Taq* DNA polymerase that results in cleavage of the *TaqMan* probes, allowing fluorescence to be emitted during PCR; the intensity of fluorescence is then measured by a Sequence Detection System. The *TaqMan* probe is located between the two PCR primers and has a melting temperature 10°C higher than that of the primers; binding of the *TaqMan* probe prior to the primers is crucial because without it PCR products would be formed without generation of fluorescence intensity and thus without being detected. The *TaqMan* probe has two fluorescent tags attached to it. One is a reporter dye, such as 6-carboxyfluorescein (FAM), which has its emission spectra quenched due to the spatial proximity of a second fluorescent dye, 6-carboxy-tetramethyl-rhodamine (TAMRA). Degradation of the *TaqMan* probe, by the *Taq* DNA polymerase, frees the reporter dye from the quenching activity of TAMRA and thus the fluorescent activity increases with an increase in cleavage of the probe, which is proportional to the amount of PCR product formed (Figure 6.4). For SNP detection, two *TaqMan* probes are designed with one for each allele, and they are labeled with different fluorescent dye (e.g., VIC dye is linked to the 5' end of the Allele 1 probe. FAM dye is linked to the 5′ end of the Allele 2 probe). *Taq* DNA polymerase cleaves only probes that are hybridized to the target. Mismatches between a probe and target reduce the efficiency of probe hybridization. Furthermore, DNA polymerase is more likely to displace a mismatched probe without cleaving it, which does not produce a fluorescent signal.

The Invader Assay

The invader assay enables simultaneous detection of two different alleles (e.g., Ryan et al. 1999, Kwiatkowski et al. 1999, Cooksey et al. 2000, Hsu et al. 2001). Two oligonucleotide probes (an allele-specific primary probe and an invader probe) hybridize in tandem to the target DNA to form a specific overlapping structure (Figure 6.5). The 5′-portion of the primary probe contains a 5′-Flap that is noncomplementary to the target DNA and therefore cannot hybridize to the target sequence. The 3′-end of the invader probe overlaps the primary probe by a single base at the SNP site. A cleavage enzyme (Flap endonuclease I) recognizes the overlapping structure and cleaves the 5′-Flap on the primary probe at the base of the overlap releasing it as a target specific product. If the probe does not hybridize perfectly at the site of interest, no overlapping structure is formed, no cleavage occurs, and the target-specific product is not released.

The invader assay consists of two primary probes and two invader probes, with each set of primary probes and invader probes specific to Allele 1 or Allele 2, respectively, generating two target-specific products.

The target specific 5′-Flap oligos are involved in a secondary reaction for quantification of the fluorescent signals. The released target-specific 5′-Flap oligos act as invader probes on a fluorescent resonance energy transfer (FRET) cassette leading to the formation of an overlapping structure that is recognized by the cleavage enzyme. When the

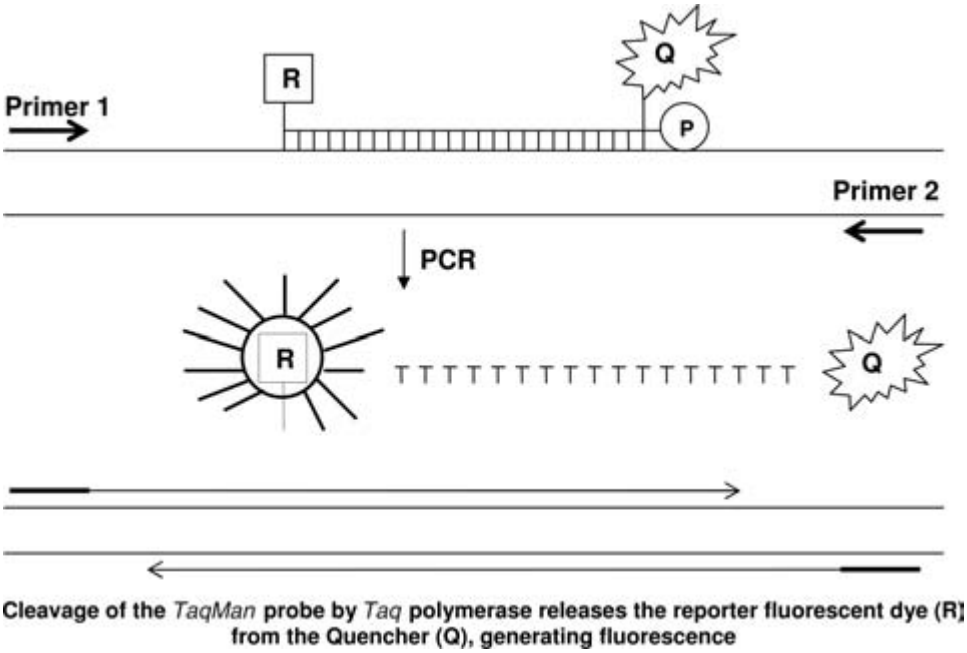


Figure 6.4. The *TaqMan* 5′-3′ nuclease assay. PCR primers 1 and 2 and a *TaqMan* probe, labeled with a reporter dye, FAM, (R) and a quencher dye, TAMRA, (Q), bind to the DNA template. The 3′ phosphate group (P) prevents extension of the *TaqMan* probe. The presence of the enzyme, *Taq* polymerase, enables extension of the primer which displaces the *TaqMan* probe. The displaced probe is cleaved by *Taq* DNA polymerase resulting in an increase in relative fluorescence of the reporter.

FRET cassette is cleaved, a fluorophore is released from a quencher, generating a fluorescence signal. There are two signal fluorophores attached to two different FRET cassettes that are spectrally distinct and specific to either allele of the biallelic system. The ratios of the two fluorescent signals then allow a genotype to be assigned.

MALDI-TOF Mass Spectrometry Platform

SNP genotyping using MALDI-TOF mass spectrometry involves PCR of the SNP region, annealing of a primer immediately ahead of the mutation spot, primer extension using dideoxy nucleotides, and mass spectrometric analysis based on molecular size of the primer extension products (Lawrence et al. 1997). It is one of the most efficient SNP genotyping methods and can perform 60,000 genotypes a day.

DNA Chip Platform for SNP Genotyping

Genotyping of SNP using the DNA chip technology can be viewed as the very high density of allele-specific oligo analysis as discussed above. A DNA chip is a small

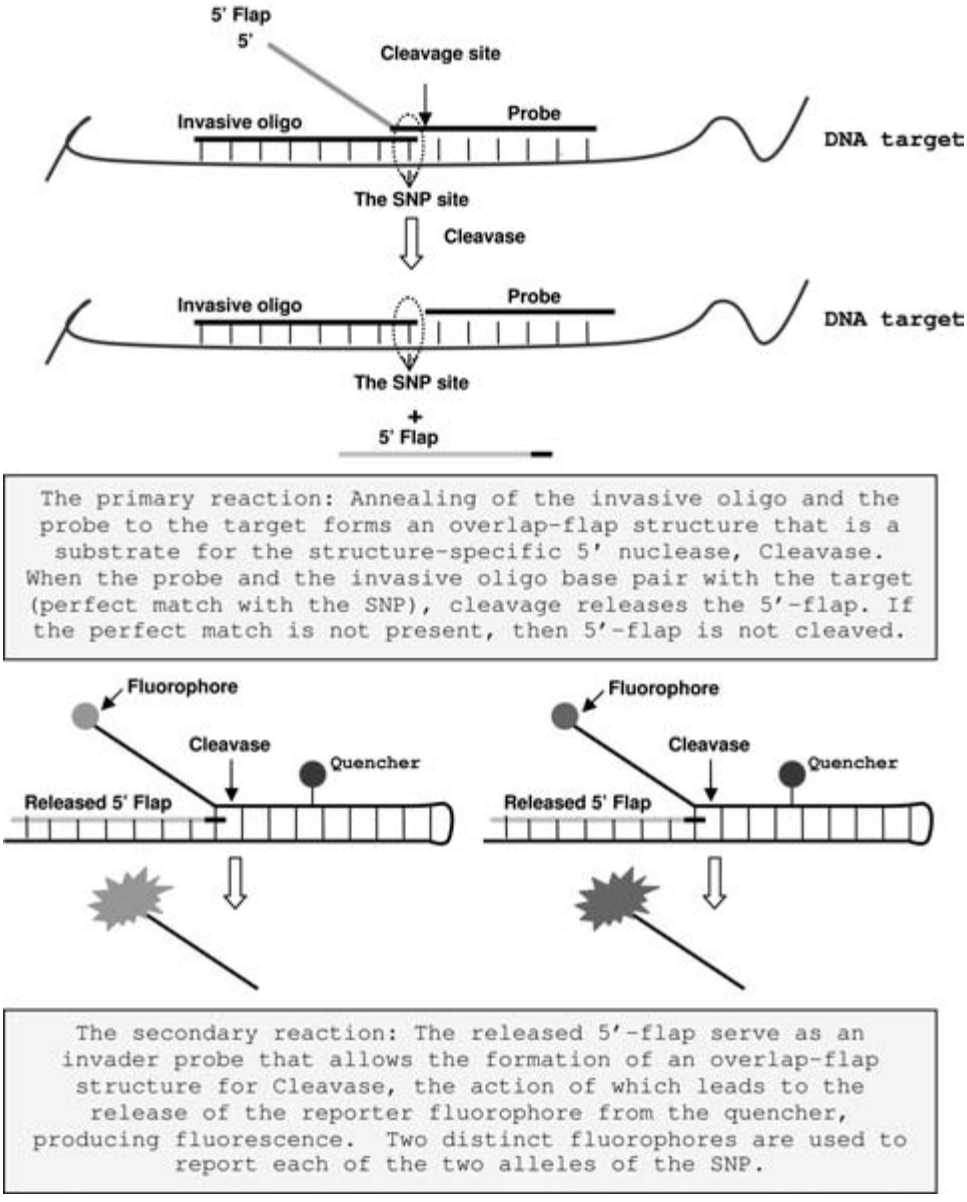


Figure 6.5. A schematic presentation of the invader assay of SNPs. (Also see color plate.)

piece of silicon glass (approximately 1 cm^2) to which a large number of synthetic, single-stranded DNA oligonucleotides ("oligos") have been chemically bonded. Oligos function as DNA probes. They anneal selectively only to those DNA molecules whose nucleotide sequences are exactly complementary—T pairs with A, and G with C. Therefore, they can be used to identify the presence of specific DNA sequences in a heterogeneous mixture of genes. DNA chips can be used to look for DNA sequences that differ by SNPs. To determine which alleles are present, genomic DNA

from an individual is isolated, fragmented, tagged with a fluorescent dye, and applied to the chip. The genomic DNA fragments anneal only to those oligos to which they are perfectly complementary.

The Beadarray Platform (The GoldenGate Assay)

Very recently, a Beadarray platform was developed by Illumina (Fan et al. 2003). As this platform provides the greatest flexibility, has the highest troughput, and is one of the most economical platforms (pennies per genotype), the Illumina's GoldenGate assay has become the most popular large-scale SNP assay. The allele discrimination at each SNP locus is achieved by using three oligos (P1, P2, and P3, each is tailed at 5′ with sequence A, B, and C, respectively, serving as universal primers for PCR), of which P1 and P2 are allele-specific and are Cy3- and Cy5-labeled. P3 is locus-specific designed several bases downstream from the SNP site. If the template DNA is homozygous, either P1 or P2 will be extended to meet P3; if the template is heterozygous, both P1 and p2 will be extended to meet P3, allowing ligation to happen. Upon ligation, the artificial, allele-specific template is created for PCR using universal primers. P3 contains a unique address sequence that targets a particular bead type with a complementary sequence to the address sequence. After downstream-processing, the single-stranded, dye-labeled DNAs are hybridized to their complement bead type through their unique address sequences. After hybridization, the BeadArray Reader is used to analyze flourescence signal on the Sentrix Array Matrix or Beadchip, which is in turn analyzed using software for automated genotype clustering and calling. Most recently, an even higher throughput system called iSelect Infinium Custom Genotyping was launched by Illumina, allowing tens of millions of genotypes to be determined simultaneously. Although the equipment performing the GoldenGate and iSelect assays are expensive, genotyping services are available. Interested readers are referred to the company's web site (http://www.illumina.com/). Undoubtedly, such efficient systems will have a tremendous impact on aquaculture genome research as well as on medical genomics.

Inheritance of SNP Markers

Theoretically, a SNP within a locus can produce as many as four alleles, each containing one of four bases at the SNP site—A, T, C, and G. Practically, however, most SNPs are usually restricted to one of two alleles (quite often either the two pyrimidines C/T or the two purines A/G) and have been regarded as biallelic. They are inherited as codominant markers in a Mendelian fashion. Obviously, their polymorphic information content (PIC) is not as high as multiallele microsatellites, but this shortcoming is balanced by their great abundance.

Conclusion

With so many approaches for SNP discovery and genotyping, it is not easy to determine the best approach. It all depends on the situation and objectives. Clearly, the genome of a large number of aquaculture species will not be sequenced and therefore, SNP

discovery in aquaculture species likely will come in the form of data mining using ESTs and BAC end sequences, although limited efforts using targeted PCR or reduced RSS are possible.

SNPs can be genotyped with a wide range of techniques and instrumentations, from small-scale, low-budget to expensive high-throughput systems. For SNP genotyping, the greatest determinants of the genotyping platform depend on the availability of equipment. Given the availability of the equipment, considerations can be made based on budget, number of markers, number of individuals, and the requirement for robustness. In spite of its current low levels of application in aquaculture genome research, SNP markers should gain in popularity as more and more sequence information becomes available in aquaculture species. Equally important, once the genetic linkage maps are well constructed, genome scans for quantitative trait loci (QTL) are expected to follow, in order to study traits important to aquaculture, which then depends on the use of well-defined association analysis. Because SNP markers are great markers for the analysis of trait-genotype associations, their application to aquaculture will become essential. It may be the case that a few laboratories working in aquaculture genomics will be able to map a large number of SNPs to genetic maps, saving the trouble for most other laboratories that can concentrate on studies involving biology and aquaculture traits.

Acknowledgments

Research in my laboratory is supported by grants from the USDA NRI Animal Genome and Genetic Mechanisms Program, the USDA NRI Basic Genome Reagents and Tools Program, the Mississippi-Alabama Sea Grant Consortium, the Alabama Department of Conservation, the USAID, and the BARD.

References

Abe K, H Noguchi, K Tagawa, M Yuzuriha, A Toyoda, T Kojima, K Ezawa, N Saitou, M Hattori, Y Sakaki, et al. 2004. Contribution of Asian mouse subspecies *Mus musculus* molossinus to genomic constitution of strain C57BL/6J, as defined by BAC-end sequence-SNP analysis. Genome Res, 14, pp. 2439–2447.

Ahmadian A, B Gharizadeh, AC Gustafsson, F Sterky, P Nyren, M Uhlen, and J Lundeberg. 2000. Single-nucleotide polymorphism analysis by pyrosequencing. Anal Biochem, 280, pp. 103–110.

Alderborn A, A Kristofferson, and U Hammerling. 2000. Determination of single-nucleotide polymorphisms by real-time pyrophosphate DNA sequencing. Genome Res, 10, pp. 1249–1258.

Altshuler D, VJ Pollara, CR Cowles, WJ Van Etten, J Baldwin, L Linton, and ES Lander. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature, 407, pp. 513–516.

Bader JS. 2001. The relative power of SNPs and haplotype as genetic markers for association tests. Pharmacogenomics, 2, pp. 11–24.

Bhattacharyya A and DM Lilley. 1989. The contrasting structures of mismatched DNA sequences containing looped-out bases (bulges) and multiple mismatches (bubbles). Nucleic Acids Res, 17, pp. 6821–6840.

Cariello NF, JK Scott, AG Kat, WG Thilly, and P Keohavong. 1988. Resolution of a missense mutant in human genomic DNA by denaturing gradient gel electrophoresis and direct sequencing using *in vitro* DNA amplification. Am J Hum Genet, 42, pp. 726–734.

Cooksey RC, BP Holloway, MC Oldenburg, S Listenbee, and CW Miller. 2000. Evaluation of the invader assay, a linear signal amplification method, for identification of mutations associated with resistance to rifampin and isoniazid in *Mycobacterium tuberculosis*. Antimicrob Agents Chemother, 44, pp. 1296–1301.

Cotton RG. 1993 Current methods of mutation detection. Mutat Res, 285, pp. 125–144.

Fan JB, A Oliphant, R Shen, BG Kermani, F Garcia, KL Gunderson, M Hansen, F Steemers, SL Butler, P Deloukas, et al. 2003. Highly Parallel SNP Genotyping. In: Cold Spring Harbor Symposium on Quantitative Biology, Volume LXVIII, pp. 69–78, Cold Spring Harbor Press, Cold Spring Harbor.

Gonen D, J Veenstra-VanderWeele, Z Yang, B Leventhal, and EH Cook Jr. 1999. High throughput fluorescent CE-SSCP SNP genotyping. Mol Psychiatry, 4, pp. 339–343.

Hacia JG, JB Fan, O Ryder, L Jin, K Edgemon, G Ghandour, RA Mayer, B Sun, L Hsie, CM Robbins, LC Brody, D Wang, ES Lander, R Lipshutz, SP Fodor, and FS Collins. 1999. Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. Nat Genet, 22, pp. 164–167.

Haff LA and IP Smirnov. 1997. Single-nucleotide polymorphism identification assays using a thermostable DNA polymerase and delayed extraction MALDI-TOF mass Spectrometry. Genome Res, 7, pp. 378–388.

Halldorsson BV, S Istrail, and FM De La Vega. 2004. Optimal selection of SNP markers for disease association studies. Hum Hered, 58, pp. 190–202.

He C, L Chen, M Simmons, P Li, S Kim, and ZJ Liu. 2003. Putative SNP discovery in interspecific hybrids of catfish by comparative EST analysis. Animal Genetics, 34, pp. 445–448.

Heaton MP, WM Grosse, SM Kappes, JW Keele, CG Chitko-McKown, LV Cundiff, A Braun, DP Little, and WW Laegreid. Estimation of DNA sequence diversity in bovine cytokine genes. Mamm Genome, 12, pp. 32–37.

Hsu TM, SM Law, S Duan, BP Neri, and PY Kwok. 2001. Genotyping single-nucleotide polymorphisms by the invader assay with dual-color fluorescence polarization detection. Clin Chem, 47, pp. 1373–1377.

John S, N Shephard, G Liu, E Zeggini, M Cao, W Chen, N Vasavda, T Mills, A Barton, A Hinks, S Eyre, KW Jones, W Ollier, A Silman, N Gibson, J Worthington, and GC Kennedy. 2004. Whole-genome scan, in a complex disease, using 11,245 single-nucleotide polymorphisms: comparison with microsatellites. Am J Hum Genet, 75, pp. 54–64.

Kalin I, S Shephard, and U Candrian. 1992. Evaluation of the ligase chain reaction (LCR) for the detection of point mutations. Mutat Res, 283, pp. 119–123.

Kim H, CM Hutter, SA Monks, and KL Edwards. 2005. Comparison of single-nucleotide polymorphisms and microsatellites in detecting quantitative trait loci for alcoholism: The Collaborative Study on the Genetics of Alcoholism. BMC Genet, 6 (Suppl 1), p. S5.

Kim H, CJ Schmidt, KS Decker, and MG Emara. 2003. A double-screening method to identify reliable candidate non-synonymous SNPs from chicken EST data. Anim Genet, 34, pp. 249–254.

Konfortov BA, VE Licence, and JR Miller. 1999. Re-sequencing of DNA from a diverse panel of cattle reveals a high level of polymorphism in both intron and exon. Mamm Genome, 10, pp. 1142–1145.

Kwiatkowski RW, V Lyamichev, M de Arruda, and B Neri. 1999. Clinical, genetic, and pharmacogenetic applications of the Invader assay. Mol Diagn, 4, pp. 353–364.

Lai E. 2001. Application of SNP technologies in medicine: lessons learned and future challenges. Genome Res, 11, pp. 927–929.

Li B, I Kadura, DJ Fu, and DE Watson. 2004. Genotyping with TaqMAMA. Genomics, 83, pp. 311–320.

Lin HF, SH Juo, and R Cheng. 2005. Comparison of the power between microsatellite and single-nucleotide polymorphism markers for linkage and linkage disequilibrium mapping of an electrophysiological phenotype. BMC Genet, 6 (Suppl 1), p. S7.

Ma Q, Y Yu, Y Meng, J Farrell, LA Farrer, and MA Wilcox. 2005. Genome-wide linkage analysis for alcohol dependence: a comparison between single-nucleotide polymorphism and microsatellite marker assays. BMC Genet, 6 (Suppl 1), p. S8.

Malmgren H, J Gustavsson, T Tuvemo, and N Dahl. 1996. Rapid detection of a mutation hotspot in the human androgen receptor. Clin Genet, 50, pp. 202–205.

Marnellos G. 2003. High-throughput SNP analysis for genetic association studies. Curr Opin Drug Discov Devel, 6, pp. 317–321.

Middleton FA, MT Pato, KL Gentile, CP Morley, X Zhao, AF Eisener, A Brown, TL Petryshen, AN Kirby, H Medeiros, et al. 2004. Genomewide linkage analysis of bipolar disorder by use of a high-density single-nucleotide-polymorphism (SNP) genotyping assay: a comparison with microsatellite marker assays and finding of significant linkage to chromosome 6q22. Am J Hum Genet, 74, pp. 886–897.

Nurmi J, M Kiviniemi, M Kujanpaa, M Sjoroos, J Ilonen, and T Lovgren. 2001. High-throughput genetic analysis using time-resolved fluorometry and closed-tube detection. Anal Biochem, 299, pp. 211–217.

Rafalski A. 2002. Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol, 5, pp. 94–100.

Ross P, L Hall, I Smirnov, and L Haff. 1998. High level multiplex genotyping by MALDI-TOF mass spectrometry. Nat Biotechnol, 16, pp. 1347–1351.

Ryan D, B Nuccie, and D Arvan. 1999. Non-PCR-dependent detection of the factor V Leiden mutation from genomic DNA using a homogeneous invader microtiter plate assay. Mol Diagn, 4, pp. 135–144.

Sachidanandam R, D Weissman, SC Schmidt, JM Kakol, LD Stein, G Marth, S Sherry, JC Mullikin, BJ Mortimore, DL Willey, et al. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature, 409, pp. 928–933.

Schaid DJ, JC Guenther, GB Christensen, S Hebbring, C Rosenow, CA Hilker, SK McDonnell, JM Cunningham, SL Slager, ML Blute, and SN Thibodeau. 2004. Comparison of microsatellites versus single-nucleotide polymorphisms in a genome linkage screen for prostate cancer-susceptibility Loci. Am J Hum Genet, 75, pp. 948–965.

Schmid M, I Nanda, M Guttenbach, C Steinlein, M Hoehn, M Schartl, T Haaf, S Weigend, R Fries, JM Buerstedde, et al. 2000. First report on chicken genes and chromosomes Cytogenet. Cell Genet, 90, pp. 169–218.

Sorrentino R, I Potolicchio, GB Ferrara, and R Tosi. 1992. A new approach to HLA-DPB1 typing combining DNA heteroduplex analysis with allele-specific amplification and enzyme restriction. Immunogenetics, 36, pp. 248–254.

Storm N, B Darnhofer-Patel, D van den Boom, and CP Rodi. 2003. MALDI-TOF mass spectrometry-based SNP genotyping. Methods Mol Biol, 212, pp. 241–262.

Stram DO. 2004. Tag SNP selection for association studies. Genet Epidemiol, 27, pp. 365–374.

Suzuki Y, M Orita, M Shiraishi, K Hayashi, and T Sekiya. 1990. Detection of *ras* gene mutations in human lung cancers by single-strand conformation polymorphism analysis of polymerase chain reaction products. Oncogene, 5, pp. 1037–1043.

Thalamuthu A, I Mukhopadhyay, A Ray, and DE Weeks. 2005. A comparison between microsatellite and single-nucleotide polymorphism markers with respect to two measures of information content. BMC Genet 2005, 6 (Suppl 1), p. S27.

Tobe VO, SL Taylor, and DA Nickerson. 1996. Single-well genotyping of diallelic sequence variations by a two-color ELISA-based oligonucleotide ligation assay. Nucleic Acids Res, 24, pp. 3728–3732.

Tost J and IG Gut. 2005. Genotyping single nucleotide polymorphisms by MALDI mass spectrometry in clinical applications. Clin Biochem, 38, pp. 335–350.

Vignal A, D Milan, M SanCristobal, and A Eggen. 2002. A review on SNP and other types of molecular markers and their use in animal genetics. Genet Sel Evol, 34, pp. 275–305.

Wilcox MA, EW Pugh, H Zhang, X Zhong, DF Levinson, GC Kennedy, and EM Wijsman. 2005. Comparison of single-nucleotide polymorphisms and microsatellite markers for linkage analysis in the COGA and simulated data sets for Genetic Analysis Workshop 14: Presentation Groups 1, 2, and 3. Genet Epidemiol, 29 (Suppl 1), pp. S7–28.

Xing C, FR Schumacher, G Xing, Q Lu, T Wang, and RC Elston. 2005. Comparison of microsatellites, single-nucleotide polymorphisms (SNPs) and composite markers derived from SNPs in linkage analysis. BMC Genet, 6 (Suppl 1), p. S29.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 7
Allozyme and Mitochondrial DNA Markers

Huseyin Kucuktas and Zhanjiang Liu

In the first five chapters of this section, various marker systems were discussed including restriction fragment length polymorphism (RFLP), random amplification of polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP), microsatellites, and single nucleotide polymorphisms (SNP) that are all important to aquaculture genome research. Here we present two additional types of markers—the allozyme markers and the mitochondrial DNA markers. In spite of the very limited uses of these two types of markers for the purpose of genome research, they have been among the most popular markers in aquaculture and fishery research in the past.

Allozyme Markers

Genetic diversity measurements in aquaculture stocks are an essential part of an effective management strategy. Historically, these measurements relied on phenotypic or qualitative markers that were used in classical genetics. Morphological differences such as body dimensions, size, and pigmentation are some examples of phenotypic markers. Genetic diversity measurements based on phenotypic markers are often indirect, and inferential through controlled breeding and performance studies (Okumus and Ciftci 2003, Parker et al. 1998). Because these markers are polygenically inherited and have low heritabilities, they may not represent the true genetic differences (Smith and Chesser 1981). Only when the genetic basis for these phenotypic markers is known, can some of them be used to measure genetic diversity. Molecular markers were developed to overcome problems associated with phenotypic markers.

Many types of molecular markers are currently available to use in both aquaculture and population genetics. These markers can be categorized in two types: protein and DNA markers (Okumus and Ciftci 2003). In terms of variation source, on the other hand, they are classified into three groups: allozyme markers, mitochondrial DNA markers, and nuclear DNA markers. Allozymes are protein products of genes that are encoded by a single gene locus. Since they represent genes of known function, they are considered to be Type I markers (Liu and Cordes 2004).

The term "isozyme" refers to multiple biochemical forms of an enzyme having identical substrate specificity (or the same catalytic activities) within the same organism. "Allozymes" or "allelic isozymes" are the different allelic forms of the same enzymes encoded at the same locus (Hunter and Market 1957, Parker et al. 1998, May 2003). Strictly speaking, allozymes represent different allelic forms of the same gene and isozymes represent different genes whose products catalyze the same reaction. However, the two terms are usually used interchangeably. The variation detected in allozymes may be the result of point mutations, insertions, or deletions (indels). It is

believed that isozymes usually form as a result of gene duplication; however, there may be other events (hybridization, polyploidization) that lead to the formation of isozymes.

The most common use of allozyme electrophoresis is to detect genetic variation in natural populations. In the last 30 years, large amounts of allelic frequency data were collected from many fish species for management purposes. Although use of allozyme data in aquaculture appears to be limited compared to population studies in fisheries, the aquaculture industry has long used this information for its development because aquaculture and fisheries can not be separated from each other (Dunham 2004). Allozyme electrophoresis in aquaculture is used for stock identification, parentage analysis, inbreeding analysis, and limited genetic mapping (Liu and Cordes 2004).

Some Considerations Related to Allozyme Analysis

In most cases, sampling for allozyme analysis is lethal. Most commonly used tissues in allozyme electrophoresis for analyzing the full range loci are muscle, liver, eye, and heart samples that are freshly obtained from individual fish. Since enzymes are heat labile, samples are either immediately processed and analyzed or properly frozen. Sample preparation requires each tissue to be mechanically ground in a buffer, but a sonicator can also be used for this purpose. However, heat generated during the sonication process often yields poor resolution due to enzyme degradation. Because many enzyme loci are used to detect genetic variability in fisheries and aquaculture, we will not provide detailed descriptions of the allozyme assays, but rather provide a comprehensive list of the most frequently used allozymes (Table 7.1) with linkage to Internet sources containing detailed descriptions and references.

Commonly used support media in allozyme electrophoresis are starch, cellulose acetate, acrylamide, and agarose. With the consideration for cost, resolving power, and electrophoresis time, starch gels are the most often used medium of support.

There is no single buffer system that will give a desirable resolving power with starch electrophoresis. Every enzyme has its own optimal buffer systems. Interested readers are referred to three published books on isozymes (Pasteur et al. 1987, Morizot and Schmidt 1990, Richardson et al. 1986). These books are excellent sources of information related to isozyme analysis.

Genotyping of allozyme gels can be complicated at times. Individual genotypes at each locus are inferred from the banding patterns observed on the gels. Allele nomenclature and the locus identification in allozyme electrophoresis used in aquaculture are based on the relative mobility of the proteins (Shaklee et al. 1990). Before using electrophoretic data for genetic variation analysis, breeding data must be used to verify observed variation (Wolf et al. 1970). The variation detected by electrophoretic data may not be limited to true genetic variation. In some exceptional cases, patterns on the gels do not always fit the simple Mendelian inheritance. One of the complications is the presence of null alleles (Stoneking et al. 1981). The second one is the sample artifact due to improper storage and processing. Some enzyme systems may give different banding patterns due to pathological or environmental differences. The pattern detected after gel staining depends on the quaternary structure of the enzyme. Diploid organisms have two copies of each gene—one maternal, the other paternal. However, in some cases there may be multiple copies of the same enzyme

Table 7.1. List of enzymes or other protein loci used in fish genetic research. The most commonly used enzymes were indicated with gray shading. The list was compiled from Shaklee et al. 1990 and BRENDA enzyme database (http://brenda.bc.uni-koeln.de/) (Schomburg et al. 2004). *Gene functions are described in the Gene Ontology Web page (http://www.geneontology.org/). By using either the E. C. number or the enzyme name, one can obtain more information including related literature.

Enzyme or protein name	E. C. number	Abbreviation	Gene ontology number*
beta-N-acetylgalactosaminidase	3.2.1.53	bGALA	-
beta-N-acetylhexosaminidase	3.2.1.30	bGLUA	16231
Acid phosphatase	3.1.3.2	ACP	3993
Aconitate hydratase	4.2.1.3	AH	3994
Adenine phosphoribosyltransferase	2.4.2.7	APRT	3999
Adenosine deaminase	3.5.4.4	ADA	46936
Adenosine kinase	2.7.1.20	ADK	4001
Adenylate kinase	2.7.4.3	AK	4017
Alanine transaminase	2.6.1.2	ALAT	4021
Alcohol dehydrogenase	1.1.1.1	ADH	4025
Alkaline phosphatase	3.1.3.1	ALP	4035
alpha-Amylase	3.2.1.1	aAMY	4556
alpha-N-arabinofuranosidase	3.2.1.55	aARAF	46556
Aspartate aminotransferase	2.6.1.1	AAT	4069
Carbonate dehydratase	4.2.1.1	CAH	4089
Carboxylesterase	3.1.1.1	ESTD	16789
Catalase	1.11.1.6	CAT	4096
Creatine kinase	2.7.3.2	CK	4111
6,7-dihydropteridine reductase	1.5.1.34	DHPR	4155
Dipeptidase	3.4. -.-	PEPA	
Esterase	3.1.1. -	EST	-
Fructose bisphosphatase	3.1.3.11	FBP	42132
Fructose-bisphosphate aldolase	4.1.2.13	FBALD	4332
alpha-L-Fucosidase	3.2.1.51	aFUC	4560
Fumarate hydratase	4.2.1.2	FH	4333
Galactokinase	2.7.1.6	GALK	4335
alpha-Galactosidase	3.2.1.22	aGAL	4557
beta-Galactosidase	3.2.1.23	bGAL	9341
General (unidentified) protein	-	PROT	-
Glucose 1-dehydrogenase	1.1.1.47	GDH	47936
Glucose-6-phosphate 1-dehydrogenase	1.1.1.49	G6PDH	4345
Glucose-6-phosphate isomerase	5.3.1.9	GPI	4347
alpha-Glucosidase	3.2.1.20	aGLU	4558
beta-Glucosidase	3.2.1.21	bGLU	8442
beta-Glucuronidase	3.2.1.31	GUS	4566
Glutamate dehydrogenase	1.4.1.-	GLUDH	-
Glutamate-ammonia ligase	6.3.1.2	GLAL	4356
Glutathione-disulfide reductase	1.8.1.7	GR	4362
Glyceraldehyde-3-phosphate dehydrogenase (phosphorylating)	1.2.1.12	GAPDH	8943
Glycerate dehydrogenase	1.1.1.29	GLYDH	8465

(*Continued*)

Table 7.1. (*Continued*)

Enzyme or protein name	E. C. number	Abbreviation	Gene ontology number*
Glycerol-3-phosphate dehydrogenase (NAD+)	1.1.1.8	G3PDH	4367
Guanine deaminase	3.5.4.3	GDA	8892
Guanylate kinase	2.7.4.8	bGUK	4385
Hekokinase	2.7.1.1	HK	4396
Hemoglobin	-	HB	-
Hydroxyacylglutathione hydrolase	3.1.2.6	HAGH	4416
3-Hydroxybutyrate dehydrogenase	1.1.1.30	HBDH	3858
Hypoxanthine phosphoribosyltransferase	2.4.2.8	HPRT	4422
Inorganic pyrophosphatase	3.6.1.1	PP	4427
Lactoylglutathione lyase	4.4.1.5	aLGL	4462
L-iditol dehydrogenase	1.1.1.14	sIDDH	3939
L-Lactate dehydrogenase	1.1.1.27	LDH	4459
lsocitrate dehydrogenase (NADP+)	1.1.1.42	IDHP	4450
Malate dehydrogenase	1.1.1.37	MDH	30060
Malate dehydrogenase (decarboxylating)	1.1.1.39	ME	4471
Malate dehydrogenase (oxaloacetate-decarboxylating)	1.1.1.38	ME	16619
Malate dehydrogenase (oxaloacetate-decarboxylating) (NADP+)	1.1.1.40	MEP	4473
Mannose-6-phosphate isomerase	5.3.1.8	MPI	4476
alpha-Mannosidase	3.2.1.24	sMAN	4559
Nucleoside-phosphate kinase	2.7.4.4	NPK (UMPK)	50145
Nucleoside-triphosphate diphosphatase	3.6.1.19	NTP	47429
Octanol dehydrogenase	1.1.1.73	ODH	4552
Ornithine carbamoyltransferase	2.1.3.3	OTC	9348
Parvalbumin	-	PVALB	-
Peptidase-C	3.4. -. -	PEPC	
Peptidase-S	3.4. -. -	PEPS	
6-phosphofructokinase	2.7.1.11	PFK	3872
Phosphoglucomutase	5.4.2.2	PGM	4614
Phosphogluconate dehydrogenase (decarboxylating)	1.1.1.44	6PGDH	4616
Phosphoglycerate kinase	2.7.2.3	PGK	4618
Phosphoglycerate mutase	5.4.2.1	PGAM	46538
Phosphoglycolate phosphatase	3.1.3.18	PGP	8967
Phosphopyruvate hydratase	4.2.1.11	ENO	4634
Proline dipeptidase	3.4.13.9	PEPD	4251
Purine-nucleoside phosphorylase	2.4.2.1	PNP	4731
Pyruvate kinase	2.7.1.40	PK	4743
Superoxide dismutase	1.15.1.1	SOD	16954
Thymidine kinase	2.7.1.21	TK	4797
Transferrin	-	TF	-
Triose-phosphate isomerase	5.3.1.1	TPI	4807
Tripeptide aminopeptidase	3.4.11.4	PEPB	45148
Tyrosine transaminase	2.6.1.5	TAT	4838
UDP-glucose-hexose-1-phosphate uridylyltransferase	2.7.7.12	UGHUT	8108

gene due to genome or gene duplication (especially in fish). An active allozyme may have more than one subunit, and both allelic forms may result in polymeric bands. These pose a great challenge in gel scoring. Assuming the subunits of enzymes detected combine in a random fashion (Utter et al. 1974), the simplest allozyme pattern with a single polypeptide chain (monomer) yields three possible genotypes: AA, AB, and BB. However, interpretation becomes much more complex when multimeric enzymes composed of two or more subunits are involved (Figure 7.1). Glucose-6-phosphate isomerase (GPI) enzyme in brook trout *Salvelinus fontinales* is a good example of tetraploidization and the confusing nature of gel scoring. Although GPI is a dimeric enzyme, the banding pattern in gel presented in Figure 7.2 does not fit the

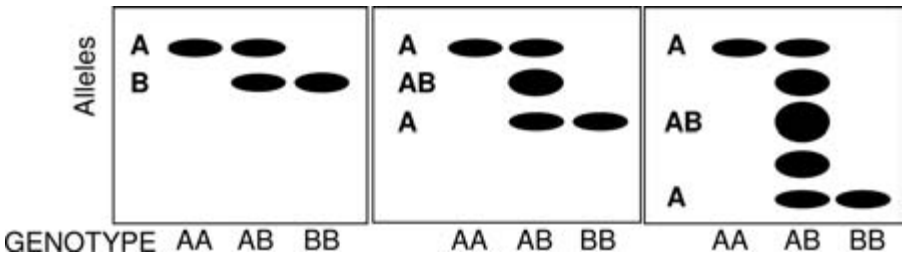


Figure 7.1. Allozyme patterns commonly observed in fishes showing a co-dominant fashion of diallelic forms. In a monomeric case, a homozygous sample with allele A should produce one band (genotype AA); similarly, a homozygous sample with allele B also produces one band (genotype BB); however, the heterozygous sample should produce two bands (genotype AB). In a dimeric case, the homozygous samples produce only one band (AA or BB), but a heterozygous sample produces three bands (AA homodimer, BB homodimer, and AB heterodimer). In a trimeric case, the situation is more complex: homozygous samples produce only one band, AA or BB; heterozygous samples produce five bands with various intensities depending on their proportion in the sample: A4 homotetramer, A3B heterotetramer, A2B2 heterotetramer, AB3 heterotetramer, and B4 homotetramer.

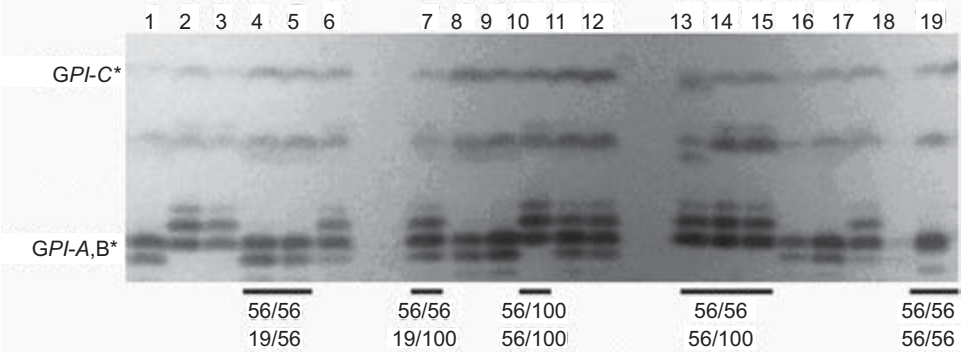


Figure 7.2. Glucose-6-phosphate isomerase (GPI) enzyme in brook trout *Salvelinus fontinales.* Sample #19 is homozygous at both loci, #13, #14, and #15 are homozygous at one and heterozygous at the other loci, and #10 is heterozygous at both loci. Note the hybrid band between duplicated loci and GPI-C*.

simple three banding pattern for a dimeric enzyme. On the other hand, a careful examination of this gel shows the duplicated nature of the enzyme. Therefore, great caution should be exercised in scoring allozyme gels.

The frequency of alleles can be calculated from the allozyme banding patterns. Allele and genotype frequencies and the relevant descriptive statistics are calculated using a variety of computer programs. BIOSYS (Version 1.7) (Swofford and Selander 1981) is the most commonly used program to analyze allozymes data in fish. The program can be used for allele frequency and genetic variability computations to test for Hardy-Weinberg equilibrium to determine substructuring, for linkage disequilibrium, calculations for F-statistics (Weir and Cockerham 1984), similarity and distance analysis (Nei 1978), and for construction of phenograms using cluster analysis (Rogers 1972). There are many other programs available to use in analyzing allozymes data. An extensive list of programs that are used for population genetics can be found at http://evolution.genetics.washington.edu/phylip/software.html.

Applications and Limitations of Allozyme Markers

Allozymes have wide applications in fisheries and aquaculture including population analysis, mixed stock analysis, and hybrid identification. In spite of limited numbers of loci, compared to other genetic tools, allozyme analysis had the most profound effect on fisheries science (May 2003). For example, inbreeding is the major concern for the aquaculture industry. Using genetic variation measured by the allozyme electrophoresis, relative degrees of inbreeding can be estimated by average heterozygosity comparisons in different broodstocks (Allendorf and Phelps 1981, Liu and Cordes 2004). Although allozyme studies did not find common application in marker assisted selection, correlations between certain allozyme markers and performance traits has been reported (Hallerman et al. 1986). Similarly, due to the limited number of polymorphic loci available, use of these markers in linkage mapping in fish is limited (Pasdar et al. 1984, May and Johnson 1993, Morizot et al. 1994). Allozyme data are also used in hybrid systems. First generation hybrids (F1) and later generation hybrids can be identified with a certain degree of probability if there are enough markers available (Avise and Van den Avyle 1984). One of the most common applications of allozyme data has been the use of mixed stock analysis, especially in salmonids (May 2003, Koljonen and Wilmot 2005). The effects of hatchery reared stocks of brown trout (*Salmo trutta*) and rainbow trout (*Oncorhynchus mykiss*) with the wild populations from which they were originated was studied by comparing the allozyme variation by Ferguson and others (1991). Use of allozyme markers in aquaculture also includes the genetic monitoring at the hatchery populations by establishing a monitoring program (Pérez et al. 2001).

The major drawback of allozyme analysis is the necessity for a large amount of fresh or frozen tissue samples. This often requires lethal sampling especially if the full array of allozyme markers are to be studied. Furthermore, although allozymes represent actual gene products, they often measure a very small portion of the genomic variation because a limited number of loci are involved (Utter et al. 1987). Mutation at the DNA level that causes a replacement of a similarly charged amino acid may not be detected by allozyme electrophoresis. Although very cheap and technically easy, the number of allozyme loci and the polymorphism is low (Agnese et al. 1997). These drawbacks will seriously limit the applications of allozymes for genome studies.

Mitochondrial DNA Markers

Mitochondria are the energy factory of living cells. Unlike the rest of the cellular functions that are determined by the nuclear DNA, these organelles have their own double-stranded circular DNA, known as mitochondrial genome (mtDNA). The size of the mitochondrial genome in animals varies among different organisms, but typically the number is around 16,000–19,000 base pairs (bp) in teleost fish (Table 7.2). Animal mitochondrial genome encodes 13 proteins, 2 ribosomal RNAs, and 22 transfer RNAs (Boore 1999). These genes are highly packed on the mitochondrial genome without introns (Burger et al. 2003). In addition to these genes, a highly variable control region, known as the D-loop (displacement loop) serves as the heavy-strand replication origin (Harrison 1989, Parker et al. 1998).

Table 7.2. List of selected aquaculture species whose mtDNA is sequenced, mtDNA size, and GenBank accession numbers.

Fish name	Size bp	Accession #	Reference
Anguilla japonica (Japanese eel)	16,685	NC_002707	Inoue et al. 2001
Conger myriaster (conger eel)	18,705	NC_002761	Inoue et al. 2001
Ictalurus punctatus (channel catfish)	16,497	NC_003489	Waldbieser et al. 2003
Cyprinus carpio (common carp)	16,575	NC_001606	Chang et al. 1994
Carassius auratus (goldfish)	16,579	NC_002079	Murakami et al. 1998
Danio rerio (zebrafish)	16,596	NC_002333	Broughton et al. 2001
Salmo salar (Atlantic salmon)	16,665	NC_001960	Hurst et al. unpublished
Oncorhynchus mykiss (rainbow trout)	16,642	NC_001717	Zardoya et al. 1995
Gadus morhua (Atlantic cod)	>10,000	Not available	Johansen et al. 1990
Oncorhynchus tshawytscha (chinook salmon)	16,644	NC_002980	Bernales et al. unpubished
Salvelinus alpinus (arctic char)	16,659	NC_000861	Doiron et al. 2002
Crassostrea virginica (eastern oyster)	17,242	NC_007175	Milbury and Gaffney 2005
Oreochromis mossambicus (Mozambique tilapia)	16641	NC_007231	Chen et al. unpubished
Tetraodon nigroviridis (green-spotted freshwater pufferfish)	16462	NC_007176	Yue et al. unpubished
Penaeus monodon (tiger prawn)	15984	NC_002184	Wilson et al. 2000
Scomber scombrus (Atlantic mackerel)	16560	NC_006398	Takashima et al. unpublished
Oryzias latipes (Japanese medaka)	16714	NC_004387	Miya et al. 2003
Esox lucius (northern pike)	16695	NC_004593	Ishiguro et al. 2003
Pagrus major (red seabream)	17031	NC_003196	Miya et al. 2001
Plecoglossus altivelis (ayu-fish)	16,537	NC_002734	Ishiguro et al. 2001
Salvelinus fontinalis (brook trout)	16,624	NC_000860	Doiron et al. 2002
Coregonus lavaretus (common whitefish)	16,737	NC_002646	Miya and Nishida (2000)
Takifugu rubripes (pufferfish)	16,442	NC_004299	Elmerot et al. 2002

Background and Principles

Mitochondrial genome evolves more rapidly than the nuclear genome. The rapid evolution of the mtDNA makes it highly polymorphic within a given species. The polymorphism is especially high in the control region (D-loop region), making the D-loop region highly useful in population genetic analysis. Since recombination is a rare event in mtDNA, common ancestry studies could be done with mtDNA analysis (Avise 1994). However, recent reports have indicated that recombination does occur in mtDNA (Harrison 1989, Rokas et al. 2003, Guo et al. 2006), but new genotypes could not be observed because recombination of identical mtDNA molecules should produce the same molecules (homoplasmy) (Maoguolas 2005). Therefore, mutations in mtDNA can be detected, but recombination of mtDNA most often cannot be noticed.

Mitochondrial DNA is maternally inherited for the most part, but there are reports of paternal leakage during fertilization (Birky et al. 1989). Each cell contains a variable number of mitochondria ranging from a few hundred to more than 10,000 mitochondria per cell depending on the cell types. The cells of the brain, the skeletal and heart muscles, and the eye contain the highest number of mitochondria (as many as 10,000 per cell) while the skin cells, which do not require much energy, contain only a few hundred of them. In spite of a large variation of mitochondria per cell, each mitochondrion contains a constant number of mtDNA molecules (Robin and Wong 1988). Since there are multiple copies of mtDNA per cell, many copies of them are transmitted to each offspring.

A variant mtDNA molecule transmitted to the offspring will cause heteroplasmy, or presence of two types of mtDNA in an organism. Variations in mtDNA are caused by mutations. Partitioning of different types of mtDNA into a single oocyte would result in heteroplasmy. For example, heteroplasmy detected in Milkfish (*Chanos chanos*) is attributed to both the 41 bp tandem repeat structure and the 48 bp indel at the control region of the mtDNA (Ravago et al. 2002). In addition to the heteroplasmy caused by mutations, high rates of heteroplasmy can also be caused by paternal mtDNA leakage (Kaneda et al. 1995, Ballard et al. 2005).

Due to the high levels of polymorphism and the ease of mitochondrial DNA analysis, mtDNA has been widely used as markers in aquaculture and fisheries settings. As discussed above, the non-Mendelian inheritance greatly limits the applications of mtDNA for genome research. However, as an identification tool often used in aquaculture, mtDNA can be used as a supplemental tool for aquaculture genome research. Here we will briefly describe the principles behind the wide application of mtDNA markers.

Mitochondrial DNA analysis is actually a restriction fragment length polymorphism (RFLP) analysis, as discussed in Chapter 2, except that the target molecule is mtDNA rather than nuclear genomic DNA (Liu and Cordes 2004). The high levels of polymorphism, the maternal inheritance, and the relatively small size of mtDNA make the RFLP analysis using mtDNA one of the easiest methods for many population studies (Okumus and Ciftci 2002, Liu and Cordes 2004, Billington 2003). The RFLP polymorphism detected in mtDNA is usually caused by a gain or loss of restriction sites. For example, striped bass (*Morone saxatilis*) exhibits a site loss or gain at *Xba* I restriction site, causing an RFLP polymorphism that could easily be detected with polymerase chain reaction (PCR) amplification of the polymorphic region followed by *Xba* I restriction digestion. However, polymorphism could also be caused by insertions or deletions leading to a length variation of mtDNA (Ravago et al. 2002). In this case,

electrophoresis of PCR products in the region should provide information on mtDNA haplotypes (Billington 2003).

mtDNA Analysis

Analysis of mtDNA polymorphism has become a useful genetic tool for studies of genetic divergence within and among populations (Avise 1994). Because mtDNA shows considerable variation among individuals, it is regarded as an effective marker for population structure and geographic variations. Distinct mtDNA lineages have been detected in many freshwater fishes in different parts of their species ranges. Because only half of the population (assuming 1:1 sex ratio) passes on their mtDNA to their offspring, effective population size for mtDNA is smaller than that of nuclear DNA (Harrison 1989).

Early studies using mtDNA analysis relied on purified mtDNA before the invention of PCR. Back in the 1970s and 1980s, the available molecular markers included allozymes, few RFLP, and mtDNA. In aquaculture species, RFLP markers were quite rare at that time. Most studies of natural populations have found greater genetic diversity in mtDNA compared to that revealed by allozyme electrophoresis. Lack of diversity with allozyme markers in striped bass, walleye, and many other fish species was the driving force for different laboratories to use mtDNA analysis to search for more genetic diversity in a variety of fish in the United States in the 1970s to 1980s (Wirgin et al. 1991, 1997; Billington and Herbert 1988). Initial studies done with mtDNA analysis were performed with a whole mtDNA molecule.

Mitochondrial DNA isolation was carried out by first purifying the mitochondria from the tissues containing high amounts of mitochondria, usually liver and gonadal tissues, and extraction of mtDNA from tissue lysates using density gradient centrifugation (Chapman and Brown 1990, Billington and Herbert 1988). A number of 4 and 6 base-cutter restriction endonucleases were used to digest the whole mtDNA molecule to search for fragment length polymorphisms using electrophoresis. The number and size of fragments obtained after digestion and electrophoresis produced a "haplotype." The comparison of haplotypes from several individuals is considered to be representative of the nucleotide differences of their whole mtDNA sequences because of the inferred restriction site loss at certain regions due to mutations (Maogulas 2003).

After the availability of PCR technology, RFLP analysis of the PCR-amplified regions of the mtDNA, rather than using the whole mtDNA, greatly improved mtDNA analysis. There is no need to purify the mtDNA for PCR; a simple total DNA extracted with a commercial kit is sufficient. Most often, the D-loop region is amplified by PCR, and then analyzed by RFLP. In some cases, other coding regions such as ND3, ND4, ND5, ND6, 12S, and 16S RNA regions of the mtDNA were also used (Merker and Woodruff 1996, Nielsen et al. 1998, Wirgin et al. 1997). Whether or not the PCR products need to be digested depends on the nature of the mtDNA RFLP. If the polymorphism is caused by length difference due to insertion or deletion, no restriction digestion is needed. After PCR amplification, the amplicon is analyzed directly by gel electrophoresis. If the polymorphism is caused by gain or loss of restriction site, the PCR amplicon needs to be digested by the restriction enzyme, and then analyzed by gel electrophoresis.

Applications, Data Analysis, and Limitations

Mitochondrial DNA markers have been used extensively to analyze genetic variation in several different aquaculture species including striped bass (Wirgin et al. 1991, Garber and Sullivan 2006), channel catfish (Waldbieser et al. 2003), walleye (Merker and Woodruff 1996), salmonids (Nielsen et al. 1998, Crespi and Fulton 2004), red snapper (Pruett et al. 2005), and bluegill (Chapman 1989). Data analysis in mtDNA studies include determining the number of mtDNA haplotypes, calculating the haplotype frequencies, and nucleotide diversity. Various computer programs such as Arlequin (Schneider et al. 2000), (http://anthro.unige.ch/arlequin/) and TFPGA (Miller 1997) (http://www.marksgeneticsoftware.net/tfpga.htm) are available to perform these analyses. A review by Labate (2000) describes the attributes of several software applications for population genetic analysis.

There are two major drawbacks of mtDNA markers. One is the non-Mendelian inheritance of mtDNA, and the second is the proportion of the total genomic variation one can observe with mtDNA alone. Additionally, mtDNA markers are subject to similar problems that exist for other DNA-based markers. For example, in back mutation cases, nucleotide sites that have already undergone substitution are returned to their original state. Mutations taking place at the same site on the mtDNA in independent lineages and unparallel rates of heterogeneity at the same region (Liu and Cordes 2004) can place limitations on the validity of using mtDNA for genetic studies.

References

Agnèse JF, B Adépo-Gourène, EK Abbans, Y Fermon. 1997. Genetic differentiation among natural populations of the Nile tilapia *Oreochromis niloticus* (Teleostei, Cichlidae). Heredity, 79, pp. 88–96.

Allendorf FW and SR Phelps. 1981. Isozymes and the preservation of the genetic variation in Salmonid Fish. In fish gene pools. Ryman N, Ed. Ecological Bulletin, Stockholm. 34, pp. 37–52.

Avise JC. 1994. Molecular Markers. Natural History and Evolution, Chapman and Hall, New York. pp. 1–511.

Avise JC and MJ Van den Avyle. 1984. Genetic analysis of reproduction of hybrid white bass × striped bass in the Savannah River. Trans American Fish Soc, 113, pp. 563–570.

Ballard JWO and DM Rand. 2005. The population biology of mitochondrial DNA and its phylogenetic implications. Annu Rev Ecol Evol Syst. 36, pp. 621–642.

Billington N. 2003. Mitochondrial DNA. Hallerman EM, Ed. Population genetics: principles and applications for fisheries scientists. American Fisheries Society, Bethesda, Maryland, pp. 59–100.

Billington N and PDN Hebert. 1988. Mitochondrial DNA variation in Great Lakes walleye (*Stizostedion vitreum*) populations. Can J Fish Aquat Sci, 45, pp. 643–654.

Birky CW, P Fuerst, and T Maruyama. 1989. Organelle gene diversity under migration, mutation, and drift: equilibrium expectations, approach to equilibrium, effect of heteroplasmic cells, and comparison to nuclear genes. Genetics, 121, pp. 613–627.

Boore JL. 1999. Animal mitochondrial genomes. Nucleic Acids Res, 27, pp. 1767–1780.

Broughton RE, JE Milam, and BA Roe. 2001. The complete sequence of the zebrafish (*Danio rerio*) mitochondrial genome and evolutionary patterns in vertebrate mitochondrial DNA. Genome Res, 11, pp. 1958–1967.

Burger G, GW Gray, and BF Lang. 2003. Mitochondrial genomes: anything goes. Trends Genet, 19, pp. 709–716.

Chang YS, FL Huang, and TB Lo. 1994. The complete nucleotide sequence and gene organization of carp (*Cyprinus carpio*) mitochondrial genome. J Mol Evol, 38, pp. 138–155.

Chapman RW. 1989. Mitochondrial and nuclear gene dynamics of introduced populations of *Lepomis macrochirus*. Genetics, 123, pp. 399–404.

Chapman RW and BL Brown. 1990. Mitochondrial DNA isolation methods. In: Whitmore DH, Ed. Electrophoretic and isoelectric focusing techniques in fisheries management. CRC Press, Boca Raton, Florida, pp. 107–129.

Crespi BJ and MJ Fulton. 2004. Molecular systematics of Salmonidae: combined nuclear data yields a robust phylogeny. Mol Phylogenet Evol, 31, pp. 658–679.

Doiron S, L Bernatchez, and PU Blier. 2002. A comparative mitogenomic analysis of the potential adaptive value of Arctic charr mtDNA introgression in brook charr populations (*Salvelinus fontinalis* Mitchill). Mol Biol Evol, 19, pp. 1902–1909.

Dunham RA. 2004. Aquaculture and fisheries biotechnology. Genetic approaches. CABI Publishing, Cambridge, MA, pp. 1–367.

Elmerot C, U Arnason, T Gojobori, and A Janke. 2002. The mitochondrial genome of the pufferfish, *Fugu rubripes*, and ordinal teleostean relationships. Gene, 295, pp. 163–172.

Ferguson MM, PI Ihsen, and JD Hynes. 1991. Are cultured stocks of brown trout (*Salmo trutta*) and Rainbow trout (*Oncorhynchus mykiss*) genetically similar to their source populations? Can J Fish Aquatic Sci, 48, pp. 118–123.

Garber AF and CV Sullivan. 2006. Selective breeding for the hybrid striped bass (*Morone chrysops*, Rafinesque *M. saxatilis*, Walbaum) industry: status and perspectives. Aquaculture Res, 37, pp. 319–338.

Guo X, S Liu, and Y Liu. 2006. Evidence for Recombination of Mitochondrial DNA in Triploid Crucian Carp. Genetics, 172, pp. 1745–1749.

Hallerman EM, R Dunham, and RO Smitherman. 1986. Selection or drift-isozyme allele frequency changes among channel catfish selected for rapid growth. Trans Am Fish Soc, 115, pp. 60–68.

Harrison RG. 1989. Animal mitochondrial DNA as a genetic marker in population and evolutionary biology. TREE, 4, pp. 6–11.

Hunter RL and CL Market. 1957. Histochemical demonstration of enzymes separated by zone electrophoresis in starch gels. Science, 125, pp. 1294–1295.

Inoue JG, M Miya, J Aoyama, S Ishikawa, K Tsukamoto, and M Nishida. 2001. Complete Mitochondrial DNA Sequence of the Japanese Eel, *Anguilla japonica.* Fish Sci, 67, pp. 118–125.

Inoue JG, M Miya, K Tsukamoto, and M Nishida. 2001. Complete mitochondrial DNA sequence of *Conger myriaster* (Teleostei: Anguilliformes): novel gene order for vertebrate mitochondrial genomes and the phylogenetic complications for Anguilliform families. J Mol Evol, 52, pp. 311–320.

Ishiguro N, M Miya, and M Nishida. 2001. Complete Mitochondrial DNA Sequence of Ayu *Plecoglossus altivelis.* Fish Sci, 67, pp. 474–481.

Ishiguro NB, M Miya, and M Nishida. 2003. Basal euteleostean relationships: a mitogenomic perspective on the phylogenetic reality of the Protacanthopterygii. Mol Phylogenet Evol, 27, pp. 476–488.

Johansen S, PH Guddal, and T Johansen. 1990. Organization of the mitochondrial genome of Atlantic cod, *Gadus morhua*. Nucleic Acids Res, 18, pp. 411–419.

Kaneda H, J Hayashi, S Takahama, C Taya, KF Lindahl, and H Yonekawa. 1995. Elimination of paternal mitochondrial DNA in intraspecific crosses during early mouse embryogenesis. Proc Natl Acad Sci USA, 92, pp. 4542–4546.

Koljonen ML and R Wilmot. 2005. Genetic Analysis: Allozymes. In Stock Identification Methods, Applications in Fishery Science. Cadrin SX, Friedland KD, Waldman JR, Eds. Elsevier, Amsterdam, pp. 295–309.

Labate JA. 2000. Software for population genetic analysis of molecular marker data. Crop Science, 40, pp. 1521–1527.

Liu ZJ and J Cordes. 2004. DNA marker technologies and their applications in aquaculture genetics. Aquaculture, 238, pp. 1–37.

Maoguolas A. 2005. Mitochondrial DNA. In Stock Identification Methods. Cadrin et al., Ed. Elsevier, New York, pp. 311–330.

May B. 2003. Allozyme variation. In Hallerman EM, Ed. Population genetics: principles and applications for fisheries scientists. American Fisheries Society, Bethesda, Maryland, pp. 23–36.

May B and KR Johnson. 1993. Composite linkage map of salmonid fishes (*Salvelinus*, *Salmo*, and *Oncorhynchus*), in O'Brien SJ, Ed. Genetic Maps: Locus Maps of Complex Genomes. Cold Spring Harbor, New York, 4, pp. 309–317.

Merker RJ and RC Woodruff. 1996. Molecular evidence for divergent breeding groups of walleye (*Stizostedion vitreum*) in tributaries to western Lake Erie. J Great Lakes Res, 22, pp. 280–288.

Milbury CA and PM Gaffney. 2005. Complete mitochondrial DNA sequence of the eastern oyster *Crassostrea virginica.* Mar Biotechnol, 7, pp. 697–712.

Miller MP. 1997. Tools for population genetic analysis (TFPGA) 1.3: A Windows program for the analysis of allozyme and molecular population genetic data. Distributed by the author.

Miya M, A Kawaguchi, and M Nishida. 2001. Mitogenomic Exploration of Higher Teleostean Phylogenies: A Case Study for Moderate-Scale Evolutionary Genomics with 38 newly determined complete mitochondrial DNA sequences. Mol Biol Evol, 18, pp. 1993–2009.

Miya M and M Nishida. 2000. Use of mitogenomic information in teleostean molecular phylogenetics: a tree-based exploration under the maximum-parsimony optimality criterion. Mol Phylogenet Evol, 17, pp. 437–55.

Miya M, H Takeshima, H Endo, NB Ishiguro, JG Inoue, T Mukai, TP Satoh, M Yamaguchi, A Kawaguchi, K Mabuchi, SM Shirai, and M Nishida. 2003. Major patterns of higher teleostean phylogenies: a new perspective based on 100 complete mitochondrial DNA sequences. Mol Phylogenet Evol, 26, pp. 121–138.

Morizot D, M Schmidt, and G Carmichael. 1994. Joint segregation of allozymes in catfish genetic crosses: designation of *Ictalurus punctatus* linkage group I. Trans Am Fish Soc, 123, pp. 22–27.

Morizot DC and ME Schmidt. 1990. Starch gel electrophoresis and histochemical visualization of proteins. In: Whitmore DH, Ed. Electrophoretic and isoelectric focusing techniques in fisheries management. CRC Press, Boca Raton, Florida, pp. 23–80.

Murakami M, Y Yamashita, and H Fujitani. 1998. The complete sequence of mitochondrial genome from a gynogenetic triploid 'ginbuna' (*Carassius auratus* langsdorfi). Zool Sci, 15, pp. 335–337.

Nielsen EE, MM Hansen, and K-LD Mensberg. 1998. Improved primer sequences for the mitochondrial ND1, ND3/4 and ND5/6 segments in salmonid fishes. Application to RFLP analysis of Atlantic salmon. J Fish Biology, 53, pp. 216–220.

Okumuş I and Y Çiftci. 2003. Fish population genetics and molecular markers: II. Molecular markers and their applications in fisheries and aquaculture. Turkish J Fish Aquatic Sci, 3, pp. 51–79.

Parker PG, AA Snow, MD Schug, GC Booton, and PA Fuerst. 1998. What molecules can tell us about populations: choosing and using a molecular marker. Ecology, 79, pp. 361–382.

Pasdar M, DP Philipp, and GS Whitt. 1984. Linkage relationships of nine enzyme loci in sunfishes (Lepomis; Centrarchidae). Genetics, 107, pp. 435–446.

Pasteur N, G Pasteur, F Bonhomme, J Catalan, and J Britton-Davidian. 1987. Practical Isosyme Genetics. Hellis Horwood Ltd, Chichester, UK, pp. 1–215.

Pérez LA, FM Winkler, NF Díaz, C Càrcamo, N Silva. 2001. Genetic variability in four hatchery strains of coho salmon, *Oncorhynchus kisutch* (Walbaum), in Chile. Aquaculture Res, 32, pp. 41–46.

Pruett CL, E Saillant, and JR Gold. 2005. Historical population demography of red snapper (*Lutjanus campechanus*) from the northern Gulf of Mexico based on analysis of sequences of mitochondrial DNA. Marine Biol, 147, pp. 593–602.

Ravago RG, VD Monje, and MA Juinio-Menez. 2002. Length and Sequence Variability in Mitochondrial control Region of the Milkfish, *Chanos chanos*. Mar Biotechnol, 4, pp. 40–50.
Richardson PR, BJ Baverstock, and M Adams. 1986. Allozyme Electrophoresis. A handbook for animal systematics and population studies. Academic Press, New York, NY, pp 1–410.
Robin ED and R Wong. 1988. Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells. J Cell Physiol, 136, pp. 507–513.
Rogers JS. 1972. Measures of genetic similarity and genetic distance. Studies in Genetics Univ. Texas Publ, 7213, pp. 145–153.
Rokas A, E Ladoukakis, and E Zouros. 2003. Animal mitochondrial DNA recombination revisited. Trends Ecol Evol, 18, pp. 411–417.
Schneider S, D Roessli, and L Excoffier. 2000. ARLEQUIN ver. 2000: a software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Geneva.
Schomburg I, A Chang, C Ebeling, M Gremse, C Held, G Huhn, and D Schomburg. 2004. BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res, 32, pp. D431–D433.
Shaklee JB, FW Allendorf, DC Morizot, and GS Whitt. 1990. Gene nomenclature for protein-coding loci in fish. Trans Amer Fish Soc, 119, pp. 2–15.
Smith MH and RK Chesser. 1981. Rationale for conserving genetic variation of fish gene pools. In Ryman N, Ed. Fish Gene Pools. Ecological Bulletin, Stockholm, 34, pp. 13–20.
Stoneking M, B May, and JE Wright. 1981. Loss of duplicate gene expression in salmonids: evidence for a null allele polymorphism at the duplicate aspartate aminotransferase locus in brook trout (*Salvelinus fontinulis*). Biochem Genet, 19, pp. 1063–1077.
Swofford DL and RB Selander. 1981. BIOSYS-1: a FORTRAN program for the comprehensive analysis of electrophoretic data in population genetics and systematics. J Heredity, 72, pp. 281–283.
Utter F, P Aebersold, and G Winans. 1987. Interpreting genetic variation detected by electrophoresis. In: Ryman N, Utter F, Ed. Population Genetics and Fishery Management. Washington Sea Grant Program, University of Washington Press, Seattle, USA.
Utter FM, HO Hodgins, and FW Allendorf. 1974. Biochemical genetic studies of fishes: potentialities and limitations. In: Biochemical and Biophysical Perspectives in Marine Biology, l, pp. 213–238.
Waldbieser GC, AL Bilodeau, and DJ Nonneman. 2003. Complete sequence and characterization of the channel catfish mitochondrial genome. DNA Sequence, 14, pp. 265–277.
Weir BS and CC Cockerham. 1984. Estimating F-statistics for the analysis of population structure. Evolution, 38, pp. 1358–1370.
Wilson K, V Neville, E Ballment, and J Benzie. 2000. The complete sequence of the mitochondrial genome of the crustacean *Penaeus monodon*: are malacostracan crustaceans more closely related to insects than to branchiopods? Mol Biol Evol, 17, pp. 863–874.
Wirgin I, C Grunwald, SJ Garte, and C Mesing. 1991. Use of DNA fingerprinting in the identification and management of a striped bass population in the southeastern United States. Trans American Fish Soc, 120, pp. 273–282.
Wirgin I, J Waldman, J Stabile, L Maceda, and V Vecchio. 1997. Mixed-stock analysis of Atlantic coast striped bass using mitochondrial DNA and nuclear DNA. Can J Fish Aquatic Sci, 54, pp. 2814–2826.
Wolf U, W Engel, and J Faust. 1970. Mechanism of diploidization in vertebrate evolution—coexistence of tetrasomic and disomic gene loci for isocitrate dehydrogenases in trout (*Salmo irideus*). Humangenetik, 9, p. 150.
Zardoya R, A Garrido-Pertierra, and JM Bautista. 1995. The complete nucleotide sequence of the mitochondrial DNA genome of the rainbow trout, *Oncorhynchus mykiss*. J Mol Evol, 41, pp. 942–951.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 8
Individual-based Genotype Methods in Aquaculture

Pierre Duchesne and Louis Bernatchez

Introduction

DNA marker technologies have revolutionized the way aquaculture genetics research is being conducted (Liu and Cordes 2004). Early on, most applications of molecular genetics in aquaculture relied on the estimation of demographic parameters of diversity and differentiation that were derived from averaging the genetic composition over populations or stocks. It has been recognized for nearly 25 years, however, that further knowledge of relevance for stock management and production may be obtained from the analysis of individual-based genotypic information (Smouse et al. 1982). The blooming development of new genetic markers over the last decade, namely variable number of tandem repeat loci (especially microsatellites), Amplified Fragment Length Polymorphism (AFLP), and Single Nucleotide Polymorphism (SNP) have revived a major interest in studies based on the definition of individual multilocus genotypes, and opened exciting avenues of research and applications. Basically, studies of relevance for aquaculture and based on the analysis of individual multilocus genotypes can be grouped into three broad categories of applications: parentage (including kinship), group allocation, and hybrid detection.

Parental allocation studies necessitate the assessment of parental relationships within populations, which may be achieved in various ways, including the use of exclusion probability, likelihood methods, and categorical and fractional parental assignment (reviewed in Wilson and Ferguson 2002, Jones and Ardren 2003). Parental allocation improves the efficiency of selective breeding programs in many ways, namely the following:

- establishing selected strains without having to keep families in separate tanks (Wilson and Ferguson 2002)
- investigating parent to offspring transmission of illness or parasitism
- assessing fertilization success (Selvamani et al. 2001)
- measuring reproductive success variance among breeders (Jackson et al. 2003)
- avoiding mating between closely related individuals and thus minimizing inbreeding (Ferguson and Danzmann 1998, Jackson et al. 2003, Norris et al. 2000)
- improving heritability estimates of desirable traits (Ferguson and Danzmann 1998, Vandeputte et al. 2004)
- allowing a higher rate of genetic improvement because it becomes possible to identify the progeny of parents with desirable or undesirable characteristics (Wilson and Ferguson 2002).

Studies of group allocation (also called "assignment methods") typically imply the determination of population membership of single individuals (Manel et al. 2005).

This consists of assigning an individual to the population in which its multilocus genotype has the highest probability of occurring. Such estimation may be relevant to more precisely quantify gene flow and the degree of differentiation between stocks, quantifying the admixture proportion of different stocks in a sample of individuals of unknown origin such as wild versus cultured (Miggiano et al. 2005), or enhancing traceability for trade control purposes in animals and products, and thus allowing consumers to obtain information on the origin and the production chain of food products (Liu and Cordes 2004, Hayes et al. 2005).

In aquaculture, genetic group allocation may be used to identify species or strain membership of specimens. Such identifications are useful both at the input and output end of production facilities. For instance, controlling for possible admixture in purebred populations can be done in an objective fashion when based on solid genetic data. Allocation can also reveal proportions of wild versus cultivated specimens in the marketplace or in a natural system undergoing invasion by farmed escapees or deliberately stocked by a nonnative strain. Coarse traceability can also be performed when distinct production organizations are associated with distinct strains.

Hybridization between or within species is both a common natural phenomenon and the consequence of mixing due to human-related activities, including aquaculture, and stocking of domesticated fish (Congiu et al. 2001, Vaha and Primmer 2006). Identification of hybrid individuals is often a necessary first step in the implementation of management strategies, such as breeding or translocation programs for threatened species since, it allows the removal of morphologically indistinguishable hybrid individuals from the wild population or the identification of indigenous individuals for breeding programs (Hansen 2002, Manel et al. 2005, Vaha and Primmer 2006). Early identification of hybrids may help reduce the impact of introgression between cultured and wild fish (Morizot et al. 1991, Young et al. 2001). Also identification of hybrids can impact trade by detecting hybrid production labeled as purebred, for example, sturgeon caviar (Congiu et al. 2001).

Because these issues have been treated in several recent reviews, our intent here is not to address the suitability of various molecular techniques, nor is it meant to review the empirical applications of individual-based genotype analyses. We do not wish to provide an exhaustive guide or detailed treatment to the existing analytical methods or related computer software packages. Instead, our main goal is to explain the basics of statistical principles and applications of specific methods that we have developed and applied in our laboratory over the recent years. In an attempt to render the chapter content easily accessible to the nonstatistician scientist, we have deliberately opted for verbal explanations rather than relying on the treatment of mathematical complexity and equations.

Parental Allocation

Definition and General Principles

The objective of a parental allocation process based on genetic information is to find parental genotypes corresponding to the true parents of each of a set of offspring genotypes. In some contexts, it is known in advance that the genotypes of all the parents

involved in the generation of the set of offspring are included in the collection of the putative parental genotypes. If that is the case, then the allocation system, comprising parental and offspring genotypes, is said to be closed. When some parental genotypes are missing, the allocation system is said to be open. Despite obvious similarities, the allocation problems for closed and open systems turn out to be quite different, the latter being more complex.

The two main factors affecting the performance of a parental allocation process are the number of potential parental pairs and the genetic contents of the genotypes. Performance decreases with the size of the parental set while it increases with genetic contents. Other important performance factors are the relatedness level of the parental set, accuracy of genotype scoring, and sexing of potential parents. Closely related potential parents tend to be more similar than unrelated parents resulting in a higher probability of misidentification. Whenever possible, it is generally advantageous to sex parents since this reduces by at least one half the number of potential parental pairs to be considered (Wilson and Ferguson 2002).

Markers

In theory, any type of marker can be used for performing parentage allocation. However, microsatellites are currently the most popular because of their potential for high variability even among individuals of the same strain (Liu and Cordes 2004). For instance, using eight highly variable microsatellite markers, Norris and others (2000) correctly allocated 95% of offspring from more than 12,000 potential parental pairs. Generally, codominant markers are best suited for parental allocation since allele transmission from parent to offspring is never masked by allelic dominance. The use of diploid codominant markers will be assumed throughout the following discussion.

Scoring Errors: Effects and Modeling

Here a scoring (transmission) error is defined as the result of mistaking a specific allele for another one. While scoring microsatellites, it is estimated that errors occur at a rate of 0.5 to 3%. Erroneous allocations due to scoring errors are not likely. The main negative effect of erroneous allele scores is possible loss of correct parental allocations. The probability that a genotype contains at least one scoring error increases rapidly with number of loci. Therefore, as one increases the information genetic contents by adding extra loci, one is also increasing the proportion of erroneous genotypes and thus leading to a larger proportion of incorrect allocations. This dilemma can be broken by integrating an appropriate scoring error model within the allocation process.

Within closed allocation systems, the negative effect of scoring errors can be completely neutralized by allowing a small nonzero probability estimate to the scoring of allele X as any distinct allele Y. The uniform error model (see definition below) provides such an error-catching mechanism. The transmission error probability (ε) estimate does not have to be accurate; estimates of say 1%, 2%, and 3% for ε will have the same effect on the allocation output.

The transmission error probability can be distributed in several ways over (erroneous) alleles. However, it is well known that scoring errors usually involve alleles that

are close to the true allele. This information can be fed into error modeling through the following formalization. Suppose the parental allele X is referred to as the focal allele. Then the distance between any allele Y and X can be measured in terms of number of offsets, that is, the difference between Y and X divided by the smallest allelic distance between any two alleles found in the locus (Figure 8.1). For instance, if a locus is of type tetra (nucleotide) then Y = 172 is −2 offsets away from X = 180.

The uniform error model is the simplest error model. It distributes e uniformly over all possible nonfocal alleles. Restricted error models distribute e over close neighbors of the focal allele. The examples of a ±1 offset model and a ±2 offset model are shown in Table 8.1.

Allocation Methods in Closed Systems

Basically, parental allocations can be based either on likelihood or on exclusion.

Likelihood

Given an offspring, the likelihood of a specific parental pair is essentially a measure of the probability that this pair has generated the offspring. There are three possible outputs associated with the allocation of a particular offspring. When only one parental pair has the largest likelihood, the offspring is allocated to the parental pair with the

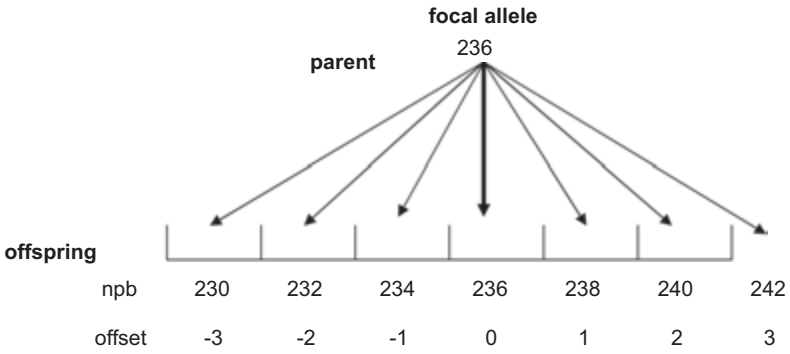


Figure 8.1. Measuring a transmission error in offset units. The distance between any allele Y and X is measured in terms of number of offsets (i.e., the difference between Y and X divided by the unit distance [smallest allelic distance between any two alleles found in the locus]).

Table 8.1. The examples of a *±1 offset model* and a *±2 offset model.*

−2 offsets	*−1 offset*	*0 offset = focal allele*	*+1 offset*	*+2 offsets*
0.002	0.01	0.98	0.01	0.002
	0.008	0.98	0.008	

largest likelihood. When several parental pairs share the largest (nonzero) likelihood, the offspring is not allocated but the output is scored as ambiguous. When all parental pairs have zero likelihood, the offspring is not allocated and the output is scored as null. Although most allocation programs do not distinguish explicitly between ambiguous and null outputs (both are scored as nonallocations), this distinction allows the computation of three system-based allocation statistics: proportions of offspring that have been scored as allocated, ambiguous, and null. These statistics turn out to be very useful in the context of the overall assessment and subsequent improvement of an allocation system. For instance, any proportion of ambiguity, except negligible, is indicative of a lack of resolution (i.e., insufficient genetic contents). In such cases, the only cure is to add one or several loci to the existing set until all ambiguity disappears.

Within closed systems, allocations should usually be performed with the uniform error model since it can absorb all kinds of errors including those generated by null alleles scored at any offset distance from the focal allele. The only drawback of the uniform model is that it may, though not necessarily, increase the proportion of offspring classified as ambiguous. This can be corrected by using a nonuniform error model but more efficiently by adding one or several loci.

Exclusion

Exclusion-based allocation is based on the idea that as information accumulates, only real parents remain after all other potential parents have turned out to be impossible candidates. Exclusion-based allocation should generally not be used in closed systems since it takes far more genetic information to exclude the set of false parents than to find the most likely pair. Unless otherwise stated, we will hereafter refer to likelihood-based allocation. Exclusion will be further discussed in the context of open system allocations.

Breeding Designs (Closed Systems)

Sometimes the offspring from blocks of breeders are put together in a single tank. Block matings generally reduce the total number of potential parental pairs as compared with allowing all adults to breed together. This reduction could translate subsequently into a reduced number of loci necessary to reach a satisfactory level of allocation correctness. Provision has been made in the last version of Package for the Analysis of Parental Allocation (PAPA) software (Duchesne et al. 2002) to allow definition of blocks of breeders reflecting breeding designs in aquaculture settings. Distinct blocks may share specimens and they may be sexed or unsexed.

Validation of Allocations in Closed Systems

Allocation to a parental pair may not always be correct. Ideally one should be able to test the correctness rate (CR), that is, the proportion of correct allocations over all

allocations, not all offspring, by checking the allocations against empirical evidence. However, under most circumstances, establishing parental connections through direct observation even in hatchery fish can prove very difficult and expensive. It is therefore customary to use simulations to estimate correctness rates. Also, simulations are useful when it comes to deciding on a set of loci sufficiently informative to reach a satisfying level of CR.

Preparental and Parental Simulations

Basically there are two types of parental allocation simulation procedures. One procedure (preparental) generates artificial parental genotypes from allelic frequencies (estimated from samples) and then artificial offspring from these parents (Figure 8.2). Another procedure (parental) uses the genotypes of real, collected parents. Preparental simulations are useful to decide on a minimal set of loci to attain the desired correctness rate even before parents and offspring have been collected. Preliminary choice of a sufficient set of loci can save lab work and resources. However, preparental simulations tend to underestimate minimal genetic information contents mainly because it generates sets of totally unrelated parents. Sets of real parents, especially when drawn from a hatchery population, may contain several subsets of highly related specimens. Therefore, it might be safer to add an extra locus to the minimal set found from preparental simulations especially when the targeted correctness level is barely reached.

To estimate correctness rates more precisely, parental simulations should be run when the set of collected parents has been genotyped. Parental simulations are not biased by the relatedness structure of the parental set.

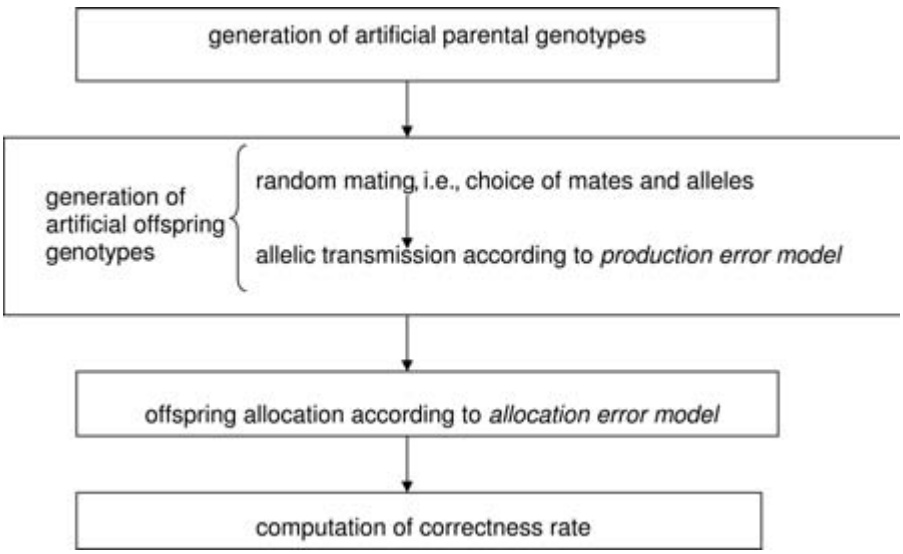


Figure 8.2. Preparental simulator procedure. The preparental procedure generates artificial parental genotypes from allelic frequencies (estimated from samples) and then artificial offspring from these parents. The parental procedure is similar except that it uses the genotypes of real, collected parents.

Production and Allocation Error Models

To estimate correctness rates more realistically, the production of artificial offspring during simulations has to mimic scoring errors. Therefore, there is a need for a production error model. The production error probabilities associated with various numbers of offsets do not have to be very accurate although they do impact on correctness rate estimations. After artificial offspring have been generated, they are processed for allocation. As with true offspring, an allocation error model is used to capture scoring errors. Ideally one should be able to define production and allocation error models separately. Allocation error models in simulations should generally be the same as the one used in allocating real offspring.

Likelihood and Exclusion Methods in Open Systems

Likelihood

Allocation in open systems poses a double problem (i.e., identify true parents that belong to the collected parental set and identify uncollected parents as uncollected). Likelihood-based allocation can be very efficient in solving the collected parent problem but is liable to mistake an uncollected parent for a collected one (i.e., overallocate). Overallocation increases sharply with the proportion of uncollected parents. With more uncollected parents, there is a higher probability that collected specimens are sufficiently similar to uncollected parents to become likely candidates for (erroneous) allocation. This problem is more acute with methods allowing a nonzero probability for any kind of scoring error, which translates into nonzero likelihood for all possible parental-offspring genotype combinations. The overallocation probability can only be assessed when a reasonably accurate estimate of the missing part of the parental set is available (Wilson and Ferguson 2002). Unfortunately, likelihood-based allocation cannot provide such an estimate on the basis of the available genotypes. In short, likelihood methods in open systems tend toward overallocation, the extent of which cannot be safely estimated without a (generally lacking) reliable estimate of the uncollected portion of the parental set.

Exclusion

The drawbacks of likelihood-based methods in open systems have led some researchers to resort to the exclusion allocation method. This method essentially compares the genotype of each potential parent with that of the offspring. Parental genotypes are excluded as soon as both offspring alleles are absent on a single locus of the parental genotype. In addition, no more than two nonexcluded parental genotypes have to remain for the allocation to be performed. The idea is that, given enough loci, nonparental collected specimens will eventually be excluded on at least one locus.

The exclusion method has several drawbacks. It is very costly in terms of genetic information since most excluded candidates would have been discarded on account of

low likelihood based on much less powerful sets of loci. Since scoring errors are more numerous with each additional locus (Jones and Ardren 2003), it is plausible that a substantial number of genotypes will contain at least one error. Such errors are very likely to provoke the loss of one or several allocations especially when parental genotypes are erroneous. Some researchers have suggested tolerance for mismatches not exceeding a predetermined number. However, mismatches may also come from a truly nonparental genotype. Therefore, this less stringent version of exclusion, while it does reduce the probability of erroneous exclusion, also increases the probability of retaining nonparental combinations (i.e., erroneous allocations). This tradeoff between two types of errors cannot be easily assessed in the absence of a sound estimate for the missing proportion of uncollected parents. Therefore, the choice of a number of tolerated mismatches is largely arbitrary.

To alleviate the stringency of the exclusion method resulting in overexclusion, another approach is sometimes used that includes rescoring a nearly perfectly matching genotype. The idea is to see if some scoring error might not be the reason for missing an allocation by so little. Although this method does make some sense, it is prone to self persuasion and is certainly not amenable to correctness analysis. Briefly stated, exclusion methods tend to miss sizable numbers of true parents and do not lend themselves to rigorous evaluations of correctness rates. They would be efficient if based on a very informative set of loci and extremely accurate genotypes. These two conditions are not generally met except in some forensic contexts.

The PASOS Approach (Open Systems)

Likelihood-based methods lean toward overallocation whereas exclusion methods tend to overexclude (i.e., eliminate true parents). The PASOS software (Duchesne et al. 2005) uses a mixed approach by first picking up the most likely parental pair(s) among all potential pairs based on a uniform scoring error model that ensures that at least one most likely pair is listed. When several most likely pairs are found, the first one in the list is retained. Then an extended exclusion method is applied to the two genotypes of the retained most likely parental pair.

Extended Exclusion Method

The extended exclusion method used by PASOS compares each of the locus genotypes of the two putative parents together with that of the offspring. From these three genotypes, a transmission scenario (Figure 8.3A) is built that associates each offspring allele to a parental allele. Such scenarios are built from a set of rules that aims at restoring the most probable allelic transmission pattern, taking the three genotypes together. Once the two most likely parent-to-offspring allele pairs have been determined, the distance in offset units is computed for each pair. Any allelic distance exceeding the maximum offset tolerance (MOT) specified by the user provokes the exclusion of the corresponding putative parent (Figure 8.3B). Therefore, there may be zero, one, or two parents excluded at each locus. It suffices that the offset tolerance be exceeded on a single locus for the putative parent, relative to the offspring currently processed, to be discarded.

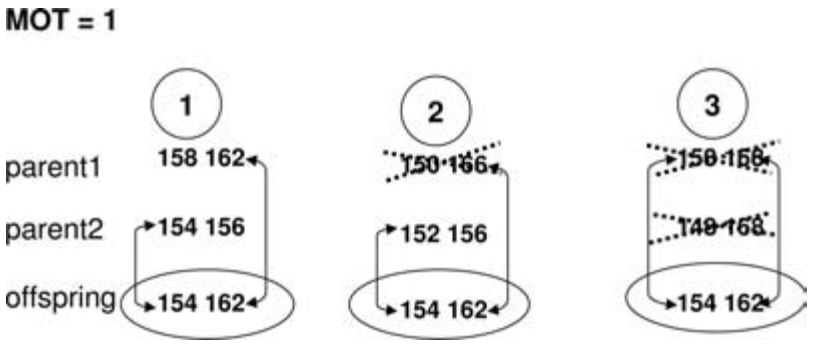


Figure 8.3A. Allelic transmission scenarios: Allelic transmission scenarios are built from a set of rules that aims at restoring the most probable allelic transmission pattern, taking the two parental and the offspring genotypes simultaneously into account.

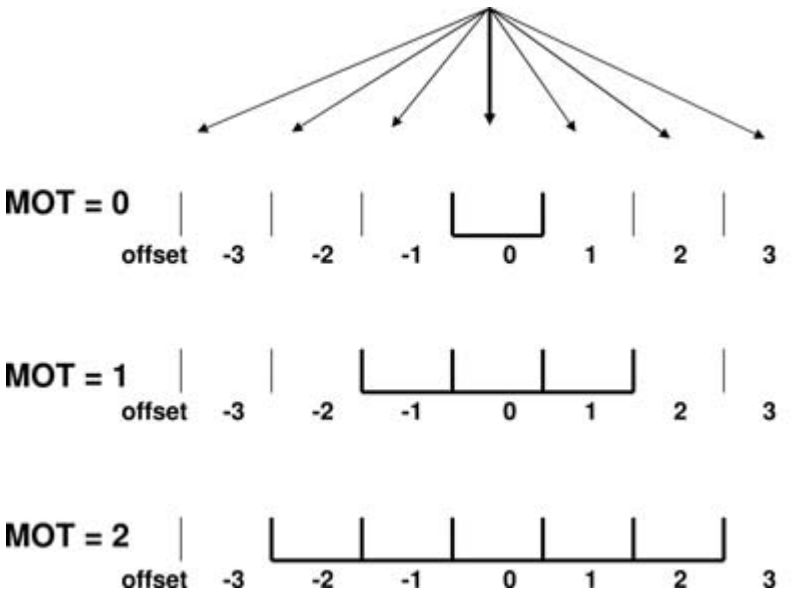


Figure 8.3B. Tolerance net as defined by MOT: Any allelic distance exceeding the maximum offset tolerance (MOT) specified by the user provokes the exclusion of the corresponding putative parent.

Rationale

The two-step allocation approach implemented in PASOS is based on the following rationale. If the two real parents of an offspring belong to the collected set of potential parents, the probability that they will be selected during the likelihood phase will increase with genetic information contents (i.e., with number of loci). If they have been genotyped with scoring errors within the bounds of the maximum offset tolerance,

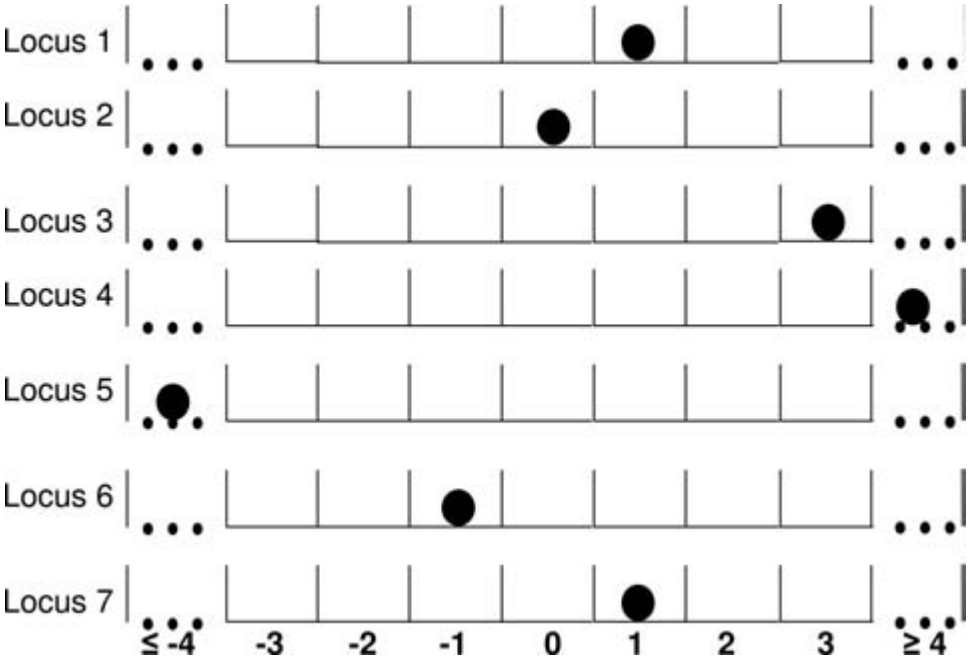


Figure 8.4. Extended exclusion of a false parent. The probability that a nonparental member of the most likely parental pair be eliminated increases with number of loci.

then they will most probably not be discarded during the exclusion phase. If only one parent belongs to the parental set, it will probably be part of each of the most likely pair(s) and thus of the first pair listed. The probability that the nonparental member of the most likely pair be eliminated during the exclusion phase will increase with number of loci (Figure 8.4). If none of the two parents belongs to the set of collected parents, then the most likely pair will contain false parents both of which will eventually be discarded as the number of loci increases.

Sequence Allocation (Allocation) and Proportion of Missing Parents

When PASOS is run sequentially with one, two, three, etc., loci from the allocation set, it makes less and less allocations and eventually reaches a stable or near stable proportion of allocations (Figure 8.5). This happens when false parents have been purged by the extended exclusion procedure. The remaining proportion of allocations may then be taken as an estimate of the proportion of missing parents. The precision of the latter estimate depends on the assumption that the collected parental set comprises specimens that have truly participated, no matter how successfully, in the breeding event at the origin of the offspring sample. If the parental set is inflated with individuals not involved in reproductive events, then the number of missing parents will likely be overestimated. Clearly, precision of the estimate should increase with the size of the offspring sample. The estimated number of missing parents must be fed into simulation runs to obtain estimates of the correction rates.

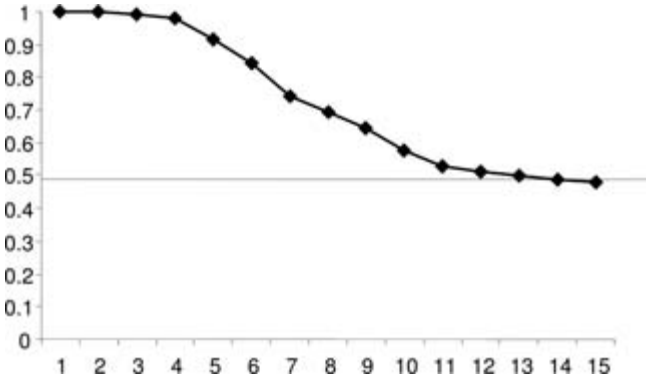


Figure 8.5. Sequence allocation curve. When PASOS is run sequentially with one, two, three . . . loci from the allocation set, it makes less and less allocations and eventually reaches a stable or near stable proportion of allocations which the user may then use to estimate the number of uncollected parents.

Automatic sequence allocation (i.e., with one, two, three or more loci) is implemented in PASOS.

Due to its use of restricted error modeling, PASOS should only be used when scoring is of good quality (i.e., does not generally exceed two offsets from focal alleles). Also, the set of loci should be tested for the presence of null alleles and all loci suspected of containing null alleles should be dropped.

Validation of Allocations in Open Systems

The estimation of the correctness rate within any open allocation system depends heavily on the estimated number of missing parents. In fact, the larger the set of missing parents, the higher the probability that some of their offspring will be mistaken for offspring from the collected parental set. Unfortunately, the number of missing parents often is difficult to estimate under most settings and so estimates have typically been guessed in the past.

However, recent developments in allocation techniques that combine likelihood with exclusion approaches (PASOS) now make it possible to obtain reliable estimates of the missing part of the parental set. Once the sequence allocation of the sample of real offspring has produced a (nearly) stable allocation rate curve, an estimate of the proportion of missing parents is available. The latter can then be fed into parental simulations for obtaining a sound estimate of the correctness rate associated with the specific allocation system.

Preparental simulations should be run whenever possible to find minimal sets of loci. Since the missing part of the parental set cannot be estimated genetically prior to parent collecting, care should be taken to use both optimistic and pessimistic scenarios corresponding to lower and higher proportions of missing parents, respectively. Again, minimal sets of loci should preferably be complemented by an extra locus in case the real parental set comprises highly related specimens.

Features to Look for in Parentage Allocation Programs

In closed as well as open allocation systems, programs should provide simulation facilities. Simulations are usually the only way to obtain a sound estimate of the correctness rate or accuracy of the system (i.e., the proportion of correct allocations among all allocations). In addition, one should be able to run the simulator on both preparental and parental modes. One should be able to run programs either with sexed or unsexed parental sets since sexing in fish cannot always be done easily and reliably.

Closed Systems

In closed systems, programs should provide distinct statistics for ambiguous and null outputs. The proportion of ambiguous outputs is a direct measure of the capacity of the set of loci to perform the allocation task under way. An error model that provides nonzero probability for any possible scoring error such as the uniform error model should suffice under most circumstances. However, with reliable scoring and absence of null alleles, the use of a restricted error model allowing for a limited number of error offsets could save on the number of loci without significantly reducing the number of allocations. A mechanism for defining blocks of breeders reflecting breeding designs in aquaculture settings is desirable. Block definition can increase resolution power of a set of loci and reduce the probability of incorrect allocations.

Open Systems

In open systems, uniform error modeling can lead to overallocation since parent-offspring mismatches can also originate from an incorrect allocation. On the other hand, zero error tolerance is very likely to provoke losses of allocations especially as the number of loci is increased. Restricted error modeling is a means to distinguish between scoring errors and erroneous allocations without dropping a significant proportion of true parents. Restricted error modeling is currently implemented in PASOS. The most important features for parental allocation programs are described in Figure 8.6.

Some Available Programs

Some of the currently available programs with respective allocation methods follow:

- CERVUS (Marshall et al. 1998) (likelihood)
- FAMOZ (Gerber et al. 2002) (likelihood)
- KINSHIP (Goodnight and Queller (1999) (exclusion), (Danzmann 1997) (exclusion)
- NEWPAT (Wilmer et al. 1999) (exclusion)
- PAPA (Duchesne et al. 2002) (likelihood/closed systems)
- PARENTE (Cercueil et al. 2002) (likelihood)
- PASOS (Duchesne et al. 2005) (likelihood + extended exclusion/open systems)

All of these freely available programs can be downloaded at http://www.bio.ulaval.ca/louisbernatchez/links.htm.

General

Both pre-parental and parental simulations are available.

Simulations and allocations may be run with sexed or unsexed parental sets.

Closed systems

Distinction is made between *null* and *ambiguous* non-allocation statistics.

Scoring error may be distributed over all non-focal alleles, e.g., uniformly.

Parental files can be structured according to block mating designs.

Restricted error models may be user-defined.

Open systems

Restricted error models are available and user-defined.

A means to estimate the number of uncollected parents is provided.

Figure 8.6. A list of the most important features for parental allocation programs.

Group Allocation (Species, Population, or Strain Identification)

Definition and General Principles

Species, population, or strain identification of individuals on the basis of genetic data is technically the same and will hereafter be referred to as group allocation. Only those allocation situations will be considered where each purebred group has been sampled so that fairly accurate estimates of allelic frequencies for each genotyped locus and each purebred group are available (baseline samples).

Some recent developments aim at allocating individuals from mixed samples without prior sampling of group purebreds. Those so-called clustering techniques essentially tend to partition a given mixed sample into subsamples to minimize (or maximize) some statistic associated with population structuring (e.g., linkage disequilibrium). Allocation from good baseline samples produces verifiable results within a small fraction of the computation time required from clustering methods. Moreover, currently used clustering methods tend to perform poorly when group differentiation is weak (Waples and Gaggiotti 2006), a very serious handicap when it comes to strain identification. Finally, they do not provide ad hoc means to estimate the accuracy of their allocations and involve considerable uncertainty (Manel et al. 2005). Given the above drawbacks of clustering methods, they will not be discussed any further since baseline samples are available for most group allocation tasks within aquaculture settings.

The idea underlying group allocation of an individual genotype (G) is rather simple. In its simplest version, the probability (likelihood) that G could be found within a group is computed for each possible group and then G is allocated to the group with highest probability. Since such probabilities are often very small, they are usually expressed in log10 format and comparisons between two populations as log-likelihood ratios. For example, if G is 1,000 times more likely to be found within

population A than it is within population B, the log-likelihood ratio of A relative to B is equal to three.

Within a given allocation task, a minimal log-likelihood ratio (threshold) between the most likely and the next most likely group is defined. If the threshold is not reached for G, it is simply not allocated and classified as nonallocated. For instance, a log-likelihood threshold of two would mean that no individual genotype should be allocated if it is not at least 100 times more probable within the most probable group. The log-likelihood threshold turns out to be an important allocation parameter. Generally, raising the threshold increases the probability of allocating correctly (accuracy) but decreases the number of genotypes being allocated (allocation rate). Care should be taken to choose an appropriate threshold for the task under way.

Another important aspect of group allocation is the question of ghost groups (i.e., groups that have not been sampled as purebreds since they have not yet been identified but which may be represented within the sample of individuals to be allocated). Ghost groups are much more likely when allocations involve wild populations. When it is suspected that ghost groups might exist, one should test whether G might not belong to such an external yet undefined group. This can be done through an exclusion procedure based on membership P values computed from simulations (see the Simulations section).

Markers

As in parentage allocation, any type of marker (RFLP, RAPD, AFLP, microsatellite) can be used for performing group allocations. However, very high polymorphism (number of alleles/ locus > 10) does not add substantial allocation resolution when compared to less variable loci. Here, the most important characteristic of a set of loci is sheer number (Ferguson and Danzmann 1998, Bernatchez and Duchesne 2000, Hayes et al. 2005). Therefore, when it comes to distinguishing several weakly differentiated groups (e.g., strains), markers available in virtually unlimited numbers are the best candidates even when each locus has low information content. For such heavy allocation tasks, AFLP markers are currently the most appropriate choice except when a sufficient set of microsatellites already exists (Campbell et al. 2003).

Scoring and Sampling Errors

With dominant markers such as AFLP, allele should be taken as an equivalent for presence/absence in the following discussion. Generally speaking, scoring errors within their usual range (0.5 to 3%) have little impact on group allocation. However, special care should be taken when scoring purebred samples especially when small (>20). As a rule of thumb, purebred samples should contain at least 20, but preferably 30, specimens to obtain reasonably accurate frequency estimates (Ruzzante 1998). Smaller samples might still be used especially when dealing with highly differentiated groups. When using highly polymorphic microsatellite loci with large numbers (>15) of low frequency alleles, sample sizes should be increased accordingly (e.g., to 50 specimens). Note that the low frequency of an allele can suddenly double following sampling of a

single extra copy (Roques et al. 1999). To obtain truly representative purebred samples, sampling should be done as randomly as possible. In particular, overrepresentation of specific families should be avoided.

A special sampling problem arises when some allele is totally absent from one or several purebred samples while present in other purebred samples or the (mixed) sample to be allocated. Customarily, the frequency of a missing allele within a purebred sample was estimated at $1/(N+1)$ (N = number of scored alleles within sample). This amounts to the expectation that the next allele would be the missing one (maybe-next-allele formula). Another approach consists of fixing the missing allele frequency at some user-defined low value (e.g., 0.01). Practically, missing allele frequency estimates have little impact on the result of an allocation task. If one favors the fixed low value approach, then this value may be seen as an allocation parameter and its value may be chosen to maximize the correct reallocation rate.

Validation of Group Allocations

The accuracy of group allocations, that is, the estimated proportion of correct allocations over all allocations (excluding non-allocated specimens), can be assessed through reallocation and simulation procedures (Figure 8.7).

Reallocation

The reallocation procedure allocates the purebred specimens among the candidate groups as if their group membership were unknown. The latter condition means that each time a purebred specimen is (re-)allocated, the allelic frequencies of its group are recalculated as if it did not belong. This precaution aims at eliminating the bias resulting from the specimen actually weighing on frequency estimates and, as a consequence, artificially increasing the probability of being allocated to its proper group. These frequency recalculations are usually referred to as the leave-one-out procedure.

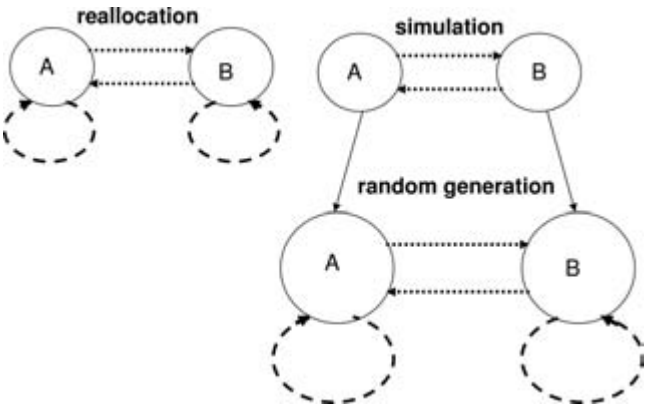


Figure 8.7. Validation procedures in group allocation. The reallocation procedure allocates the purebred specimens among the candidate groups as if their group membership were unknown. To estimate accuracy from simulations, artificial specimens are generated randomly, based on the allelic frequencies derived from purebred samples.

Reallocation of purebreds is usually a very reliable way of estimating accuracy. One important advantage of reallocation oversimulations is that it takes scoring errors automatically into account. On the other hand, accuracy estimates from reallocation may be biased upward when purebred samples include highly inbred specimens (e.g., full and half-siblings). Thus, the quality of accuracy estimates from reallocation is somewhat sensitive to the quality of purebred group samples. Low reallocation rates may result from very poor scoring, a lack of resolution due to poor genetic content relative to group differentiation, or even from an absence of real differentiation (i.e., from samples not actually representing distinct biological entities).

Simulations

Estimations of accuracy can also be obtained from simulations. Artificial specimens are generated randomly, based on the allelic frequencies derived from purebred samples. The simulators currently built into population (group) allocation programs do not allow mimicking of scoring errors. Consequently, accuracy may sometimes be slightly overestimated from simulations since scoring errors do increase the probability of misallocating real genotypes. One important advantage of simulations over reallocation is their potential for spanning a very large range (e.g., tens of thousands of possible genotypes from each group). Therefore, genotypes from prospective mixed samples get a more complete coverage by simulations than they do from reallocation.

Besides accuracy estimations, simulations are sometimes used to obtain likelihood distributions from each purebred sample. Each group likelihood distribution is obtained by producing a large number of artificial genotypes, based on the group allelic distributions, and then the likelihoods associated with the genotypes. Thereafter, the group-specific likelihood distributions may be used to produce a group membership P value for each genotype of a mixed sample. Some allocation programs actually use group membership P values by excluding each candidate group with membership P value lower than a predefined threshold. When the allocation procedure is based on likelihood ratios, membership P values can still be useful to detect ghost groups: when membership P values are very low (e.g., <0.001) for all potential groups considered, the presence of at least one ghost group may be suspected.

Another usage of simulations is the adjustment of the likelihood ratio allocation threshold. Sometimes a proportion of artificial genotypes are misallocated indicating that there is a nonnegligible probability that real genotypes may also be misallocated. This problem can be solved to a large extent by raising the likelihood ratio allocation threshold until misallocation of simulated genotypes vanishes. Note, however, that this will generally be associated with a rise in the proportion of nonallocated real and simulated genotypes.

Reallocation Versus Simulation Accuracy Estimates

Accuracy estimates from reallocation and simulations should be close. However, if the estimated accuracy from reallocation is substantially lower than that from simulations, it is probably due to unusually numerous scoring errors. On the other hand, higher accuracy estimates from reallocation could reflect highly inbred portions of samples (e.g., families).

Features to Look for in Group Allocation Programs

Reallocation of purebred genotypes, allocation of mixed samples, and simulations are the three basic procedures that should be provided by group allocation programs. The leave-one-out procedure should be used in reallocating purebred samples.

The log-likelihood ratio allocation threshold should be user defined. Calculation of membership P values for each genotype should be possible even when the allocation procedure is based on likelihood ratio values (i.e., not on low P value exclusion). Membership P values are especially important when there are grounds to believe that some members of the mixed sample may come from a ghost group. Group log-likelihoods for each real genotype should be made available to the user rather than just the allocation or nonallocation decision. Preferably, the user should be able to choose missing allele frequency values either as constants or as the classical maybe-next-allele formula.

Some Available Programs

Currently the three most widely used programs for group allocation based on purebred genotype samples are GENECLASS2 (Piry et al. 2004), WHICHRUN (Banks and Eichert 2000) for codominant markers (microsatellites), and AFLPOP (Duchesne and Bernatchez 2002) for dominant markers (AFLP). These freely available programs can be downloaded at http://www.bio.ulaval.ca/louisbernatchez/links.htm.

Specifics of Hybrid Identification

Definition and General Principles

Hybrids may involve two distinct species, two strains, or two populations within a single species. Genetic identification of either type of hybrids is technically the same problem. However, intraspecific hybrids are typically more difficult to detect due to less genetic differentiation and therefore require considerably more information (i.e., more genotyped loci). Given two source breeds/species, a diagnostic allele (presence/absence) is one that has 100% frequency within one breed and 0% frequency in the other breed/species. Historically, genetic identification of hybrids was associated with the simultaneous presence of diagnostic alleles (presence/absence) of both source breeds/species within a single genotype (Morizot et al. 1991). Indeed genotypes with diagnostic alleles of mixed origin are easily observable and, without any calculation, can be safely attributed to hybridization assuming no other breed/species has contributed to the purported hybrid's genotype. The 100% versus 0% frequency diagnostic criterion has been somewhat relaxed in recent literature and loci with an allele differing by >99% have sometimes been considered diagnostic (Young et al. 2002). However, there has been an increasing awareness that all loci showing a frequency difference beyond sampling error could contribute to distinguish between purebreds and hybrids (Bjornstad and Roed 2002). Even though loci with 10%

frequency differential, for example, have far less hybrid detection power than diagnostic loci, they can still be cumulated to attain any power level.

Hybrid Identification as Group Identification: the Virtual-Hybrid-Group Method

Thus, hybrid identification is technically the same problem as group identification except that preidentified samples of hybrids are usually not available as one of the potential allocation groups. However, F1 hybrid allelic frequency distributions can be directly computed from purebred frequencies, say f1 and f2. For codominant loci such as microsatellites, a straightforward estimate of any hybrid allelic frequency fh is simply the average (f1 + f2)/2 of the two purebred frequencies. For dominant markers (e.g., aflp), fh = 1 − sqrt (1 − f1)*sqrt (1 − f2). This means that purebred samples are sufficient for allocation tasks including purebred and F1 hybrid groups. Again, sets of nondiagnostic loci can be used successfully for hybrid detection. Following the same idea, purebred samples also suffice to identify second-generation hybrids (F2 and backcrosses).

Special Sampling Care

Although hybrid identification is technically the same as any other type of group allocation, it requires special sampling care for two reasons. First, differentiation is weaker between F1 hybrids and purebreds than between two distinct purebreds. Second, since allelic frequencies are computed from the two purebred frequency estimates, sampling errors in the latter will be passed along to the hybrid estimates. Consequently, when hybridization is suspected, sample sizes should be increased (>30), sampling performed as randomly as possible and alleles (or presence/absence in the case of AFLP) scored with extra precaution. Clearly, all of the above is even more important when second generation hybrids are considered (Epifanio and Phillipp 1997).

Efficiency and Accuracy in Hybrid Identification

There are two ways to look at the performance of a hybrid identification procedure. One important measure is the probability that, given a specimen classified as hybrid, this specimen is in fact a hybrid. Another important measure is the probability that, given a true hybrid, it was classified (allocated) as a hybrid. Following Vähä and Primmer (2006), we use the words accuracy and efficiency to denote the first and second of these two measures, respectively. The product of these two measures can be seen as the overall performance of the hybrid identification procedure.

If the likelihood distribution for purebreds and hybrids are not (nearly) perfectly disjoint, then there is an unavoidable tradeoff between accuracy and

Low accuracy and high efficiency hybrid ID

	number of	specimens	among
allocated to	popA	popB	popA X popB
popA	38	0	0
popB	0	35	0
popA X popB	12	15	50
None	0	0	0

High accuracy and low efficiency hybrid ID

	number of	specimens	among
allocated to	popA	popB	popA X popB
popA	35	0	0
popB	0	38	0
popA X popB	0	0	11
None	15	12	39

LOD THRESHOLD

Figure 8.8. The accuracy versus efficiency tradeoff in hybrid identification. One way to strike the desired balance between accuracy and efficiency is to fix the log-likelihood allocation threshold by running allocation simulations. Raising the LOD threshold generally decreases efficiency while increasing accuracy.

efficiency (Figure 8.8). Some users will prefer to make sure that any possible hybrid be identified (i.e., to raise the efficiency component). For instance, when there exists independent data bearing on intermediate morphological traits, uncertain hybrid genetic classification may be used in a cross-validation fashion. On the other hand, in the absence of any control data and especially when there is only a suspicion that hybrid specimens might exist, it is preferable to obtain highly confident hybrid detection (i.e., enhance the accuracy component of performance). One way to strike the desired balance between accuracy and efficiency is to fix the log-likelihood allocation threshold by running allocation simulations. For instance, raising the threshold sufficiently will virtually eliminate false hybrid classification (i.e., accuracy will become close to 100%). Of course, this will be at the expense of a higher rate of nonallocations of both purebreds and hybrids.

So far, we have discussed hybrid identification based on purebred samples. However, as with general group allocation procedure, there exist clustering methods for

hybrid identification. Two such methods have been implemented in STRUCTURE (Pritchard et al. 2000) and NEWHYBRIDS (Anderson and Thompson 2002) and have been recently assessed by Vähä and Primmer (2006). It was found that both programs, unless run with very large numbers (n = 48) of codominant loci, showed high rates of misclassification of purebred as F1 hybrids even with moderately high Fst (0.12). Also backcrosses were often misclassified as purebred. Briefly, there are accuracy and efficiency problems with currently available programs performing hybrid allocation without baseline samples. Unfortunately, these methods do not provide any inbuilt mechanism, such as simulation tools, to assess the accuracy and efficiency levels associated with the user's own specific data. Therefore, it is usually much safer in hybrid studies to rely on good quality samples of purebred groups.

Markers

In principle, any type of marker (RFLP, RAPD, AFLP, microsatellite, SNP) can be used for performing hybrid identification. However, correct detection of hybrids takes more genetic information and so, roughly speaking, more loci than allocation of purebred specimens. This is especially true when purebred individuals belong to distinct, but weakly differentiated, strains. Detection of intraspecific hybrids necessitates large numbers of loci and so AFLP markers should be considered until SNP markers can be obtained in large numbers and analyzed at low cost in nonmodel species.

Available Programs

The virtual-hybrid-group method based on purebred samples has been implemented in AFLPOP (Duchesne and Bernatchez 2002) for dominant markers (AFLP). NEWHYBRIDS (Anderson and Thompson 2002) and STRUCTURE (Pritchard et al. 2000) are additional software that provides posterior distribution that shows that individuals fall into different hybrid categories between populations using dominant or codominant markers. These programs can be downloaded at http://www.bio.ulaval.ca/louisbernatchez/links.htm.

Conclusion

The current context in the applications of molecular genetic techniques, particularly as pertaining to individual-based genotype analyses, is extremely positive. There is a wealth of powerful genetic markers that are being developed for an increasing number of cultured species, both vertebrates and invertebrates, and many efficient analytical tools are readily accessible, free of charge for the most part. It is our hope that we have provided a better understanding of the principles underlying some of the most versatile methods currently available for performing parentage, strain/population assignment, and hybrid analyses, as well as useful guidelines for choosing proper efficient analytical software.

Acknowledgments

We thank Dr. John Liu for kindly inviting us to contribute to this book. L.B.'s research is financially supported by FQRNT (Québec), NSERC (Canada), and a Canadian Research Chair in Genomics and Conservation of Aquatic resources.

References

Anderson EC and EA Thompson. 2002. A model-based method for identifying species hybrids using multilocus genetic data. Genetics, 160, pp. 1217–1229.

Banks MA and W Eichert. 2000. WHICHRUN (version 3.2): A computer program for population allocation of individuals based on multilocus genotype data. J Hered, 91, pp. 87–89.

Bernatchez L and P Duchesne. 2000. Individual-based genotype analysis in studies of parentage and population allocation: how many loci, how many alleles? Can J Fish Aquat Sc, 57, pp. 1–12.

Bjornstad G and KH Roed. 2002. Evaluation of factors affecting individual allocation precision using microsatellite data from horse breeds and simulated breed crosses. An Gen, 33, pp. 264–270.

Campbell D, P Duchesne, and L Bernatchez. 2003. AFLP utility for population allocation studies: analytical investigation and empirical comparison with microsatellites. Mol Ecol, 12, pp. 1979–1992.

Cercueil A, E Bellemain, and S Manel. 2002. PARENTE: Computer program for parentage analysis. J Hered, 93, pp. 458–459.

Congiu L, I Dupanloup, T Patarnello, F Fontana, R Rossi, G Arlati, and L Zane. 2001. Identification of interspecific hybrids by amplified fragment length polymorphism: the case of sturgeon. Mol Ecol, 10, pp. 2355–2359.

Danzmann RG. 1997. PROBMAX: A computer program for allocating unknown parentage in pedigree analysis from known genotypic pools of parents and progeny. J Hered, 88, pp. 333.

Duchesne P and L Bernatchez. 2002. AFLPOP: a computer program for simulated and real population allocation, based on AFLP data. Mol Ecol Notes, 2, pp. 380–383.

Duchesne P, T Castric, and L Bernatchez. 2005. PASOS (parental allocation of singles in open systems): a computer program for individual parental allocation with missing parents. Mol Ecol Notes, 5, pp. 701–704.

Duchesne P, MH Godbout, and L Bernatchez. 2002. PAPA (package for the analysis of parental allocation): a computer program for simulated and real parental allocation. Mol Ecol Notes, 2, pp. 191–193.

Epifanio JM and D Phillipp. 1997. Sources for misclassifying genealogical origins in mixed hybrid populations. J Hered, 88, pp. 62–65.

Ferguson MM and RG Danzmann. 1998. Role of genetic markers in fisheries and aquaculture: useful tools or stamp collecting? Can J Fish Aquat Sc, 55, pp. 1553–1563.

Gerber S, S Mariette, R Streiff, C Bodénès, and A Kremer. 2000. Comparison of microsatellites and amplified fragment length polymorphism markers for parentage analysis. Mol Ecol, 9, pp. 1037–1048.

Goodnight KF and DC Queller. 1999. Computer software for performing likelihood tests of pedigree relationship using genetic markers. Mol Ecol, 8, pp. 1231–1234.

Hansen MM. 2002. Estimating the long-term effects of stocking domesticated trout into wild brown trout (*Salmo trutta*) populations: an approach using microsatellite DNA analysis of historical and contemporary samples. Mol Ecol, 11, pp. 1003–1015.

Hayes B, AK Sonesson, and B Gjerde. 2005. Evaluation of three strategies using DNA markers for traceability in aquaculture species. Aquaculture, 250, pp. 70–81.

Jackson TR, DJ Martin-Robichaud, and ME Reith. 2003. Application of DNA markers to the management of Atlantic halibut (*Hippoglossus hippoglossus*) broodstock. Aquaculture, 220, pp. 245–259.

Jones AG and WR Ardren. 2003. Methods of parentage analysis in natural populations. Mol Ecol, 12, pp. 2511–2523.

Liu ZJ and JF Cordes. 2004. DNA marker technologies and their applications in aquaculture genetics. Aquaculture, 238, pp. 1–37.

Manel S, OE Gaggiotti, and RS Waples. 2005. Allocation methods: matching biological questions with appropriate techniques. TREE, 20, pp. 136–142.

Marshall TC, J Slate, LEB Kruuk, and JM Pemberton. 1998. Statistical confidence for likelihood-based paternity inference in natural populations. Mol Ecol, 7, pp. 639–655.

Miggiano et al. 2005. AFLP and microsatellites as genetic tags to identify cultured gilthead seabream escapees: data from a simulated floating cage breaking event. Aqua Intern, 13, pp. 137–148.

Morizot DC, SW Calhoun, LL Clepper, and ME Schmidt. 1991. Multispecies hybridization among native and introduced centrarchid basses in central Texas. Trans Am Fish Soc, 120, pp. 283–289.

Norris AT, DG Bradley, and EP Cunningham. 2000. Parentage and relatedness determination in farmed Atlantic salmon (*Salmo salar*) using microsatellite markers. Aquaculture, 182, pp. 73–83.

Piry S, A Alapetite, JM Cornuet, D Paetkau, L Baudouin, and A Estoup. 2004. GENECLASS2: A software for genetic allocation and first-generation migrant detection. J Hered, 95, pp. 536–539.

Pritchard JK, M Stephens, and P Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics, 155, pp. 945–959.

Roques S, P Duchesne, and L Bernatchez. 1999. Potential of microsatellites for individual assignment: the North Atlantic redfish (genus *Sebastes*) species complex as a case study. Mol Ecol, 8, pp. 1703–1717.

Selvamani MJP, A Sandie, and M Degnan. 2001. Microsatellite Genotyping of Individual Abalone Larvae: Parentage Assignment in Aquaculture. Mar Biotech, 3, pp. 478–485.

Smouse PE, RS Spielman, and MH Park. 1982. Multiple-locus allocation of individuals to groups as a function of the genetic variation within and differences among human populations. Am Nat, 119, pp. 445–463.

Vähä JP and CR Primmer. 2006. Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. Mol Ecol, 15, pp. 63–72.

Vandeputte M, M Kocour, S Mauger, M Dupont-Nivet, D De Guerry, M Rodina, D Gela, D Vallod, B Chevassus, and O Linhart. 2004. Heritability estimates for growth-related traits using microsatellite parentage assignment in juvenile common carp (*Cyprinus carpio L.*). Aquaculture, 235, pp. 223–236.

Waples RS and O Gaggiotti. 2006. What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity, 15, pp. 1419–1439.

Wilmer JW, PJ Allen, PP Pomeroy, SD Twiss, and W Amos. 1999. Where have all the fathers gone? An extensive microsatellite analysis of paternity in the grey seal (*Halichoerus grypus*). Mol Ecol, 8, pp. 1417–1429.

Wilson AJ and MM Ferguson. 2002. Molecular pedigree analysis in natural populations of fishes: approaches, applications, and practical considerations. Can J Fish Aquat Sc, 59, pp. 1696–1707.

Young WP, CO Ostberg, P Keim, and GH Thorgaard. 2001. Genetic characterization of hybridization and introgression between anadromous rainbow trout (*Oncorhynchus mykiss irideus*) and coastal cutthroat trout (*O-clarki clarki*). Mol Ecol, 10, pp. 921–930.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 9
Application of DNA Markers for Population Genetic Analysis

Eric M. Hallerman, Paul J. Grobler, and Jess W. Jones

Introduction

Remarkable progress has been made in the past 20 years in the ability to detect DNA-level genetic variations. The development of polymerase chain reaction (PCR), automated sequencing, and various marker systems (see Chapters 2–7) has provided unlimited availability of genetic markers. Much progress has also been made in the area of statistical methods for data analysis (Luikart and England 1999, Nei and Kumar 2000). At the same time, availability of high-performance microcomputers, development of software packages for powerful application of population genetic and phylogenetic analysis, and ready access to the software through the Internet have placed rigorous analysis of molecular data within the reach of any research group. These advances in molecular and computational biology have been applied across the breadth of life sciences, including our own field of aquaculture genomics. Against the background of all the chapters in this book, it seems fair to say that the development and use of genetic markers and analytical methodologies has revolutionized our view of genetic resources in wild and cultured populations of aquatic organisms.

Genetic analysis of both wild and cultured populations is relevant to the interests of aquaculture. Genetic analysis of wild populations is useful for understanding phylogenetic (i.e., evolutionary) relationships within and among species, establishing population genetic differentiation within species, collecting genetic resources representing the full range of variation in the species, and detecting the escape of cultured fish into wild populations. Screening cultured populations is useful for understanding population genetic differentiation among cultured stocks within the species, inferring parentage in mixed-family assemblages, maintaining genetic variability in the population, estimating the genetically effective size of the population, and inferring the effects of selection within the cultured stock (see Chapter 8). Here, we explain each of these applications and their relevance to aquaculture scientists. We support our discussion with case studies drawn from the published literature in aquaculture genetics. We refer to key software packages, providing a literature citation and an Internet URL to encourage further inquiry and use by interested readers. We hope that this approach increases appreciation of population genetics and phylogenetics theory and use of associated tools by the aquaculture genomics community.

Genotyping

The principles and applications of various marker systems have been discussed in Chapters 2–7, and therefore, we will focus on the analysis of marker genotypes for population genetic analysis. For the most part, we will focus on the use of microsatellite markers. The first task facing a geneticist is to manipulate and score the raw data. The nature of these tasks and associated software tools of choice will depend upon the genetic markers being analyzed.

Codominant Markers

Among DNA markers, microsatellites have become the markers of choice in many laboratories because they are highly variable, reproducible, codominant, and are easy to score using PCR and automated fragment analysis (Chistiakov et al. 2006, Schlotterer and Tautz 1992). Microsatellites tend to add or subtract a single repeat. Under the step-wise mutation model (SMM) (Kimura and Ohta 1978, Bell and Jurka 1997), the evolutionary affinity of alleles can be inferred, a property useful for population genetics and phylogenetic inference, as discussed below. Further, the potential for cross-species application of microsatellite primers means that even smaller laboratories without the capacity to develop new primer libraries can use the technique (see Chapter 5). With codominant markers, both alleles can be detected phenotypically on the gel. To manipulate raw data, the point of departure is accurate scoring of fragment sizes and recording of genotypes. This is generally followed by calculation of allelic and genotypic frequencies. Genemapper (ABI 2006) is a program useful for visualizing and scoring microsatellite fragment data generated by an Applied Biosystems automated sequencer. Software packages that are useful in these contexts include Excel Microsat Toolkit (Park 2001), POPGENE (Yeh et al. 1999), GENEPOP (Raymond and Rousset 1995), and Arlequin (Excoffier et al. 2006). MICROSAT (Minch et al. 1999) is a useful program, but is no longer supported by the authors.

Dominant Markers

Dominant markers are those for which homozygotes and heterozygotes cannot be distinguished phenotypically, including RAPD (see Chapter 3) and AFLP markers (see Chapter 4). The dominant mode of marker expression complicates calculation of allele frequencies somewhat, and hence requires special treatment statistically. Such data usually are analyzed using a binary data matrix based on a presence/absence approach, for which there is a small but applicable suite of statistical analyses. A commonly used coefficient is genetic similarity (GS), which is based on the number of bands shared in different taxa (Nei and Li 1979). Software packages available for analyzing data on markers with a dominant mode of expression include REAP (McElroy et al. 1992) which calculates a range of relevant metrics. Phyltools (Buntjer 2004) is a package of utilities that is particularly useful for large data sets, converting data files,

and manipulation of data to account for missing data points and monomorphisms. TFPGA (Miller 1987) performs a wide variety of population genetic analyses for either codominant or dominant markers. Winboot (Yap and Nelson 2006) reads binary data in either PHYLIP format (Felsenstein 2004) or an Excel-like format and constructs population trees. POPGENE (Yeh et al. 1999) allows use of input files with a presence/absence format and can be used to calculate coefficients of genetic diversity and differentiation. Several packages are available for reading and comparing gel images for DNA fingerprinting-related applications and computing relevant statistics. Such packages include DENDRON (Soll Technologies, 2006) and GelCompar II (Applied Maths 2006).

DNA Sequence Variation

With the advent of widespread access to automated sequencers, analysis of DNA sequence variation has become common, and a variety of software packages, both commercial and freeware, has been developed. CLUSTALX (Thompson et al. 1997) is a Macintosh-based alignment program for aligning DNA sequences. It is probably the most commonly used program for the purpose, especially for sequences that are more divergent and thus more difficult to align. Sequencher (Gene Codes Corporation 2006) is a widely used, commercial program that is useful for sequence assembly and editing, contig assembly, determination of consensus sequences, SNP detection, and restriction mapping. MEGA (Kumar et al. 2004) can be employed for automatic or manual sequence alignment, as well as for various phylogenetic analyses. BioEdit (Hall 1999) is a sequence alignment and analysis program that operates within the Windows environment. Although not as efficient and easy to use as Sequencher, it is freeware that can be downloaded from the Internet. BioEdit allows the user to align sequences by visual inspection, edit nucleotides, manually manipulate regions as needed, and perform many of the same functions as Sequencher. The PolyBayes program is also widely used for the detection of SNPs. Many other programs for manipulating and scoring DNA sequence data also are available.

For genetic dominant and sequence markers, the mutational history of alleles cannot generally be inferred, and geneticists must assume that every mutation creates a new allele, an assumption underlying the infinite allele model (Kimura and Crow 1964) mentioned below.

Genetic Variation within Populations

After Mendel's landmark work, phenotypes could be related to genotypes, and their frequencies could be used to make predictions of their ratios within families. What, though, of their ratios within populations? A model is needed to relate genotype frequencies to allele frequencies, which can be used to reach inferences about processes acting upon populations. The Hardy-Weinberg model (Hardy 1908, Weinberg 1908) has several underlying assumptions: population size is large and constant between generations, mating is random (i.e., the population is panmictic), the organism is

diploid, generations do not overlap, reproduction is sexual, and the effects of mutation, migration, and selection are negligible. For autosomal loci (i.e., those not located on sex chromosomes) with two alleles, the Hardy-Weinberg model is expressed as: $(p + q)^2 = p^2 + 2pq + q^2 = 1$, where p is the frequency of the more common allele A, and q is the frequency of the less common allele a and $p + q = 1$. A and a are alleles for a particular trait for which we can distinguish carriers of the genotypes AA, Aa, and aa on a phenotypic basis. The frequency of the more common homozygote, AA, is p^2; the frequency of the heterozygote, Aa, is $2pq$; and the frequency of the less common homozygote, aa, is q^2. The Hardy-Weinberg model not only relates frequencies of phenotypes and genotypes at the population level, but also predicts constancy of allele frequencies between generations, and may be used to predict genotype frequencies given present allele frequencies, subject to the satisfaction of the model's assumptions. Key applications of the model follow:

1. Infer frequency of recessive allele (solve for q)
2. Infer frequency of "carriers" of a trait, and
3. Testing for genotype frequency equilibrium

Tests of departure from expected genotype frequencies can be categorized into two groups (Guo and Thompson 1992). One group consists of large-sample goodness-of-fit tests such as χ^2, likelihood ratio statistic G^2 and the conditional χ^2 test (Li 1955). The other approach involves exact tests (Levene 1949, Haldane 1954, Chapko 1976). Exact tests are performed when sample sizes are small and hence, expected frequencies of some genotypes are small, but such tests are computationally intensive. Screening of highly polymorphic markers necessitates use of exact tests. Guo and Thompson (1992) proposed two algorithms to estimate significance levels for exact tests of Hardy-Weinberg proportions. They involve (computerized) resampling procedures, one a Monte Carlo and one a Markov Chain method (Box 9.1) GENEPOP (Raymond and Rousset 1995) may be used to implement this method.

Genetic diversity at a single locus is characterized by three parameters: A = allelic diversity, the number of alleles observed; H_o = observed heterozygosity, the total number of heterozygotes/sample size; and H_e = expected heterozygosity. For a single locus with two alleles, $H_e = 2pq$, also called gene diversity. When there are more than two alleles, $H_e = 1 - \sum_{i-1}^{A} p_i^2$, where i = frequency of the ith allele. Often, H_e is reported in preference to H_o because it is less affected by sampling. To reliably estimate heterozygosity, multiple loci must be sampled. To extend these metrics to multiple loci, A = total number of alleles over all loci/number of loci. H_o and H_e can be averaged over loci.

Application of the Hardy-Weinberg model provides a basis for assessment of evolutionary forces affecting the array of genotypes in a population. That is, when a fit of observed data to Hardy-Weinberg expectations is found, all assumptions underlying the model were at least approximately met. When there is significant departure, then one or more of the underlying assumptions were not met. Deviations may include inbreeding, assortative or disassortative mating, random genetic drift, selection, or population differentiation and mixing. Independent lines of evidence will be needed to infer the cause(s) of the departure, inferences often reached using statistical approaches described in subsequent sections of this chapter.

Sewall Wright (1965) developed a statistical approach to partitioning the departure of genotype frequencies in a system of populations into within- and between-population

Box 9.1. Methods underlying certain population genetic data analytic procedures

Several methods underlie population analytic methods used for drawing inferences from population genetics data. Key methods briefly summarized here are more fully explained for population genetics contexts by Luikart and England (1999) and for phylogenetics contexts by Nei and Kumar (2000).

Maximum likelihood methods find the genetic parameters that maximize the likelihood of obtaining the observed data under a certain model. For example, a statistical model is developed that gives the probability of obtaining the observed allele frequencies given a set of demographic and mutational parameters (Luikart and England 1999). It then can be determined which parameter values maximize the likelihood of obtaining the observed data. Maximum likelihood methods are difficult to understand and computationally intensive, complications addressed to a large degree by the availability of specialized software packages.

Coalescent methods assess a genealogical view of intraspecific variation, looking backward through history past occasional mutations to infer pathways by which extant haplotypes "coalesce" to common ancestors (Avise 2000). The demographic history of a species leaves signatures on the coalescence events in population trees. Examination of patterns of pairwise genetic differences can lead to inference of the rate and date of changes in population size (Rogers and Harpending 1992, Avise 2000).

Bayesian methods provide a probability distribution (a "posterior distribution") for a parameter of interest (e.g., N_e) by using the data and incomplete data or expert opinion about probability distribution (a "prior distribution") of one or more input parameters underlying the model (Luikart and England 1999). For example, a Bayesian approach might incorporate a prior distribution of mutation rates ranging from 1×10^{-2} to 1×10^{-6}, with 1×10^{-4} being most probable, even though our knowledge of the microsatellite mutation rate in our species of interest is incomplete. Although the subjective aspect of Bayesian inference is disturbing to some users, Bayesian methods can yield reasonably precise estimates of genetic parameters, especially when prior information is available.

Markov chain Monte Carlo (MCMC) algorithms are computation-intensive stochastic simulation methods for solving complex functions, for example, the mathematical integration needed to calculate the posterior distribution for Bayesian analyses (Luikart and England 1999).

components. The Hardy-Weinberg model is revised to add a parameter, F, the fixation index, which captures the deviation from expected genotype frequencies:

$$\begin{aligned} \text{frequency of } (AA) &= \bar{p}^2 + F\bar{p}\bar{q} \\ \text{frequency of } (Aa) &= 2\bar{p}\bar{q} + (1 - F) \\ \text{frequency of } (aa) &= \bar{q}^2 + F\bar{p}\bar{q} \\ F &= \mathrm{v}_p / (\bar{p}\bar{q}) \end{aligned} \tag{9.1}$$

The overall deviation from Hardy-Weinberg expectation (termed F_{IT}) can be partitioned into: F_{IS} = the deviation due to "inbreeding" (i.e., to processes occurring within populations), and F_{ST} = the deviation due to differentiation (i.e., to processes occurring between populations). In this partitioning:

$$\begin{aligned} H_i &= 1 - (p_i^2 - q_i^2) = \text{expected heterozygosity in population } i. \\ H_S &= (H_i)/n = \text{expected heterozygosity among } n \text{ populations in a group of related populatons (in fisheries, often termed a genetic stock).} \\ H_T &= 1 - (\bar{p}^2 + \bar{q}^2) = \text{expected total heterozygosity.} \\ F_{ST} &= 1 - (H_S / H_T) = \text{genetic deviation due to differentiation among populations} \\ (1 - F_{IT}) &= (1 - F_{IS})\,(1 - F_{ST}). \end{aligned} \tag{9.2}$$

Note that F_{ST} is calculated for each allele at a locus. With two alleles, the values will be equal. With more than two alleles, the values will differ. To extend this approach to multiple loci, after calculating F_{ST} for all alleles, the mean value can be calculated. If a population is divided into many breeding units, then the frequency of homozygotes tends to be higher than Hardy-Weinberg expectation, a phenomenon termed the Wahlund effect.

A variety of software packages may be used for calculating statistics quantifying within-population genetic variation, including Arlequin (Excoffier et al. 2006), POPGENE (Yeh et al. 1999), and FSTAT (Goudet 2002). Analysis of genetic diversity and fixation indices for DNA sequences can be performed using DNAsp (Rozas and Rozas 1995).

Assessment of Inbreeding

Inbreeding is the mating of related individuals. In the aquaculture context, inbreeding results from crossing related individuals in the hatchery. Every individual carries recessive alleles that are not expressed because they are "masked" in the heterozygous state. Some of these alleles are deleterious, and were they expressed, they would negatively impact the fitness of the carrier. Related individuals are more likely to share the same recessive alleles than unrelated individuals. In the classical, dominance-mediated model of inbreeding depression, inbreeding increases the rate of homozygosity by pairing alleles that are identical by descent, thereby increasing the likelihood that deleterious recessive alleles will be expressed in the homozygous state. Expression of deleterious recessive alleles can impact the fitness of the carriers as decreased larval viability, survival through key life cycle events, growth rate, or reproductive ability. The frequency of abnormalities may be increased. Decreased fitness due to inbreeding is referred to as inbreeding depression. Increasing levels of inbreeding tend to be associated with more pronounced inbreeding depression. Inbreeding depression in aquaculture stocks has been demonstrated for a range of fitness and production traits in rainbow trout (Kincaid 1976a, 1976b, 1983), growth rate and other economic traits in channel catfish (Bondari 1984, Bondari and Dunham 1987), and larval viability and growth rate in eastern oyster (Longwell and Stiles 1973).

Classical Approaches

An analytical technique called path analysis (Gall 1987, Tave 1993, Hallerman 2003) (detailed in Chapter 10) is used to identify the lines of descent from a common ancestor to parents of an individual of interest and to estimate the contribution of each line of descent to the resulting level of inbreeding of the individual. Essentially, the approach estimates the likelihood that two homozygous alleles in an individual are identical by descent from the common ancestor. The inbreeding coefficient (F_x) of an individual is calculated as $F_x = \Sigma\ [(½)^N(1 + F_A)]$, where F_x is the inbreeding coefficient for individual X, N is the number of individuals in a given path, and F_A is the inbreeding coefficient for the common ancestor (if nonzero). Closed aquaculture stocks may have complex pedigrees. A number of computational methods have been developed for calculating inbreeding coefficients for large populations for which pedigrees are known (Emik and Terrill 1949, Cruden 1949 as cited by Gall 1987). However, pedigrees for individuals in most cultured stocks are unknown.

In the absence of pedigree information, we cannot calculate F for individuals. We can, however, estimate the mean rate of inbreeding for the population if we make a key simplifying assumption. Assuming that each individual contributed equally to the progeny generation, a mean F for the population can be estimated when the numbers of breeders of each sex used to propagate a population are known (Tave 1993, Hallerman 2003), as described in Chapter 10. Since this simplifying assumption is unlikely to be met fully in most situations, the resulting estimated rate of inbreeding per generation should be regarded as the lower bound for that actually occurring in the population. The rate of inbreeding for a population also can be estimated as a function of its inbreeding effective population size, Ne, as $F = 1/2N$e, procedures for estimation of which are discussed below.

A classical approach for inferring that inbreeding may have occurred in a population of interest is to demonstrate that frequencies of homozygotes significantly exceed those expected under Hardy-Weinberg equilibrium. Wright (1931) developed an approach for estimating a parameter f quantifying how much of a population's departure from Hardy-Weinberg equilibrium might be attributed to inbreeding. This approach, however, is not particularly powerful for drawing inferences, because departures from Hardy-Weinberg expectations may occur for other reasons (e.g., mixing of populations) or for multiple reasons (selection, drift, mixing).

A large body of studies has tested the hypothesis that individual heterozygosity at isozyme or microsatellite markers is correlated with chosen metrics of fitness (Hallerman 2003) (see Chapter 10). Most studies invoke either overdominance or inbreeding as the underlying explanation for results supporting the hypothesis. The implicit assumption in such studies is that an individual's degree of inbreeding can be estimated reliably using on the order of 10 markers. Balloux and others (2004) used individual-based simulations to examine the conditions under which heterzygosity and inbreeding are likely to be correlated. The results indicated that for such a correlation to prove strong inbreeding must be severe and frequent and the number of loci screened must be large (approximately 200). Observed correlations of heterozygosity and fitness likely reflect linkage of marker and fitness-related loci.

Recent Approaches

Newer approaches to assessing inbreeding in individuals or populations are based on the stepwise mutation model (SMM) for microsatellite markers. Under the SMM, because mutation tends to add or subtract one microsatellite repeat, alleles with very different lengths are less related than alleles with similar lengths.

Coulson and others (1998) inferred the relative level of inbreeding within individuals by considering individual heterozygosity (likely to reflect recent inbreeding) and the relative lengths of their microsatellite alleles at a locus (likely to reflect inbreeding earlier in the pedigree). Regarding microsatellite markers, heterozygotes with alleles of very different length were considered less inbred than heterozygotes with alleles of similar length. A method for calculating an individual-specific internal distance measure (d^2) was described. Outbred neonatal red deer (Coulson et al. 1998) and harbor seals (Coltman et al. 1998) were heavier at birth than relatively inbred individuals. Coulson et al. (1998) concluded that both heterozygosity and mean d^2 were useful for characterizing inbreeding in populations.

Ayres and Balding (1998) and Saccheri and others (1999) developed likelihood-based methods for estimating population-level inbreeding coefficients from microsatellite data. Ayres and Balding (1998) proposed a MCMC Bayesian method (see Box 9.1) for assessing inbreeding-mediated departures from Hardy-Weinberg equilibrium. The method incorporates the effect of uncertainty regarding allele frequencies and constraints on f, the parameter that measures departure for Hardy-Weinberg equilibrium due to inbreeding. The advantages of the method were illustrated by considering data on the *Plasmodium* malaria parasite and on humans. Saccheri and others (1999) showed that heterozygosity decreased more than expected in experimentally bottlenecked populations of a butterfly. They found it more informative to estimate the probably distribution of σ^2, the variance in the number of descendants left per gene, a measure of inbreeding. These likelihood-based methods may perform better than classical F-statistics and support more explicit testing of different inbreeding models and their underlying assumptions (Luikart and England 1999).

Estimation of Effective Population Size

When an aquaculture stock is propagated, allele frequencies may be changed by random genetic drift (i.e., by nonselective processes driven by sampling error). Drift tends to decrease genetic variation; the change in genetic variance is relatively low for large populations and high for small populations. Drift can have a cumulative effect over a run of generations. To assess the effect of small population size, Wright (1931, 1938) developed the concept of the effective population number, N_e.

Demographic Approaches

While several different concepts of N_e have been developed (Hallerman 2003), in our context we are most concerned with the amount of allele frequency drift, and hence

with the variance effective number. Three mechanisms decreasing N_e are especially relevant in the aquaculture context.

1. Unequal numbers of male and female breeders. In such cases, the effective population size is related directly to the sex ratio as: $N_e = 4N_mN_f / (N_m + N_f)$, where: N_m = the number of breeding males, and N_f = the number of breeding females.
2. Nonrandom family size. If census population size N is constant and mating is random, N_e can be estimated as: $N_e = 4N/(2 + V_k)$, where V_k = variance in family size. Note that $N_e = N$ if $V_k = 2$. In many contexts, family size is likely to be different for the sexes, in which case $N_e = 8N/(4 + V_{km} + V_{kf})$, where V_{km} = variance in family size among females, and V_{kf} = variance in family size among males (Hill 1979).
3. Different numbers of parents in successive generations. If loss of variability occurs because a restricted number of individuals was used to found a new stock, it is referred to as a founder effect. If loss of variation occurs due to a restricted number of individuals reproducing within an existing stock, it is termed a bottleneck effect. The effect of a reduction in the number of breeders on the effective size of the population is estimated as $1/N_e = 1/t\ (1/N_{et=1} + 1/N_{et=2} + \ldots + 1/N_{et=t})$, where t = the number of generations, and $N_{et=n}$ = the effective size of the population in the generation specified. Since N_e appears in the denominator of the fractions, it follows that small values of N_e make a large impact on the mean N_e over a period of generations.

These mechanisms can, of course, all be at issue in a cultured population, simultaneously or at different times in the breeding history of the stock. The practical implication of these considerations is that aquaculturists should not allow a stock to become bottlenecked. The effects of random genetic drift upon N_e were seen by Allendorf and Phelps (1980), Cross and King (1983), Vuorinen (1984), Hallerman and others (1986), Brown and others (2005), who showed loss of considerable genetic variation after a few generations of captive propagation. However, the data requirements for demographics-based methods of estimating N_e frequently restrict their usefulness for aquaculture applications. In particular, demographic methods require numerous assumptions regarding the relative success of breeders. Only in exceptional instances do we have access to data on who mated with whom and how successfully (see Chapter 8 on parentage inference).

Marker-based Approaches

Because the impact of random genetic drift is a function of effective population size, molecular genetics-based approaches have been developed for using genetic data to obtain indirect estimates of N_e. Some such models estimate N_e using inferences based on the correlation of genetic variation with population size. Large populations will generally exhibit greater genetic variation; further, alleles at unlinked loci will be randomly associated. Smaller populations will exhibit less variation; because of the limited number of allelic combinations occurring in a small population, some unlinked alleles will be inherited together by chance alone. Applying this approach, Hill (1981) estimated N_e by assessing linkage disequilibrium (D) and correlation of alleles at different loci (r) in a sample drawn from a population. The correlation among alleles,

r, is estimated as $r = D/[p(1 - p) + q(1 - p)]^{1/2}$, where p = frequency of allele A at locus 1, q = frequency of allele B at locus 2, and D = Burrow's composite measure of equilibrium, a measure of linkage. A matrix of such correlation values is constructed. N_e (D), the N_e as estimated from linkage disequilibrium data, is estimated as N_e (D) = $1/[3(r^2 - 1/s)]$, where s = sample size. This yields an estimate of N_e (D) for each polymorphic locus; to obtain a single value of N_e (D), the arithmetic mean value of the r^2s and the harmonic mean of the s's are calculated and substituted into the equation. Equations are presented by Bartley and others (1992) for calculating confidence intervals and the variance of N_e (D). Both D and r will be zero in an ideal, infinite, randomly mating population, but will depart from zero in real, finite populations due to drift, migration, selection, and linkage. The gametic disequilibrium approach to estimating N_e can be applied using the LINKDOS option in GENEPOP (Raymond and Rousset 1995). Bartley and others (1992) demonstrated use of Hill's approach using isozyme marker data. One case study considered a hatchery enhancement program for white sea bass (*Atractoscion nobilis*), in which juveniles produced by 25 adults spawning in a tank year after year were sampled. The ratio of N_e (D)/N was below 1.0 because of unequal sex ratio among spawners and unequal contribution of adults to spawn. Another case study examined the rainbow trout (*Oncorhynchus mykiss*) stock at Shasta Hatchery, California, an intensively managed hatchery-breeding program tracking family data, sex ratios, and genetic contributions among 17–30 single-pair matings. N_e (D) was found to approximate N, arguing for good breeding practice at the hatchery and the validity of the method used for estimating N_e.

Other models (e.g., Nei and Tajima 1981, Pollack 1983) use temporal variation in allele frequencies to estimate an average N_e per generation over the time interval of interest. That is, these models use direct measurement of the effects of random drift to infer what population size would have caused the observed change in allele frequencies. The approach is particularly robust over intervals of 2 to 10 generations and when N_e is small (Waples 1989). Modifications of such models for estimating effective sizes of salmonid populations were made by Waples (1989, 1990), Waples and Teel (1990), and Tajima (1992). Jorde and Ryman (1995) showed how to estimate N_e for brown trout populations with overlapping generations. Temporal variation models have been shown useful in estimating N_e in hatchery populations of salmon (Waples and Teel 1990) and shellfish (Hedgecock and Sly 1990, Hedgecock et al. 1992, Appleyard and Ward 2006). The disadvantage of the temporal variance approach is that two temporally separated samples of the population (at least two generations apart) are needed. The software package TM2.exe (Beaumont 1999) implements maximum likelihood, coalescent, Bayesian, and MCMC algorithms for estimating current N_e from the temporal variance of allele frequencies.

A third approach to estimating N_e quantifies heterozygosity excess (Pudovkin et al. 1996, Cornuet and Luikart 1996). The approach is based on the observation that when populations are small, binomial sampling error produces genotype frequency differences between males and females, resulting in an excess of heterozygotes in their progeny relative to Hardy-Weinberg expectations. By quantifying the magnitude of heterozygote excess in their progeny, one can estimate N_e. A primary advantage of this approach is that only one sample is needed to estimate N_e. However, the method may be valid only for breeding systems with a random union of gametes, as in broadcast spawners, and confidence intervals about N_e estimates are very large. However, Luikart and Cornuet (1999) evaluated the accuracy and precision of the heterozygote

excess method using simulated and empirical data for monogamous, polygynous, and polygamous mating systems using realistic sample sizes of 15–120 individuals and 5–30 markers with varying levels of polymorphism. Estimates of the effective number of breeders, N_{eb}, were nearly unbiased for all mating systems. However, confidence intervals were acceptably small only for polygynous and polygamous systems with fewer than 10 effective breeders unless samples included more than 60 individuals and 20 multiallelic marker loci. The heterogeneity excess approach for estimating N_e may be applied using the BOTTLENECK software (Piry et al. 1999).

Comparing among gametic disequilibrium, temporal variance, and heterozygote excess methods for estimating N_e, Schwartz and others (1998) regarded the temporal variance method the most promising because it provides increasingly precise estimates of N_e as the number of alleles per locus increases. They noted, however, that the performance of the three methods had not been compared under the same conditions. Schwartz and others (1999) noted that coalescent, maximum likelihood, and Bayesian statistics (Nielsen et al. 1998, O'Ryan et al. 1998) can be used to estimate current N_e from changes in allele frequencies assuming a known divergence time from common ancestral populations, a context that may be relevant to aquaculture genetics.

Phylogenetic and coalescent theory have provided a number of estimators for inbreeding effective population numbers based upon nucleotide sequence and microsatellite data (Felsenstein 1992; Fu 1994a, 1994b; Kuhner et al. 1995; Beerli and Felsenstein 1999). However, the inbreeding effective population number is a more retrospective statistic, while the variance effective population number reflects more recent demographic and population genetics processes (Crandall et al. 1999). A recent bottleneck would result in a large inbreeding N_e and small variance N_e, while rapid increase in population size would result in small inbreeding N_e and large variance N_e. Relating to our context, phylogenetic and coalescence-based methods, while appropriate for estimating N_e over an evolutionary time-scale (hundreds of generations), would not seem particularly relevant to aquaculture genetics.

Inferring Selection

Selection, the differential perpetuation of genes and gene complexes (Dobzhansky 1970), is the evolutionary force mediating adaptation of populations to their environment. In the aquaculture context, domestication selection occurs as genotypes not suited to culture conditions are eliminated and those suited to culture conditions increase in frequency. Breeders purposefully use artificial selection to develop aquaculture stocks that are more productive under aquaculture conditions. Yet, selection is not the only source of change in allele frequencies in cultured stocks. Nonselective processes include random genetic drift (mediated through founder effects, limited numbers of breeders, and unequal breeding success), mutation, and introduction of genetic material from outside the cultured stock. For theoretical or practical selective breeding purposes, aquaculture scientists may want to infer which loci are subject to selection. In practice, however, this is not a simple problem (Nei and Kumar 2000, Ford 2002, Hedrick 2005).

Many contributions to the population genetics literature seek to attach selective significance to allozyme polymorphisms, as discussed in Chapter 7 (Kimura 1983, Nei

and Graur 1984), and methods were developed to test the departure of observed allozyme frequencies from expectations based upon assumptions under selective neutrality (Ewens 1972, Lewontin and Krakauer 1973, Watterson 1978). However, these tests had low experimental power and their underlying assumptions frequently were violated, making strong inferences about selection difficult (Ford 2002). For example, Hallerman and others (1986) were able to infer that selection had affected allele frequencies in cultured stocks of channel catfish but were not able to identify the affected loci.

Because of the relatively great information content of DNA sequence data, statistical tests of the fit of observed data to neutral models has proven more successful, and many such tests have been developed (Ford 2002). Application of selection models to DNA-level data, however, are complicated by issues of synonymous and nonsynonymous nucleotide substitutions, codon usage bias, and coding and noncoding regions of nuclear DNA. A complete discussion of these issues as related to inference of selection is beyond the scope of this chapter, and may be found in Brookfield and Sharp (1994) and Kreitman and Akashi (1995). We will restrict our discussion to three models that have been applied to aquatic organisms.

Tajima (1989) developed a test statistic, D, based on the difference between two estimators of the neutral polymorphism parameter $4N_e\mu$, where μ is the mutation rate. One estimator is based on the total number of polymorphic nucleotide sites observed, and the other based on the average number of differences between all pairs of sequences sampled. Under neutrality, both estimators are expected to be equal, while under selection they differ. Negative D values indicate an excess of rare variants, and positive D values indicate an excess of intermediate frequency variants. The test for selection entails quantification of how likely the difference between the estimators is attributable to chance alone. Tajima (1989) demonstrated testing of the D statistic using a data set on *Drosophila*, and it has since been applied to microbes, plants, a number of animals, and humans. Fu (1997) discusses a number of tests similar to Tajima's D.

Synonymous mutations in a DNA sequence are expected to be selectively neutral because they do not result in a change in the expressed protein. Nonsynonymous mutations, because they do result in a change in the expressed protein, are more likely to be subject to selection. Hence, the rates of nonsynomymous (d_n) to synonymous (d_s) mutations among DNA sequences can be used as a test of positive selection (Hill and Hastie 1987, Hughes and Nei 1988). If most nonsynonymous mutations are deleterious, then d_n/d_s will be less than 1; if most are beneficial, d_n/d_s will be greater than 1. Application of the approach to a piscine system is perhaps best shown by studies of salmonid transferrin, an iron-binding protein that plays a role in iron metabolism and resistance to bacterial infection. Comparison of transferrin sequences from four salmonids showed that the rate of evolution at nonsynonymous sites was significantly faster than the rate at synonymous sites (Ford et al. 1999), suggesting positive selection for new alleles. In contrast, there was no evidence of greater differentiation among Chinook salmon populations at nonsynonymous than at synonymous sites, nor of reduction of variation due to the hitchhiking effect at the transferrin gene (Ford 2000). Within the salmonids, roughly 13% of transferrin codon sites were inferred to be subject to positive selection; evidence of positive selection was limited to salmonids, however, and was not seen in other fish, amphibians, or mammals (Ford 2001). The molecular locations of sites subject to selection in salmonids supported the

view that selection is driven by competition for iron with pathogenic bacteria. Other work using the d_n/d_s test approach in piscine systems included analysis of nucleotide sequences for odorant receptor genes of channel catfish, which suggested that positive selection generated enhanced diversity within the putative odorant domains (Ngal et al. 1993).

McDonald and Kreitman (1991) put forward a two-by-two contingency test comparing the numbers of nonsynonymous and synonymous polymorphisms within and between species. Under neutrality, the ratio of nonsynonymous/synonymous polymorphisms within species is expected to equal that between species. Among applications in aquatic species, analysis of nucleotide sequences of 124 pantophysin I alleles showed evidence of both balancing and directional selection, but no evidence of a geographic basis for the varying selection (Pogson 2001). Heterogeneity in the frequencies of recently derived mutations suggested that two selective sweeps may be occurring among populations. Positive selection on *pan I* subsequently was found in other members of the cod family (Pogson and Mesa 2004, Canino and Bentzen 2004).

Thorough analyses of the properties and application of selective neutrality tests are beyond the scope of this chapter, but can be found in Li (1997), Nei and Kumar (2000), Ford (2002), and Hedrick (2005). Software for executing five tests of departures from selective neutrality is available in DNAsp (Rozas and Rozas 1995).

Although the results would be of both theoretical and applied interest, we are not aware of applications of models of selective neutrality to DNA-level data for aquaculture stocks. We encourage such applications. Particularly interesting would be improved understanding of the relation between host-pathogen or parasite interactions and variation at the major histocompatibility complex, which potentially could be applied in selective breeding. Alleles or haplotypes that confer resistance to key pathogens or parasites in resource lines could be introgressed into production lines by marker-assisted selection.

Population Differentiation

A fundamental issue on population genetics is detection, quantification, and explanation of the forces driving differentiation of populations. The issue also arises in aquaculture. Differentiation of cultured stocks from wild populations has been considered for species including masu salmon (Nakajima et al. 1986), Atlantic salmon (Skaala et al. 2004), arctic char (Lundrigan et al. 2005), puffer fish (Cui et al. 2005), and pearl oyster (Yu and Chu 2006). Differentiated populations may be identified for possible crossbreeding, as has been considered for Asian sea bass (Zhu et al. 2006). Breeders may have an interest in identifying unique wild populations that are candidate resource stocks for introgression of valued traits into cultured stocks.`

Classical Approaches

The first question regarding genetic variation in a collection of populations is how much occurs within and how much occurs among populations. The classical population genetics literature provides a number of analytic approaches. Some predate,

while others were developed following the impetus provided by the development of allozyme markers.

One of the key assumptions underlying the Hardy-Weinberg model is that genotype frequencies are being considered across a single, panmictic population. If genotype frequencies depart from expectations, one possible explanation is that the data under consideration actually are drawn from differentiated populations. Tests of departure of genotype frequencies from Hardy-Weinberg equilibrium are discussed above in the section on genetic variation within populations.

The observed average heterozygosity in a pool of populations with different allele frequencies is reduced relative to the heterozygosity in the pooled population if all individuals were randomly mating. Wright (1965) developed a theory useful for quantifying population genetic differentiation based upon the fixation index, F_{ST}, which he defined as the correlation of two alleles chosen at random from the total population. The fixation index is calculated as:

$$F_{ST} = (H_T - \bar{H}_S) / H_T, \tag{9.3}$$

where H_T is the expected heterozygosity in a randomly mating pooled population, and $\bar{H}_S$ is the average of the expected heterozygosity in randomly mating subpopulations.

Genetic variation may be hierarchical (e.g., populations within rivers within major drainages within regions within the total range), and the analysis can be structured to estimate partitioning of genetic variance at multiple levels. Terms of reference for what comprises biologically meaningful genetic differentiation must be made anew for every species.

Wright's classic formulation for fixation indices was developed for the case of two alleles, leaving a need to extend it to multiple alleles. Further, in Wright's definitions, F_{IS} and F_{IT} are the correlations between two uniting gametes relative to the subpopulation and the total population, respectively, and F_{ST} is the correlation of two gametes drawn at random from each subpopulation. The underlying worldview is that all the populations or subpopulations are derived from a common ancestral population and that all populations are equally related to one another. This pattern of relatedness almost never applies. Populations almost always have some pattern of phylogenetic relatedness. Population size varies, and migration links some but not all populations. Hence, Nei (1977) redefined *F* indices in terms of heterozygosities (i.e., without reference to uniting gametes and proposed a new metric, G_{ST}, the coefficient of gene differentiation) (Nei 1973). The two approaches are related; as for F_{ST} analyses, the relative magnitude of genetic differentiation among populations is:

$$G_{ST} = (\bar{H}_T - \bar{H}_S) / \bar{H}_T \tag{9.4}$$

Often, a matrix of F_{ST} or G_{ST} values among populations is constructed and subjected to a clustering algorithm to construct a population tree showing graphically the patterns of differentiation among populations. FSTAT (Goudet 2002) is a useful program for conducting such analyses.

Genetic distance is the degree of genomic difference between species or populations that is measured by some numerical method (Nei 1987a). Over the past several decades, various measures of genetic distance have been proposed. In some of these metrics, such as, Cavalli-Sforza and Edwards' (1967) chord distance, populations are

regarded as points in multidimensional space, and genetic distance is measured as the geometric distance between these points. The absolute values of these distances have no particular biological meaning; rather, the relative values are used to infer genetic relationships among populations. Using a contrasting approach, Nei (1972, 1973b) quantified the number of gene substitutions per locus that have occurred since divergence of two populations under consideration, providing an absolute measure with biological meaning. Minimum, standard, and maximum distance metrics were proposed to take account of molecular genetic methodological and statistical complexities. A thorough discussion of distance metrics is provided by Nei (1987b). Distance measures may be estimated and analyzed using a variety of general-purpose population genetics analysis programs mentioned below in this chapter.

In a rather different approach, allele frequencies are regarded as random variables and a statistical model for these variables is developed. Parameters in the model are estimated for allele frequencies observed in samples of subpopulations. The parameters in the model are regarded as components of variance. Cockerham (1969, 1973) developed the approach for a single, two-allele locus. Weir and Cockerham (1984) extended the approach to multiple loci and alleles. Subsequently, Excoffier and others (1992) incorporated information on DNA haplotype divergence into an analysis of variance format. This analysis of molecular variance, termed AMOVA, produces estimates of variance components and *F*-statistic analogs, designated as ϕ statistics. These metrics reflect the correlation of haplotypic diversity at different hierarchical levels of population genetic organization, for example, within populations, among populations within groups, and among groups. Excoffier and others (1992) presented permutational procedures to provide significance tests for each of the hierarchical variance components. AMOVA may be run using Arlequin (Excoffier et al. 2006).

It can be useful to determine whether differentiation among populations reflects real differences in selective forces or is simply the result of geographic distance or stream distance (i.e., isolation by distance). Most coefficients of divergence, such as, F_{ST}, R_{ST}, and genetic distance values, can be correlated to physical distance using a Mantel test, a correlation between two data matrixes. Suitable software for such analysis is PASSAGE (Rosenberg 2003).

Recently Developed Analyses for Highly Variable Markers

Analyses using F_{ST} or G_{ST} generally assume that mutation follows an infinite allele model or that mutation is negligible and can be disregarded. However, mutation of microsatellite markers may be too rapid to be disregarded, and may follow a stepwise mode of mutation. Several new statistical approaches have been developed to deal with the high mutability and stepwise mutational mode of microsatellite markers.

R-statistics (Slatkin 1995) were developed to account for loci undergoing stepwise mutation. R_{ST} may be calculated as:

$$R_{ST} = (S - S_W) / S, \tag{9.5}$$

where S_W and S are the average sum of squares of the differences in allele size within a population and in the pooled population, respectively. A matrix of R_{ST} values

among the populations studied often is used to construct a population tree. Because it captures the evolutionary history of microsatellite markers, the R_{ST} metric can prove more powerful for quantifying population genetic differentiation than IAM-based metrics such as F_{ST} or G_{ST} (Balloux and Lugon-Moulin 2002). R_{ST} metrics can be calculated using the RSTCALC software (Goodman 1997).

ρ is a metric that accounts for stepwise, multilocus variation of microsatellite alleles that is also correct for variation of sample size (Michalakis and Excoffier 1996). ρs, the proportion of shared alleles, is a measure of similarity between the multilocus genotypes of two individuals (Bowcock et al. 1994). ρs is calculated as the number of shared alleles summed over loci divided by (2 × number of loci compared). A genetic distance measure *D*ps between two individuals may be calculated as 1-ρs, which can in turn be averaged within or among populations. *D*kf is a metric of population relationship based on allele frequencies (Bowcock et al. 1994), with 1-*D*kf providing a measure of genetic distance among populations. The program MICROSAT (Minch et al. 1999) is useful for estimating pairwise genetic distances both among individuals and populations using the *D*kf and *D*ps metrics.

Most coefficients of differentiation are based on population-level data, such as allele frequencies and diagnostic alleles within populations. With highly variable markers such as microsatellites, differentiation of individuals becomes feasible. Cornuet and others (1999) noted that a simple study consisting of five loci and three alleles per locus can yield 7,776 possible genotypes. With this degree of variability, the individual can become the ultimate taxon, with the individual multilocus genotype replacing allele frequency as the basis for data analysis. The underlying principle of assignment tests is that individual multilocus genotypes are assigned to populations where their expected frequency is greatest. Individuals whose genotypes are more likely in populations other than the one in which they are found in are said to be "misassigned." Procedurally, assignment tests provide the following:

1. Identify an individual's population of origin
2. Add a measure of confidence to the estimate
3. Exclude populations when necessary

Methods for assignment testing can be broadly divided into likelihood-based methods and distance-based methods. The former can be further approached using frequency methods and Bayesian methods. Further, all these methods can be modified to be exclusion methods rather than assignment methods, noting, however, that it is possible to use exclusion as an assignment method in itself. Limitations to assignment tests should be noted. Some methods of assignment testing will always find a "closest" population to which the individual can be assigned. However, the set of populations included in a study may not necessarily contain the correct population. Clearly, it may be better not to know an individual's population of origin at all rather than to assign it to the wrong population. A measure of confidence is therefore needed. This can be done by comparing the value of the criterion for the individual (relative to the given population) with values of the criterion for individuals known to belong to the population. Assignment tests are useful for studying population differentiation, immigration and dispersal, the success of deliberate introductions, for detecting introgression (although only for a limited number of generations), and for forensic applications. Frequentist, Bayesian, and distance-based approaches to assignment tests have been developed (Rannala and Mountain 1997). Software

packages for applying the assignment test include GENECLASS (Cornuet et al. 1999, Piry et al. 2004), STRUCTURE (Pritchard et al. 2000), WHICHRUN (Banks and Eichert 2000), Spam (Debevec et al. 2000), Doh (Brzustowski 2002), and Genetic Mixture Analysis (Kalinowski 2003), as well as more general purpose packages such as Arlequin (Excoffier et al. 2000), GENEPOP (Raymond and Rousset 1995), and TFPGA (Miller 1997).

Software packages useful for applying a range of approaches for quantifying differentiation of populations include POPGENE (Yeh et al. 1999), GENEPOP (Raymond and Rousset 1995), Arlequin (Excoffier et al. 2006), TFPGA (Miller 1997), and DNAsp (Rozas and Rozas 1995). The theory and methods for study of population genetic differentiation are well elaborated, too well elaborated to be adequately described here. Interested readers are referred to Shaklee and Currens (2003) and Hedrick (2005).

Identifying Escaped Aquaculture Fish and Mixing of Cultured and Wild Populations

Routine aquaculture operations frequently involve the loss of cultured fish to the natural environment, with occasional catastrophic losses of larger numbers of fish. Entry of cultured fish into natural population and interbreeding poses concerns to many ecologists, population geneticists, and fisheries managers (Waples 1991, Utter 2003). The best elaborated case study of such concerns regards Atlantic salmon. Cultured Atlantic salmon stocks are genetically and behaviorally differentiated from natural populations (Einum and Fleming 1997, Gross 1998). Interbreeding with cultured stocks poses genetic risks to wild populations. At the individual population level, interbreeding may lead to loss of fitness through breakdown of coadapted gene complexes. A two-generation experiment comparing fitness traits among wild, cultured, F_1, F_2, and backcross salmon showed that cultured and hybrid salmon exhibited reduced survival, but faster growth than wild fish, and their parr displaced wild parr competitively (McGinnity et al. 2003). In an independent experiment, the lifetime reproductive success of farmed salmon was 16% that of native salmon, and the productivity of the native population was reduced by more than 30% by interbreeding (Fleming et al. 2000). At the landscape level, escapes of cultured Atlantic salmon over time from many culture sites and interbreeding with native populations may homogenize genetic differentiation among native populations (Hindar et al. 1991). Case studies involving nonsalmonid species are less numerous. One instance is hybrid catfish (*Clarias macrocephalus x C. gariepinus*) escaping from farms in central Thailand interbred with native populations of *C. macrocephalus*, giving rise to introgressive hybridization with both wild and cultured stocks (Senanan et al. 2004). In contrast, a survey of channel catfish (*Ictalurus punctatus*) populations in Alabama, USA, showed no evidence of genetic impact from loss of cultured fish into natural populations (Simmons et al. 2006).

Detection of individuals that have escaped from culture operations is relatively straightforward if diagnostic markers are available. In some cases, it may be possible to distinguish cultured individuals by means of phenotypes, including morphometrics, size of fins, or scale characteristics (Lund et al. 1989, Fleming et al. 1994, Hard et al. 2000), or internal characters such as postvaccination intra-abdominal adhesions

(Lund et al. 1997). Molecular markers, such as ratios of stable isotopes (Dempson and Power 2004) or presence of synthetic pigments such as astaxanthin (Lura and Saegrov 1991), have been used to identify escapees. However, these characters are useful only for identifying escapees themselves, and are not useful for identifying their descendents. This raises the need for screening molecular genetic markers, which are most useful in cases where allele frequencies differ among cultured stocks and wild populations. Hence, molecular markers will be the tool of choice for inferring the occurrence and extent of mixing of escaped or released cultured stocks and wild populations. A number of different statistical approaches for inferring the mixing of populations are available, as described below.

As noted above, if genotype frequencies in a collection depart from Hardy-Weinberg equilibrium expectations, because of the Wahlund effect, there will be a heightened frequency of homozygotes. Although this approach was widely used in the allozyme literature to infer mixing of stocks, the advent of microsatellite markers and development of more powerful analytic approaches offer greater experimental power for detecting such mixtures.

As noted above, assignment tests can be used to screen databases comprised of collections of multilocus genotypes to assess how many populations are represented and to identify the collection to which a particular individual's multilocus genotype most likely belongs (Waser and Strobeck 1998). For purposes of identifying escaped aquaculture fishes within a wild population, assignment tests are useful for identifying individuals who are themselves likely migrants among populations (Luikart and England 1999). Programs useful for applying the assignment test are mentioned in the previous section.

A coalescence theory-based method has been developed for determining the proportion of different breeding populations that contributed to a mixed or hybridized population (Bertorelle and Excoffier 1998). A computer program was developed to implement the procedure (Bertorelle 1998).

Molecular Phylogenetics

Aquaculturists may value knowledge of the evolution of a gene in order to understand the expression of a valued trait in a cultured species. They may value knowledge of the evolutionary relationships among wild and cultured stocks, as has been considered for common carp (Kohlmann et al. 2005). Knowledge of evolutionary relationships among wild populations may inform selection of prospective resource stocks for production or conservation aquaculture. For example, while the Asian oyster (*Crassostrea ariakensis*) is being considered for introduction as an aquaculture species into the Chesapeake Bay, there is considerable confusion regarding definition of the species. Wang and others (2004) collected oysters from five locations along China's coast and analyzed morphological, mitochondrial and nuclear DNA characters, distinguishing red- and white-meat forms as distinct species. With the development of methods for propagation and culture of imperiled freshwater mussels, questions arise regarding which potential donor stocks to propagate for outplanting into depleted populations. Grobler and others (2006) examined mitochondrial and nuclear DNA characters among populations of the slabside pearlymussel *Lexingtonia dollabelloides*

and found that upper Tennessee drainage and Duck River populations should not be intermixed by demographic augmentation programs. Analysis of DNA marker data from the evolutionary perspective is the domain of molecular phylogenetics. That is, similarities and differences in DNA markers may be analyzed using sophisticated algorithms to infer evolutionary relationships. The principle underlying the inference is that the greater the time since divergence from a common ancestor, the greater the number of molecular genetic differences at selectively neutral marker loci.

When DNA sequences are derived from a common ancestor, they gradually diverge by nucleotide substitution (Nei and Kumar 2000), as discussed in Chapter 3. However, the rate of substitution varies among sites, for example, among protein-encoding and nonencoding sequences. Hence, we must base our estimate of the number of substitutions upon a well-chosen mathematical model of nucleotide substitutions. Many such models have been developed. Among the most frequently used, under the Jukes-Cantor (1969) model, nucleotide substitution occurs at any site with equal frequency, and at each site a nucleotide changes to one of the three other nucleotides with a probability of α per year. Under Kimura's (1980) two-parameter model, the rate of transitional substitutions per site per year (α) is assumed to differ from that for transversional substitutions (2β). Tajima and Nei's (1984) model is based on an assumption that the rate of nucleotide substitution is equal for all nucleotide pairs. The decision of which model is most appropriate for a given data set can prove difficult. Users may select among 56 models of nucleotide substitution using the MODEL TEST software (Posada and Crandall 1998, Posada 2006).

The goal of phylogenetic inference is to construct a "tree" showing how sampled individuals, extant populations, or species evolved from a common ancestor. Phylogenetic inference may be based on distance, maximum parsimony, or maximum likelihood methods. In distance methods, evolutionary distances are estimated for all pairs of taxa, and phylogenetic trees are constructed by considering the relationships among these distance values (Nei and Kumar 2000), as discussed in Chapter 6. UPGMA (unweighted pair-group method using arithmetic averages), least squares, minimum evolution, and neighbor-joining methods have been elaborated for tree construction. Nei and Kumar (2000) suggested guidelines for constructing distance-based trees, and was also discussed in Chapter 6. Originally developed for morphological characters, maximum parsimony (MP) methods later were applied to molecular genetic data. In MP methods, four or more aligned nucleotide sequences are considered, and the nucleotides of ancestral taxa are inferred separately at each site (Nei and Kumar 2000), as also discussed in Chapter 7. The smallest numbers of nucleotide substitutions that explain the entire evolutionary process for the topology are computed, and the topology with the smallest number of substitutions is chosen as the best tree. Using maximum likelihood (ML) methods, the likelihood of observing a given set of sequence data for a specific model of nucleotide substitution is maximized for each topology, and the topology that gives the highest likelihood is chosen as the final tree (Felsenstein 1981, Nei and Kumar 2000) (see Chapter 8). Bayesian methods that incorporate prior information also may be used to infer phylogenies (Huelsenbeck et al. 2001, Felsenstein 2004a).

When there are more than a few taxa, the number of possible tree topologies becomes huge, and determining statistical support for parts of a tree becomes important. The most common approach for estimating statistical support for a node on a tree is to calculate bootstrap values by resampling the data (Felsenstein 1985, 2004a).

The computational demands are such that phylogenetic inference is virtually impossible without specially developed software packages. Key software packages for performing phylogenetic analyses include PAUP (Swofford 1998), PHYLIP (Felsenstein 2004b), and MEGA (Kumar et al. 2004). MRBAYES (Ronquist and Huelsenbeck 2003) is the best software available for running a Bayesian-based phylogenetic analysis. LAMARC (Kuhner et al. 2004) combines maximum likelihood and coalescence approaches to phylogeny reconstruction. Key software for tree construction, plotting, and drawing include PAUP (Swofford 1998), PHYLIP (Felsenstein 2004b), and Treeview (Page 2000). Hall (2004) is an excellent, step-by-step manual for constructing and printing phylogenetic trees, guiding the reader through the use of PAUP, PHYLIP, and Treeview software packages.

Phylogenetics is a well-elaborated and fast-moving field. We cannot possibly describe all applications, key publications, and supporting software. Key references that we recommend are Hillis and Moritz (1990), Nei and Kumar (2000), and Felsenstein (2004a). For references to useful software packages and other Web sites, we highly recommend visiting the Phylogeny Programs Web site (Felsenstein 2006), which describes and provides hotlinks to 265 phylogeny packages. Other useful resources include the Phylogenetics Software Resources (UCMP 2006) and Bio Net-Book (Institute Pasteur 2006) Web sites.

Perspective

Aquaculture as a field of science dates back to about 1970. A self-identifying field of "genomics" dates back to about the mid-1980s. Development of our workhorse procedure, polymerase chain reaction, dates to the late 1980s (Saiki et al. 1988). Development of microsatellite markers dates to the 1990s, giving rise to a burst of development of newer, more powerful statistical analyses starting in the late 1990s, with subsequent development of new statistical software packages. The upshot is that we, as aquaculture geneticists, find ourselves with scientific investigative power that was unimaginable only a few years ago. We see no reason that the rapid growth at the interface of aquaculture and genomics should not continue. While the underlying population genetic and phylogenetic principles that we convey here will remain current, the latest developments in computational resources will change rapidly. Hence, we encourage interested readers to keep current on new developments by searching for papers and software that cite and build on the key articles that we have described. We hope that our presentation here sparks application of these exciting, new tools to aquaculture problems.

Acknowledgments

We thank John Liu for challenging us to present our thinking on developments at the interface of aquaculture genomics and population genetics/phylogeny. Eric Hallerman's research in this area is supported, in part, by the U.S. Department of Agriculture CSREES Hatch program.

References

ABS. 2006. Genemapper, version 4.0. Applied Biosystems, Inc., Foster City, CA. https://products.appliedbiosystems.com.

Allendorf FW and SR Phelps. 1980. Loss of genetic variation in a hatchery stock of cutthroat trout. Trans Amer Fish Soc, 109, pp. 537–543.

Appleyard SA and RD Ward. 2006. Genetic diversity and effective population size in mass selection lines of Pacific oyster (*Crassostrea gigas*). Aquaculture, 254, pp. 148–159.

Applied Maths. 2006. GelCompar II. Applied Maths, BVBA, Sint-Martins-Latem, Belguim. http://www.applied-maths.com/gelcompar/gelcompar.htm.

Avise JC. 2000. Phylogeography: the history and formation of species. Harvard University Press, Cambridge, MA.

Ayres KL and DJ Balding. 1998. Measuring departures from Hardy-Weinberg: a Markov chain Monte Carlo method for estimating the inbreeding coefficient. Heredity, 80, pp. 769–777.

Balloux F, W Amos, and T Coulson. 2004. Does heterozygosity estimate inbreeding in natural populations? Molec Ecol, 13, pp. 3021–3031.

Balloux F and N Lugon-Moulin. 2002. The estimation of population differentiation with microsatellite markers. Molec Ecol, 11, pp. 155–165.

Banks M and W Eichert. 2000. WHICHRUN, ver. 3.2: a computer program for population assignment of individuals based on multilocus genotype data. J Hered, 91, pp. 87–89. http://www.bml.ucdavis.edu/whichrun.htm.

Bartley D, M Bagley, G Gall, and B Bentley. 1992. Use of linkage disequilibrium data to estimate effective size of hatchery and natural fish populations. Cons Biol, 6, pp. 365–375.

Beaumont MA. 1999. TM2.exe. m.beaumont@ucl.ac.uk.

Beerli P and J Felsenstein. 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics, 152, pp. 763–773.

Bell GI and J Jurka. 1997. The length distribution of perfect dimmer repetitive DNA is consistent with its evolution by an unbiased singe-step mutation process. J Molec Evol, 44, pp. 414–421.

Bertorelle G. 1998. ADMIX1_0: computing admixture coefficients from molecular data. Department of Integrative Biology, University of California—Berkeley, http://web.unife.it/progetti/genetica/Giorgio/docadmix.html.

Bertorelle G and L Excoffier. 1998. Inferring admixture proportions from molecular data. Molec Biol Evol, 10, pp. 1298–1311.

Bondari K. 1984. Growth comparison of inbred and randombred catfish at different temperatures. Proc Southeast Assoc Fish Wildl Agencies, 35(1981), pp. 547–553.

Bondari K and RA Dunham. 1987. Effects of inbreeding on economic traits of channel catfish. Theoret Appl Genet, 74, pp. 1–9.

Bowcock AM, A Ruiz-Linares, J Tomfohrdo, E Minch, JR Kidd, and LL Cavalli-Sforza. 1994. High resolution of human evolutionary trees with polymorphic microsatellites. Nature, 368, pp. 455–457.

Brookfield JF and PM Sharp. 1994. Neutralism and selectionism face up to DNA data. Trends Genet. 10, pp. 109–111.

Brown RC, JA Wooliams, and BJ McAndrew. 2005. Factors influencing effective population size in commercial populations of gilthead sea bream, *Sparus aurata*. Aquaculture, 247, pp. 219–225.

Brzustowski J. 2002. Doh assignment test calculator. Department of Biology, University of Alberta, Edmonton. http://www2.biology.alberta.ca/jbrzusto/Doh.php.

Buntjer J. 2004. Phyltools. Laboratory of Plant Breeding, Wageningen University, Wageningen, the Netherlands. http://www.dpw.wau.nl/pv/PUB/pt.

Canino MF and P Bentzen. 2004. Evidence for positive selection at the pantophysin I locus in walleye pollack, *Theragra chalcogramma*. Molec Biol Evol, 21, pp. 1391–1400.
Cavalli-Sforza LL and AWF Edwards. 1967. Phylogenetic analysis: models and estimation procedures. Amer J Hum Genet, 19, pp. 233–257.
Chapko W. 1976. An exact test of the Hardy-Weinberg law. Biometrics, 32, pp. 183–189.
Chistiakov DA, B Hellemans, and FAM Volckaert. 2006. Microsatellites and their genomic distribution, evolution, function, and applications: a review with special reference to fish genetics. Aquaculture, 255, pp. 1–29.
Cockerham CC. 1969. Variance of gene frequencies. Evolution, 23, pp. 72–84.
Cockerham CC. 1973. Analysis of gene frequencies. Genetics, 74, pp. 679–700.
Coltman DW, WD Bowen, and JM Wright. 1998. Birth weight and neonatal survival of harbor seal pups are positively correlated with genetic variation measured by microsatellites. Proc Royal Acad Lond Ser B, 256, pp. 803–809.
Cornuet J-M and G Luikart. 1996. Description and power analysis of two tests for detecting recent demographic bottlenecks from allele frequency data. Genetics, 144, pp. 2001–2014.
Cornuet J-M, S Piry, G Luikart, A Estoup, and M Solignac. 1999. New methods employing multilocus genotypes to select or exclude populations as origins of individuals. Genetics, 153, pp. 1989–2000.
Coulson TN, JM Pemberton, JD Albon, M Beaumont, TC Marshall, J Slate, FE Guiness, and TH Clutton-Brock. 1998. Microsatellites reveal heterosis in red deer. Proc Royal Acad Lond Ser B, 256, pp. 489–495.
Crandall KA, D Posada, and D Vasco. 1999. Effective population sizes: missing measures and missing concepts. Anim Cons, 2, pp. 317–319.
Cross TF and J King. 1983. Genetic effects of hatchery rearing in Atlantic salmon. Aquaculture, 33, pp. 33–40.
Cruden D. 1949. The computation of inbreeding coefficients. J Hered, 40, pp. 248–251.
Cui JZ, XY Shen, GP Yang, QL Gong, and QQ Gu. 2005. Characterization of microsatellite DNA in *Takifugu rubripes* genome and their utilization in the genetic diversity analysis of *T. rubripes* and *T. pseudommus*. Aquaculture, 250, pp. 129–137.
Debevec EM, RB Gates, M Masuda, J Pella, JJ Reynolds, and LW Seeb. 2000. SPAM, version 3.2: Statistic program for analyzing mixtures. J Hered, 91, pp. 509–511. http://www.cf.adfg.state.ak.us/geninfo/research/genetics/Software/SpamPage.htm.
Dempson JB and M Power. 2004. Use of stable isotopes to distinguish farmed from wild Atlantic salmon, *Salmo salar*. Ecol Freshwat Fish, 13, pp. 176–184.
Dobzhansky T. 1970. Genetics of the evolutionary process. Columbia University Press, New York.
Einum S and IA Fleming. 1997. Genetic divergence and interactions in the wild among native, farmed, and hybrid Atlantic salmon. J Fish Biol, 50, pp. 634–651.
Emik LO and CE Terrill. 1949. Systematic procedures for calculating inbreeding coefficients. J Hered, 40, pp. 51–55.
Ewens WJ. 1972. The sampling theory of selectively neutral alleles. Theoret Pop Biol, 3, pp. 87–112.
Excoffier L, PE Smouse, and JM Quattro. 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: Applications to human mitochondrial DNA restriction data. Genetics, 131, pp. 479–491.
Excoffier LG, S Schneider, and D Roessli. 2006. Arlequin, version 3.0: A software for population genetics data analysis. http://lgb.unige.ch/arlequin.
Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Molec Evol, 17, pp. 368–376.
Felsenstein J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution, 39, pp. 783–791.
Felsenstein J. 1992. Estimating effective population size from samples of sequences: a bootstrap Monte Carlo integration method. Genet Res Cambridge, 60, pp. 209–220.

Felsenstein J. 2004a. Inferring Phylogenies. Sinauer Associates, Sunderland, MA. 664 pp.
Felsenstein J. 2004b. PHYLIP (Phylogeny Inference Package) version 3.6. Department of Genome Sciences, University of Washington, Seattle. http://evolution.gs.washington.edu/phylip.html.
Felsenstein J. 2006. Phylogeny programs. http://evolution.genetics.washington.edu/phylip/software.html.
Fleming IA, K Hindar, IB Mjolnerod, B Jonsson, T Balstad, and A Lamberg. 2000. Lifetime success and interactions of farm salmon invading a native population. Proc Royal Acad Lond Ser B, 267, pp. 1517–1523.
Fleming IA, B Jonsson, and MR Gross. 1994. Phenotypic divergence of sea-ranched, farmed and wild salmon. Canad J Fish Aquat Sci, 51, pp. 2808–2824.
Ford MJ. 2000. Effects of natural section on patterns of DNA sequence variation at the transferring, somatolactin, and *p53* genes within and among Chinook salmon (*Oncorhynchus tshawhtscha*) populations. Molec Ecol, 9, pp. 843–855.
Ford MJ. 2001. Molecular evolution of transferring: evidence for positive selection in salmonids. Molec Biol Evol, 18, pp. 639–647.
Ford MJ. 2002. Applications of selective neutrality tests to molecular ecology. Molec Ecol, 11, pp. 1245–1262.
Ford MJ, PJ Thornton, and LK Park. 1999. Natural selection promotes divergence of transferrin among salmonid species. Molec Ecol, 8, pp. 1055–1061.
Fu YX. 1994a. A phylogenetic estimator of effective population size or mutation rate. Genetics, 136, pp. 685–693.
Fu YX. 1994b. Estimating effective population size or mutation rate using the frequencies of mutations of various classes in a sample of DNA sequences. Genetics, 138, pp. 1375–1386.
Fu YX. 1997. Statistical tests of neutrality of mutations against population growth, hitchhiking, and background selection. Genetics, 147, pp. 915–925.
Gall GAE. 1987. Inbreeding. Pages 47–87 In N. Ryman and F. Utter, Eds. Population Genetics and Fishery Management. University of Washington Press, Seattle.
Gene Codes Corporation. 2006. Sequencher, version 4.6. Gene Codes Corporation, Anne Arbor, MI. http://www.genecodes.com/sequencher.
Goodman SJ. 1997. RSTCALC: A collection of computer programs for calculating estimates of genetic differentiation from microsatellite data and determining their significance. Molec Ecol, 6, pp. 881–885. http://helios.bto.ed.ac.uk/evolgen/rst/rst/html.
Goudet J. 2002. FSTAT. Department of Ecology and Evolution, Lausanne University, Lausanne, Switzerland. http://www2.unil.ch/popgene/softwares/fstat.htm.
Grobler PJ, JW Jones, NA Johnson, B Beaty, J Struthers, RJ Neves, and EM Hallerman. 2006. Patterns of genetic differentiation and conservation of the slabside pearlymussel, *Lexingtonia dollabelloides* (Lea 1840) in the Tennessee River drainage. J Mollusc Stud, 72, pp. 65–75.
Gross M. 1998. One species with two biologies: Atlantic salmon (*Salmo salar*) in the wild and in aquaculture. Canad J Fish Aquat Sci, 55(Suppl. 1), pp. 131–144.
Guo SW and EA Thompson. 1992. Performing the exact test of Hardy-Weinberg proportions for multiple alleles. Biometrics, 48, pp. 361–372.
Haldane JBS. 1954. An exact test for randomness of mating. J Genet, 52, pp. 631–635.
Hall BG. 2004. Phylogenetic trees made easy: a how-to manual, second edition. Sinauer Associates, Sunderland, MA.
Hall TA. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser, 41, pp. 95–98. http://www.mbio.ncsu.edu/BioEdit/bioedit.html.
Hallerman EM, Ed. 2003. Population Genetics: Principles and Applications for Fisheries Scientists. American Fisheries Society, Bethesda, MD.
Hallerman EM, RA Dunham, and RO Smitherman. 1986. Selection or drift—isozyme allele frequency changes among channel catfish selected for rapid growth. Trans Amer Fish Soc, 115, pp. 60–68.

Hard JJ, BA Berejikian, EP Tezak, SL Schroder, CM Knudsen, and LT Parker. 2000. Evidence for morphometric differentiation of wild and captively reared adult coho salmon: a geometric analysis. Env Biol Fish, 58, pp. 61–73.

Hardy GH. 1908. Mendelian proportions in a mixed population. Science, 28, pp. 49–50.

Hedgecock D, V Chow, and RS Waples. 1992. Effective population numbers of shellfish broodstocks estimated from temporal variance in allele frequencies. Aquaculture, 108, pp. 215–232.

Hedgecock D and FL Sly. 1990. Genetic drift and effective population size of hatchery-propagated stocks of the Pacific oyster *Crassostrea gigas*. Aquaculture, 88, pp. 21–38.

Hedrick PW. 2005. Genetics of populations, third edition. Jones and Bartlett Publishers, Sudbury, MA.

Hill RE and ND Hastie. 1987. Accelerated evolution in the reactive centre regions of serine protease inhibitors. Nature, 326, pp. 96–99.

Hill WG. 1979. A note on effective population size with overlapping generations. Genetics, 92, pp. 317–322.

Hill WG. 1981. Estimation of effective population size from data on linkage disequilibrium. Genet Res, 38, pp. 209–216.

Hillis DM and C Moritz, Eds. 1990. Molecular systematics. Sinauer Associates, Sunderland, MA.

Hindar K, N Ryman, and F Utter. 1991. Genetic effects of cultured fish on natural fish populations. Canad J Fish Aquat Sci, 48, pp. 945–957.

Huelsenbeck JP, F Ronquist, R Nielsen, and JP Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science, 294, pp. 2310–2314.

Hughes AL and M Nei. 1988. Patterns of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature, 335, pp. 167–170.

Institute Pasteur. 2006. Bio NetBook. http://www.pasteur.fr/recherche/BNB/bnb-en.html.

Jorde PE and N Ryman. 1995. Temporal allele frequency change and estimation of effective size in populations with overlapping generations. Genetics, 139, pp. 1077–1090.

Jukes TH and CR Cantor. 1969. Evolution of protein molecules. Pages 21–132 In HN Munro, Ed. Mammaliam Protein Metabolism. Academic Press, New York.

Kalinowski S. 2003. Genetic mixture analysis. Department of Ecology, Montana State University, Bozeman. http://www.montana.edu/kalinowski/GMA/kalinowski_GMA.htm.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Molec Evol, 16, pp. 111–120.

Kimura M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, UK.

Kimura M and JF Crow. 1964. The number of alleles that can be maintained in a finite population. Genetics, 49, pp. 725–738.

Kimura M and T Ohta. 1978. Stepwise mutational model and distribution of allele frequencies in a finite population. Proc Nat Acad Sci USA, 75, pp. 2868–2872.

Kincaid HL. 1976a. Effects of inbreeding on rainbow trout populations. Trans Amer Fish Soc, 105, pp. 273–280.

Kincaid HL. 1976b. Inbreeding in rainbow trout (*Salmo gairdneri*). J Fish Res Bd Canad, 33, pp. 2420–2426.

Kincaid HL. 1983. Inbreeding in fish populations used in aquaculture. Aquaculture, 33, pp. 215–227.

Kohlmann K, K Kersten, and M Flajhans. 2005. Microsatellite-based genetic variability and differentiation of domesticated, wild and feral common carp (*Cyprinus carpio* L.) populations. Aquaculture, 247, pp. 253–266.

Kreitman M and H Akashi. 1995. Molecular evidence for natural selection. Annu Rev Ecol Systemat, 26, pp. 403–422.

Kuhner MK, J Yamato, and P Beerli. 2004. LAMARC, version 1.2.1. University of Washington, Seattle. http://evolution.gs.washington.edu/lamarc.html.

Kuhner MK, J Yamato, and J Felsenstein. 1995. Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics, 140, pp. 1421–1430.

Kumar S, K Tamura, and M Nei. 2004. MEGA 3: Integrated software for molecular evolutionary genetics analysis and sequence alignment. Briefings Bioinformat, 5, pp. 150–163. http://www.megasoftware.net/.

Levene H. 1949. On a matching problem arising in genetics. Annals Mathemat Stat, 20, pp. 91–94.

Lewontin RC and J Krakauer. 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics, 75, pp. 175–195.

Li CC. 1955. Population genetics. University of Chicago Press, Chicago, IL.

Li WH. 1997. Molecular evolution. Sinauer Associates, Sunderland, MA.

Longwell AC and S Stiles. 1973. Gamete cross incompatibility and inbreeding in the commercial oyster, *Crassostrea virginica* Gmelin. Cytologia, 38, pp. 521–533.

Luikart G and J-M Cornuet. 1999. Estimating the effective number of breeders from heterozygote excess in progeny. Genetics, 151, pp. 1211–1216.

Luikart G and PR England. 1999. Statistical analysis of microsatellite DNA data. Trends Ecol Evol, 14, pp. 253–256.

Lund RA, LP Hansen, and T Jarvi. 1989. Identification of reared and wild salmon by external morphology, size of fins and scale characteristics. NINA Forskningsrapp, 1, pp. 1–54.

Lund RA, PJ Midtlyng, and P Hansen. 1997. Post-vaccination intra-abdominal adhesions as a marker to identify Atlantic salmon, *Salmo salar* L., escaped from commercial fish farms. Aquaculture, 154, pp. 27–37.

Lundrigan TA, JD Reist, and MM Ferguson. 2005. Microsatellite genetic variation within and among Arctic charr (*Salvelinus alpinus*) from aquaculture and natural populations in North America. Aquaculture, 244, pp. 63–75.

Lura H and H Saegrov. 1991. A method of separating offspring from farmed and wild Atlantic salmon (*Salmo salar*) based on different ratios of optical isomers of astaxanthin. Canad J Fish Aquat Sci, 48, pp. 429–433.

McDonald JH and M Kreitman. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature, 354, pp. 114–116.

McElroy D, P Moran, E Bermingham, and I Kornfield. 1992. REAP: an integrated environment for the manipulation and phylogenetic analysis of restriction data. J Hered, 83, pp. 157–158. http://bioweb.wku.edu/faculty/mcelroy.

McGinnity P, P Prodohl, A Ferguson, R Hynes, NO Maoilieidigh, N Baker, D Cotter, B O'Hea, D Cooke, G Rogan, J Taggart, and T Cross. 2003. Fitness reduction and potential extinction of wild populations of Atlantic salmon, *Salmo salar*, as a result of interactions with escaped farmed salmon. Proc Royal Acad Lond Ser B, 270, pp. 2443–2450.

Michalakis Y and L Excoffier. 1996. A generic estimation of population subdivision using distances between alleles with special interest to microsatellite loci. Genetics, 142, pp. 1061–1064.

Miller M. 1997. Tools for population genetic analyses (TFPGA) 1.3. http://marksgenetic software.net/tfpga.htm.

Minch E, A Ruiz-Linares, D Goldsetin, M Feldman, and LL Cavalli-Sforza. 1999. MICROSAT, version 1.5. http://hpgl.stanford.edu/projects/microsat.

Nakajima M, A Kita, and Y Fujio. 1986. Genetic features of natural and cultured populations in masu salmon (*Oncorhynchus masou*). Tohuku J Agric Res, 37, pp. 31–42.

Nei M. 1972. Genetic distance between populations. Amer Nat, 106, pp. 283–292.

Nei M. 1973a. Analysis of gene diversity in subdivided populations. Proc Nat Acad Sci USA, 70, pp. 3321–3323.

Nei M. 1973b. The theory and estimation of genetic distance, pp. 45–54 In NE Morton, Ed. Genetic structure of populations. University Press of Hawaii, Honolulu.

Nei M. 1977. *F*-statistics and analysis of gene diversity in subdivided populations. Annals Hum Genet, 41, pp. 225–233.

Nei M. 1987a. Genetic distance and molecular phylogeny, pp. 193–223 In N Ryman and F Utter, Eds. Population genetics and fishery management. University of Washington Press, Seattle.

Nei M. 1987b. Molecular Evolutionary Genetics. Columbia University Press, New York.

Nei M and D Graur. 1984. Extent of protein polymorphism and the neutral mutation theory. Evol Biol, 17, pp. 73–118.

Nei M and S Kumar. 2000. Molecular Evolution and Phylogenetics. Oxford University Press, New York.

Nei M and WH Li. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Nat Acad Sci USA, 76, pp. 5269–5273.

Nei M and F Tajima. 1981. Genetic drift and estimation of effective population size. Genetics, 98, pp. 625–640.

Ngal J, MM Dowling, I Buck, R Axel, and A Chess. 1993. The family of genes encoding odorant receptors in the channel catfish. Cell, 72, pp. 657–666.

Nielsen R, JL Mountain, JP Huelsenbeck, and M Slatkin. 1998. Maximum likelihood estimation of population divergence times and population phylogeny in models without mutation. Evolution, 52, pp. 660–677.

O'Ryan C, EH Harley, MW Bruford, M Beaumont, RK Wayne, and MI Cherry. 1998. Microsatellite analysis of genetic diversity in fragmented South African buffalo populations. Anim Cons, 1, pp. 85–94.

Page RDM. 2000. TREEVIEW, version 1.5, Division of Environmental and Evolutionary Biology, Institute of Biomedical and Life Sciences, University of Glasgow, Glasgow, Scotland (visited April 24, 2006). http://taxonomy.zoology.gla.ac.uk/rod/treeview/treeview.html.

Park SDE. 2001. Trypanotolerance in West African cattle and the population genetic effects of selection. Ph.D. Dissertation, University of Dublin. http://acer.gen.tcd.ie/~sdepark/ms_toolkit.

Piry S, A Alapetite, J-M Cornuet, D Paetkau, L Bauouin, and A Estoup. 2004. GeneClass2: a software for genetic assignment and first generation migrant detection. J Hered, 95, pp. 536–539.

Piry S, G Luikart, and J-M Cornuet. 1999. Bottleneck: a computer program for detecting recent reductions in the effective population size using allele frequency data. J Hered, 90, pp. 502–503.

Pogson GH. 2001. Nucleotide polymorphism and natural selection at the pantophysin (*Pan I*) locus in the Atlantic cod, *Gadus morhua* (L.). Genetics, 157, pp. 317–330.

Pogson GH and KA Mesa. 2004. Positive Darwian selection at the pantophysin (*Pan I*) locus in marine gadid fishes. Molec Biol Evol, 21, pp. 65–75.

Pollack E. 1983. A new method for estimating the effective population size from allele frequency changes. Genetics, 104, pp. 531–548.

Posada D. 2006. MODELTEST: a tool to select the best-fit model of nucleotide substitution. http://darwin.uvigo.es/software/modeltest.html.

Posada D and KA Crandall. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics, 14, pp. 817–818.

Pritchard JK, M Stephens, and P Donelly. 2000. Inference of population structure using multilocus genotype data. Genetics, 155, pp. 945–959.

Pudovkin AI, DV Zaykin, and D Hedgecock. 1996. On the potential for estimating the effective number of breeders from heterozygote-excess in progeny. Genetics, 144, pp. 383–387.

Rannala B and JL Mountain. 1997. Detecting immigration by using multilocus genotypes. Proc Nat Acad Sci USA, 94, pp. 9197–9201.

Raymond M and F Rousset. 1995. GENEPOP, version 1.2: population genetics software for exact tests and ecumenicism. http://www.cefe.cnrs-mop,fr/.

Rogers AR and H Harpending. 1992. Population growth makes waves in the distribution of pairwise genetic differences. Mol Biol Evol, 9, pp. 552–569.

Ronquist F and JP Huelsenbeck. 2003. MRBAYES3: Bayesian phylogenetic inference under mixed models. Bioinformatics, 19, pp. 1572–1574. http://mrbayes.csit.fsu.edu.

Rosenberg MS. 2001. PASSAGE. Pattern analysis, spatial statistics, and geographic exegenis, version 1.0. Department of Biology, Arizona State University, Tempe, AZ. http://www.passagesoftware.net.

Rozas J and R Rozas. 1995. DnaSP, DNA sequence polymorphism: an interactive program for estimating population genetics parameters from DNA sequence data. Comp Appl Biosci, 11, pp. 621–625. http://www.ub.es/dnasp/.

Saccheri IL, IJ Wilson, RA Nichols, MW Bruford, and PM Brakefield. 1999. Inbreeding of bottlenecked butterfly populations: estimation using the likelihood of changes in marker allele frequencies. Genetics, 151, pp. 1053–1063. r.a.nichols@qmw.ac.uk.

Saiki RK, DH Gelfland, S Stoffel, SJ Scharf, R Higuchi, GT Horn, KB Mullis, and HA Erlich. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science, 239, pp. 487–491.

Schlotterer C and D Tautz. 1992. Slippage synthesis of simple sequence DNA. Nucl Acids Res, 20, pp. 211–215.

Schwartz MK, DA Tallmon, and G Luikart. 1998. Review of DNA-based census and effective population size estimators. Anim Cons, 1, pp. 293–299.

Schwartz MK, DA Tallmon, and G Luikart. 1999. Using genetics to estimate the size of wild populations: many methods, much potential, uncertain utility. Anim Cons, 2, pp. 321–323.

Senanan W, AR Kapuscinski, U Na-Nakorn, and LM Miller. 2004. Genetic impacts of hybrid catfish farming (*Clarias macrocephalus* x *C. gariepinus*) on native catfish populations in central Thailand. Aquaculture, 235, pp. 167–184.

Shaklee JB and KP Currens. 2003. Genetic stock identification and risk assessment, pp. 291–328 In EM Hallerman, Ed. Population genetics: principles and applications for fisheries scientists. American Fisheries Society, Bethesda, MD.

Simmons M, K Mickett, H Kucuktas, P Li, R Dunham, and Z Liu. 2006. Comparison of domestic and wild channel catfish (*Ictalurus punctatus*) populations provides no evidence for genetic impact. Aquaculture, 252, pp. 122–146.

Skaala O, B Hoyheim, K Glover, and G Dahle. 2004. Microsatellite analysis in domesticated and wild Atlantic salmon (*Salmo salar* L.): allelic diversity and identification of individuals. Aquaculture, 240, pp. 131–143.

Slatkin M. 1995. A measure of population subdivision based on microsatellite allele frequencies. Genetics, 139, pp. 457–462.

Soll Technologies. 2006. DENDRON. Soll Technologies, Iowa City, Iowa. http://www.geocities.com/solltech/dendron/.

Swofford DL. 1998. PAUP*: Phylogenetic Analysis Using Parsimony (and other methods). Sinauer Associates, Sunderland, MA. http://paup.csit.fsu.edu/.

Tajima F. 1989. Statistical methods for testing the neutral mutation hypothesis by DNA polymorphisms. Genetics, 123, pp. 585–595.

Tajima F. 1992. Statistical method for estimating the effective population size in Pacific salmon. J Hered, 83, pp. 309–311.

Tajima F and M Nei. 1984. Estimation of evolutionary distance from between nucleotide sequences. Molec Biol Evol, 1, pp. 269–285.

Tave D. 1993. Genetics for Fish Hatchery Managers, second edition. Van Nostrand Reinhold, New York.

Thompson JD, TJ Gibson, F Plewniak, F Jeanmougin, and DG Higgins. 1997. ClustalW multiple sequence alignment for DNA and proteins. Nucl Acids Res, 24, pp. 4786–4882.

UCMP (University of California Museum of Paleontology). 2006. Phylogenetics software resources. http://www.ucmp.berkeley.edu/subway/phylo/phylosoft.html.

Utter F. 2003. Genetic impacts of fish introductions, pp. 357–378 In EM Hallerman, Ed. Population Genetics: Principles and Applications for Fisheries Scientists. American Fisheries Society, Bethesda, MD.

Vuorinen J. 1984. Reduction of genetic variability in a hatchery stock of brown trout, *Salmo trutta*. J Fish Biol, 24, pp. 339–348.

Wang H, X Guo, G Zhang, and F Zhang. 2004. Classification of jinjiang oysters *Crassostrea rivularis* (Gould, 1861) from China, based on morphology and phylogenetic analysis. Aquaculture, 242, pp. 137–155.

Waples RS. 1989. A generalized method for estimating population size from temporal changes in allele frequency. Genetics, 121, pp. 379–391.

Waples RS. 1990. Conservation genetics of Pacific salmon. 3. Estimating effective population size. J Hered, 81, pp. 277–289.

Waples RS. 1991. Genetic interactions between hatchery and wild salmonids: lessons from the Pacific Northwest. Canad J Fish Aquat Sci, 48(Suppl. 1), pp. 124–133.

Waples RS and DJ Teel. 1990. Conservation genetics of Pacific salmon. 1. Temporal changes in allele frequency. Cons Biol, 4, pp. 144–156.

Waser P and C Strobeck. 1998. Genetic signatures of interpopulation dispersal. Trends Ecol Evol, 13, pp. 43–44.

Watterson GA. 1978. The homozygosity test of neutrality. Genetics, 88, pp. 415–417.

Weinberg W. 1908. Uber den nachweis der vererbung beim menschen. Jahreshefte Verein, Naturk Wurtemberg, 64, pp. 368–382.

Weir BS and CC Cockerham. 1984. Estimating *F*-statistics for the analysis of population structure. Evolution, 38, pp. 1358–1370.

Wright S. 1931. Evolution in Mendelian populations. Genetics, 16, pp. 97–159.

Wright S. 1938. Size of populations and breeding structure in relation to evolution. Science, 87, pp. 430–431.

Wright S. 1965. The interpretation of population structure by *F*-statistics with special regard to systems of mating. Evolution, 19, pp. 395–420.

Yap I and R Nelson. 2006. Winboot. International Rice Research Institute. http://www.irri.org/science/software/winboot.asp.

Yeh FC, T Yang, and T Boyle. 1999. POPGENE, ver. 3.31. Microsoft Windows based freeware for population genetic analysis. http://www.ualberta.ca/~fyeh/.

Yu DH and KH Chu. 2006. Genetic variation in wild and cultured populations of the pearl oyster *Pictada fucata* from southern China. Aquaculture, 258, pp. 220–227.

Zhu ZY, G Lin, LC Lo, YX Xu, F Feng, R Chou, and GH Yue. 2006. Genetic analysis of Asian seabass stocks using novel polymorphic microsatellites. Aquaculture, 256, pp. 167–173.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Part 2
Mapping Genomes

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 10
Linkage Mapping in Aquaculture Species

Roy G. Danzmann and Karim Gharbi

Introduction

Linkage mapping is an exercise in 'jigsaw puzzle × n assembly,' where the n refers to multiple separate linkage groups. Although complications in linkage mapping can arise (e.g., anomalies in the modes of genetic transmission of certain genetic markers [i.e., segregation distortion], or differences in the rates of genetic recombination between individuals or between the sexes within a species), linkage mapping is a relatively simple three-step process, and adherence to these steps can result in the effective 'piece-meal' construction of the jigsaw. In addition, a jigsaw puzzle is perhaps not the best analogy to use for this process, given the fact that jigsaw puzzles are two-dimensional unknowns, while linkage maps are essentially one-dimensional structures. The three iterative processes that through time build a complete linkage map follow:

1. The calculation of the minimum two-point recombination distances between all pairs of genetic markers (i.e., identification of linkage groups)
2. The ordering of these markers in relation to their proper phase vector along the linear length of a linkage group
3. The calculation of the map distances separating the markers along the length of the linkage group

The first process facilitates the assignment of new markers to their respective linkage groups. The second process may use information on the orientation or phase of marker alleles that have already been placed within a linkage group to assign the phase of newly genotyped markers, or the phase may be established through knowledge of the transmission of alleles from grandparents. The determination of phase assignment is based upon observed recombination distances between marker pairs, and recombination distances form the basis for establishing map distances among the sets of markers within the linkage group. Linkage mapping could therefore be considered a process of ordering the linkage phase of a series of markers along the length of a linkage group to achieve a minimum number of recombination events along the length of the linkage group. Thus, among several different orders of markers that may be obtained for any given linkage group, the one that minimizes the total map length (i.e., reduces the total number of recombination events) of the linkage group can generally be taken as the most correct order for the placement of markers.

Researchers can produce different types of genetic linkage maps. Classical genetic mapping procedures involve genotyping parents and offspring, while more elaborate methods may involve map construction based upon radiation hybrid mapping panels. This chapter will focus on the former method of linkage map construction. Classical

linkage mapping can involve the use of normal genomic segregation patterns derived from backcross, recombinant inbred, F_2 intercross, outcross mapping panels, and pedigreed multigenerational line information. It may also involve the generation of artificial genomes through the processes of haploid gynogenesis and/or androgenesis (Babiak et al. 2002), whereby two homozygous lines may be crossed and a resulting hybrid line between two parental homozygous lines is in turn gynogenetically or androgenetically manipulated to produce homozygous doubled haploids with all-maternal or all-paternal inheritance. These progeny can then be used to directly score meiotic events between the two parental haploid vectors that are present in the hybrid parent (see Pandian and Kirankumar 2003 for a review of androgenesis in fishes). The biggest advantage to using doubled haploid mapping panels is that all doubly heterozygous marker positions in the genome will be fully informative with respect to the construction of a single genetic map, regardless of the marker type (i.e., dominant versus codominant markers). With other types of classical genetic mapping data, both parental genomes contribute variable allelic information to the progeny used in a mapping panel, and thus sex-specific genetic maps may need to be constructed that are a heterogeneous mixture of all of the available polymorphic genetic data. Some markers will be informative in both parents while others will only be informative in either the male or female parent.

Most species that are intensively used for aquaculture production have large numbers of progeny. Therefore, it is possible to use single mapping panels for the construction of genetic maps in these species. The discussion that follows in this chapter is based on the assumption that large progeny datasets are available for the construction of genetic maps, and hence will focus on the methods that can be applied to either doubled haploid, or normal bisexual cross progeny datasets. Classical genetic mapping methods are equally applicable to both types of family structures. It is also assumed by the authors that the readers of this chapter will be familiar with the basic biological principles of the genetic transmission alleles from parents to offspring and understand the process of meiosis and recombination. Readers who are unfamiliar with these concepts are strongly encouraged to consult an introductory genetics textbook for an overview of these principles.

To begin a linkage study, it is necessary to produce one or more mapping panel families. A mapping panel can be produced by mixing the gametes from a single pair of parents. It is of course necessary to obtain a DNA sample from both parents. While not an absolute requirement, it is also recommended that both sets of grandparents be sampled in order to determine the phase of the marker alleles in the parents contributing the mapping panel. This sample may be obtained from any tissue in the individual, and it is recommended that a reserve tissue supply be frozen for future extraction of DNA. Additionally, the progeny produced for the mapping panel will need to be sampled at some point in their development to obtain tissue samples for future DNA extraction. The reader is referred to other sources for methods on DNA preparation and analysis. It is assumed that most of the genotypic data that will be analyzed for linkage analysis will be obtained using modern molecular methods involving polymerase chain reaction (PCR) of source DNA samples. Readers are referred to Part I of this book for DNA marker-related issues.

The minimal requirement for conducting a linkage test is obtaining genotypic data in the parents and progeny for two markers that are both heterozygous (i.e., have two

alleles present) in at least one of the two parents in the family. In practice of course, many more than two markers are required to build up knowledge on the chromosomal positions and linkage arrangement affinities of molecular markers. An important target of linkage mapping is to provide a framework for future in-depth molecular work involving DNA sequencing and physical mapping in the genome of the research species. To achieve this goal it may be necessary to genotype and map several thousand genetic markers. After a researcher has accumulated the genotypes of the parents involved in their mapping panels along with the genotypes of mapping panel progeny and arrayed them according to the input specifications of the software needed to perform the linkage analyses, it is possible to begin a formal analysis. Software resources available for conducting linkage analyses along with an overview of various associated features present within each package are given in the last section of this chapter.

Estimation of Recombination Distances and Mapping Functions

Using the simple principles of Mendelian inheritance, biologists have known since the beginning of the twentieth century that genetic markers tend to assort independently from one another, unless they are physically linked on a chromosome. Alfred Sturtevant, a graduate student in T. H. Morgan's laboratory in the early 1900s was the first individual to recognize that exceptions to Mendel's independent assortment of alleles were likely due to the physical proximity of genetic elements to one another that somehow prevented their independent segregation. He rationalized that two alleles would remain in combination with one another throughout the meiotic process, the closer they were to one another. Thus, recombination events between any given allele to any other chosen allele in the genome would be proportional to the distance they were apart on a chromosome. For unlinked genetic markers, this would simply approximate a ratio of 50% between any source allele at a polymorphic locus. For example, if we had two alleles, 'A' and 'a' at a heterozygous locus and 'B' and 'b' at another heterozygous locus, the 'A' allele would have an equal probability of being associated with allele 'B' or 'b' following meiosis II, if the two loci are unlinked. For physically linked alleles, however, this independent assortment expectation would not be obtained. Using the principle that crossover events along a chromosome that could reverse the phase of allelic combinations for a pair of linked marker alleles (i.e., produce a recombination event) were less likely the closer two alleles were to one another, it was possible for Sturtevant to devise a scheme whereby genetic markers could be arrayed along a chromosome according to their pair-wise recombination differences. More closely linked alleles would show fewer recombination events between one another than more distantly linked alleles. In our example above, if allele 'B' is tightly linked to allele 'A,' then AB and ab genotypic combinations would be seen in much higher frequency than their expectation according to random segregation. If allelic segregation occurs randomly at both loci and there is no linkage present, then AB and ab genotypic combinations are each expected to occur approximately 50% of the time in the progeny of a family, while the other two genotypic classes (i.e., Ab and aB) should also constitute approximately 50% of the progeny genotypes. If these latter two classes are largely underrepresented while the AB and ab genotypes constitute almost

all of the progeny genotypes, then it can be established that these two markers are indeed linked and that the likely parental phase of the linkage is A–B and a–b.

Thus, by simply ranking pair-wise recombination events between genetic markers on a scale from low (with zero recombination obviously demonstrating the tightest linkage) to high (approximating 50% recombination), it is possible to obtain the linear order of genes/markers along a chromosome, which is similar to the concept of 'beads-on-a-string' chromosome structure that Morgan and his coworkers envisioned for the early structure of genes arrayed on chromosomes. This beads-on-a-string metaphor is in fact a very good analogy to the true relationships among genetic elements along chromosomes, because genomic studies have amply demonstrated the linear one-dimensional relationships among genetic markers.

To obtain an estimate of the number of recombinant genotypes between any pair of genetic markers it is necessary to calculate the recombination fraction/ratio between the pair of markers being considered. This is simply a matter of summing the counts for the two most abundant genotypic classes (parentals) (remembering that both markers must be heterozygous to test for linkage, and thus four allelic configurations are expected in the progeny), and similarly summing the counts for the two least abundant genotypic classes (recombinants), as an estimate of recombination fraction. The recombination ratio/fraction, which is often termed theta (θ), can then simply be calculated as: recombinants/(parentals + recombinants), or as: recombinants/N, where N is the sample size of the mapping panel. An estimate of θ may also be obtained by directly looking at the number of 'breakpoints' in a genetic phase map, along the length of ordered markers in a linkage group. For example, if the best marker order for a linkage group is established, and the allelic phase of the markers within each mapping panel individual is obtained and arrayed according to the linear order of markers, then recombination points are easily observed (Figure 10.1) at each location where the allelic phase alternates within an individual. This is then counted as a recombination event (r), and the value of θ is obtained as: $\Sigma r/N$. Of course, if the phase of the alleles is known directly (i.e., obtained from knowledge of the grandparental alleles transmitted) then the recombinant classes can be established *a priori*.

The strength or likelihood of a true linkage is contingent upon two factors: (a) the sample size, or informative number of meioses, that contributed to the estimate; and (b) the observed θ for the marker pair being considered. The Logarithm of Odds Score or LOD score is a statistic that can be used to express the likelihood that two markers are linked using information based upon the two parameters just mentioned. The LOD score is calculated as:

$$\text{LOD} = nL_{10}(2) + rL_{10}(\theta) + (n - r)L_{10}(1 - \theta), \text{ if } 0<\theta\leq 0.50 \qquad (10.1)$$

or

$$\text{LOD} = nL_{10}(2), \text{ if } \theta = 0 \qquad (10.2)$$

where, r = observed number of recombinants
n = the pair-wise sample size used to obtain the linkage estimate
θ = the observed recombination fraction
L_{10} = log (to the base 10)

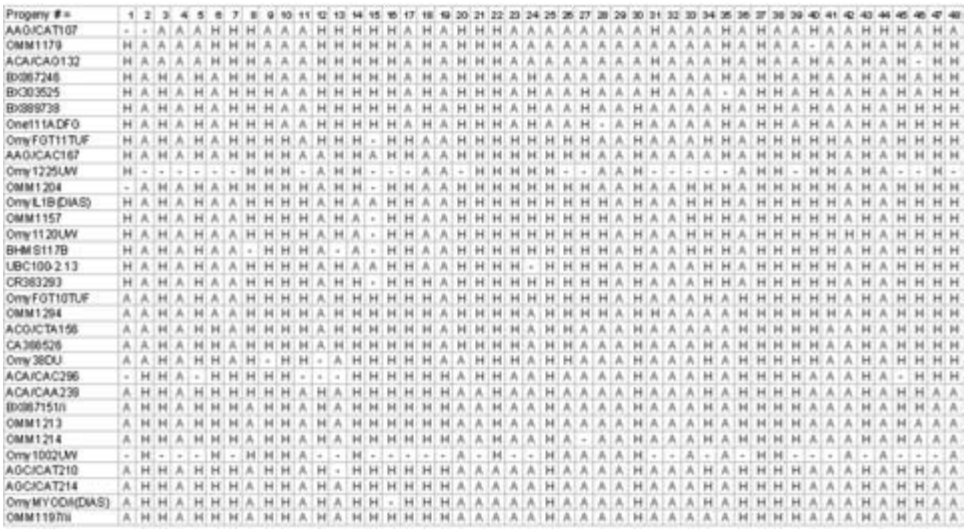


Figure 10.1. A phase map of linkage group 10 from the female genetic map in rainbow trout obtained from one of the mapping panels used by Danzmann and others (2005). The linear order of markers along the linkage group is indicated on the left hand side of the figure, and the two allelic phases for all the markers in the linkage group are scored either as 'A' or 'H' (corresponding to the two alleles that may be contributed by the female parent at any given genetic marker position). Each column represents the genotype obtained from a single offspring in the family for each marker, so that the entire column represents the phase vector for that individual. Adjacent markers along the length of the map will have a tendency to show greater affinities of one allele versus the other with their immediate adjacent neighbors (i.e., share a higher incidence of either the 'A' or 'H' allele with the neighboring marker). Recombination levels can be directly obtained from the phase map by simply summing the number of instances where a phase change occurs across all the progeny sampled. For example, the recombination level between marker OMM1204 and OmyIL1B(DIAS) would be 3/48 = 0.0625. Changes in phase along the phase vector of an individual represent the number of cross-over points that a particular chromosome experiences during meiosis prior to transmission in the offspring. (Also see color plate.)

Thus, when comparing two mapping panels, both mapping panels may have equivalent recombination fractions, but the panel with the larger sample size will have a larger associated LOD score with any given θ fraction. See Figure 10.2. Therefore, when planning the assignment of LOD thresholds for linkage analyses, consideration should be given to the sample size of the mapping panel. To minimize the instances where corrections to map distances may need to be made (i.e., to correct for the incidence of double and multiple crossover events that will reduce observed recombination estimates, see section below), LOD score thresholds should be chosen so that observed recombination fractions greater than 0.20–0.25 will not be placed within a linkage group cluster. Recombination fractions estimated to be around 0.25 may often underestimate the true map distance due to the increased probability of multiple crossovers occurring as map distances increase. A LOD threshold of 4.0 will be sufficient to minimize the chances of including markers with increased recombination

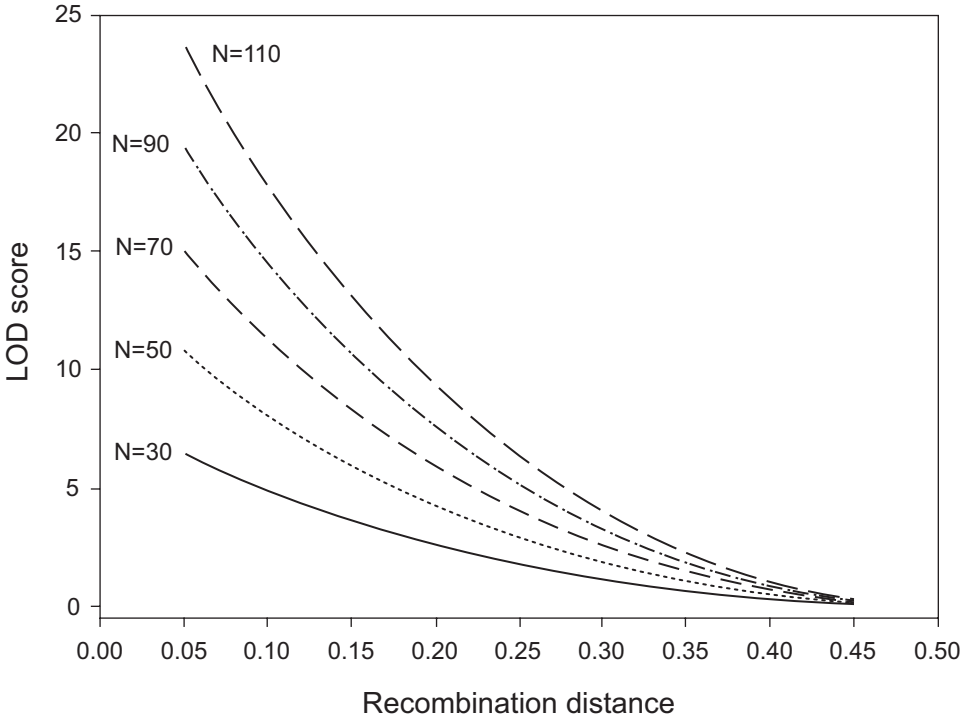


Figure 10.2. Distribution of LOD scores (y-axis) associated with varying levels of recombination (x-axis) that will be observed in mapping panels of varying sizes. Progeny sizes of 30, 50, 70, 90, and 110 progeny are shown.

fractions (i.e., $\theta > 0.25$) within moderate-sized mapping panels (i.e., N~50, or N>50). For smaller-sized mapping panels, an LOD threshold greater than 4.0 should be applied. See Figure 10.2.

The above discussion suggests that the estimation of genetic map distances will simply be a matter of calculating the cumulative recombination fractions between any point source marker and all other markers along the linkage group. This is of course not true given the fact that distantly linked markers will appear to be unlinked. Even for more proximal markers along the linkage group, consideration must be given to events such as double and higher orders of crossovers. Double crossovers present the problem of underrepresenting true genetic distances between a pair of markers since two crossover events will bring the parental phase alleles back into the same linkage phase and make it appear as if no recombination events have taken place along the length of the chromosome between these two alleles, while in fact, two distinct crossover events have occurred between these linked markers. Thus, as nearest neighbor recombination distances are obtained from a mapping panel, and are used to construct the linear order of markers in the linkage group, it is often observed that the cumulative map distances along the string of markers will exceed the recombination distance that is estimated between the two distal markers. For example, if d_1, d_2, d_3 represent the pair-wise recombination estimates between A–B, B–C, and C–D, respectively, and d_4 represents the recombination fraction between A–D, with A and D being the flanking pair of markers in the linkage cluster, it is often observed that: $(d_1 + d_2 + d_3) > d_4$. The degree markers

to which these distances differ from one another is dependent upon the frequency of multiple even crossovers between A and D. This phenomenon may also in turn be influenced to some extent by the actual size and physical structure of the chromosome (Qumsiyeh 1994). The degree to which chromosome structure, size, pairing affinities, and regulatory factors (Heyting 1996) can influence crossover events along a chromosome is termed interference and is recognized as any crossover event having an influence or inhibiting the occurrence of another nearby crossover event. This topic will be dealt with in the next section of the chapter.

In the above discussion it is important to note that an odd number of multiple crossovers will not reverse the allelic phase between the two markers being considered. However, an odd number of multiple crossovers between the most distal pair of markers may actually produce an even number of crossover events between markers that are intermediate to these two distal markers, which can in turn lead to a decrease in recombination estimates between markers intercalary to the flanking markers. See Figure 10.1.

The physical occurrence of crossover points or chiasma was historically thought to largely be a random process throughout the length of a chromosome, so that corrections for larger mapping intervals would be necessary to correct for the occurrence of double or multiple even numbered crossover events between a pair of markers. Several mapping function corrections have been proposed by past researchers, with the Haldane and Kosambi mapping functions still widely used by modern researchers. When map distances are simply reported as θ, or the observed recombination distance, this is known as the Morgan mapping distance. A review of these mapping functions is given in Ott (1999). The modified mapping distances (md) that can be estimated, from the two most common corrections available in most linkage mapping software programs follow:

The Haldane mapping function is given as:

$$\text{md} = -\tfrac{1}{2} \bullet \ln(1 - 2\theta), \text{ if } 0 \leq \theta < \tfrac{1}{2} \tag{10.3}$$

The Kosambi mapping function is given as:

$$\text{md} = \tfrac{1}{4} \bullet \ln((1 + 2\theta)/(1 - 2\theta)) \tag{10.4}$$

An additional mapping function that accounts for the incidence of crossover events along a chromosome is the binomial function (Liberman and Karlin 1984).

$$\text{md} = \tfrac{1}{2} \bullet x \bullet [1 - (1 - 2\theta)^{1/x}] \tag{10.5}$$

where, x = the modal number of crossover events that are detected along a linkage group. This mapping function provides an estimate that is similar to the Kosambi function when $x = 2$ and increases the map distance estimates as x increases (Figure 10.3).

Mapping functions serve to increase the mapping distance estimates between a pair of markers as a function of the overall θ distance that is estimated between the marker pairs. This is done in an effort to try to compensate for the anticipated multiple crossover events that are theoretically expected to occur in the interval that would reduce the estimate of the true genetic distances. For recombination distances in the region of 0–20%, θ, Kosambi, and Haldane mapping functions give fairly equivalent estimates. However,

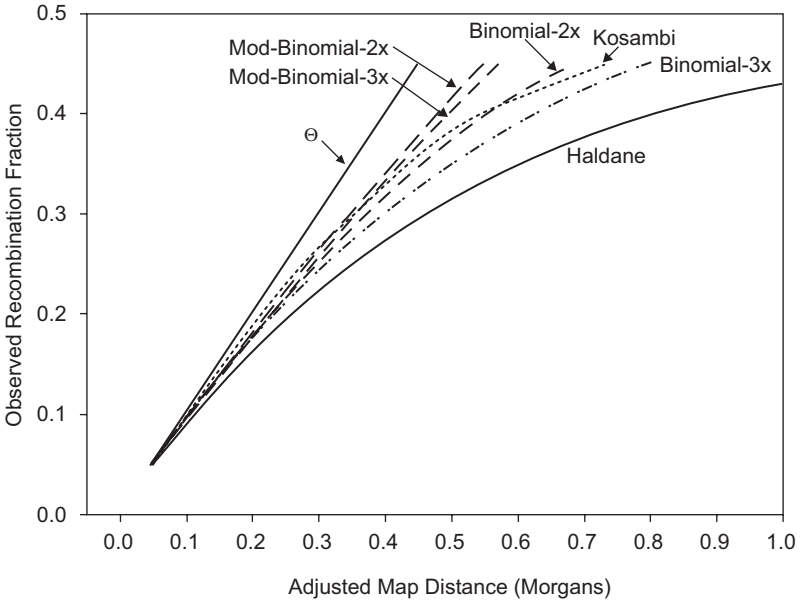


Figure 10.3. Estimation of readjusted recombination distances that will be obtained with various mapping functions.

one can see by examining Figure 10.3 that for distances exceeding $\theta = 0.20$, reestimates of genetic distance quickly increase, especially for the Haldane mapping function.

Given the fact that several recent cytogenetic studies have revealed that chiasmata distributions are not uniformly distributed throughout the lengths of chromosomes, and can be influenced by chromosome structure and length (Kaback et al. 1992, Heyting 1996, Turney et al. 2004), and DNA composition within chromosomal regions (Jansen-Seaman et al. 2004), it may not be advisable to apply mapping function corrections to a new species in the absence of a more rigorous analysis of recombination fractions within the species. As an example, a comparison of the observed recombination fractions in female and male rainbow trout from two mapping panels that were used by Danzmann et al. (2005) reveal that the observed recombination ratios versus plotted map distances do not approximate either the Kosambi or Haldane mapping functions for recombination fractions less than 0.35, and for recombination fractions larger than this, the Kosambi function may be appropriate only for correcting female mapping distances. However, neither the Kosambi nor Haldane function appears appropriate for correcting male recombination fractions (Figures 10.4A and 10.4B). Observed male recombination fractions and map distances were more similar to one another across all the recombination intervals examined, than were those of the female distributions.

This dataset highlights the importance of constructing male and female-specific linkage maps in instances where large recombination differences exist between the sexes. Salmonid fish may represent an extreme example of this condition, because these fish have the largest reported sex-specific differences in recombination ratios for any known vertebrate (Sakamoto et al. 2000). Although most aquaculture species are unlikely to have such extreme differences, it may still be informative to plot sex-specific

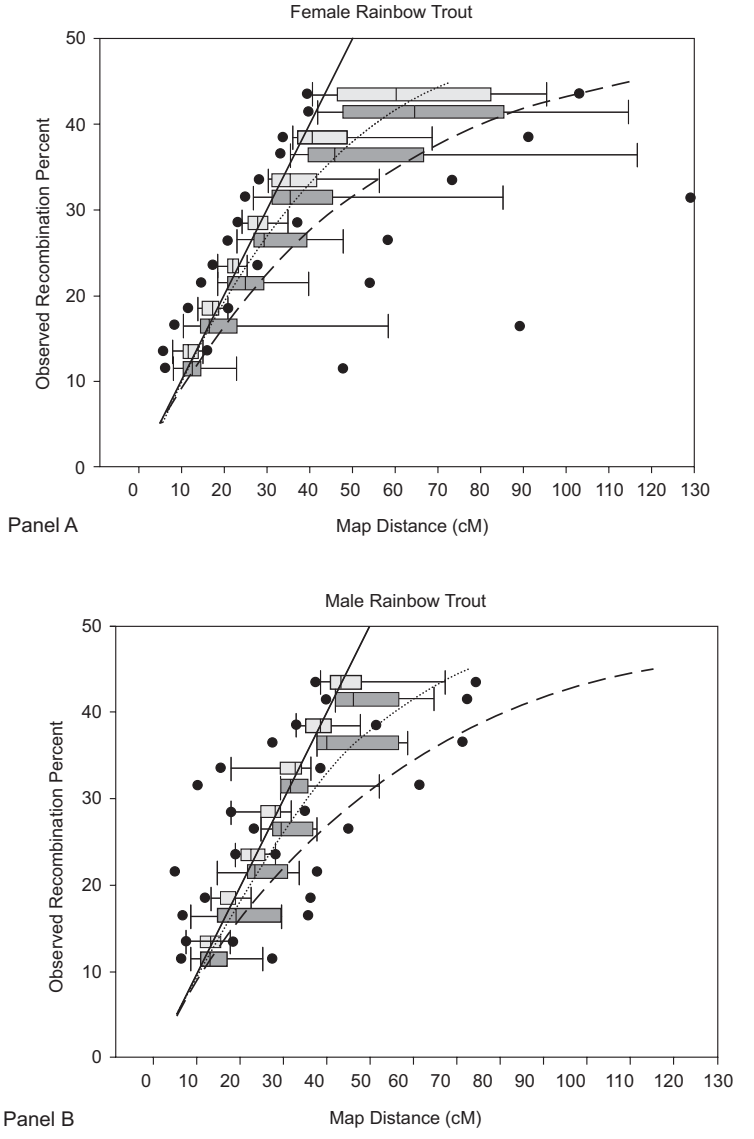


Figure 10.4. Distribution of observed recombination levels versus calculated map distances for pairs of linked genetic markers in female (panel A) and male (panel B) rainbow trout. Data are obtained from two mapping panels (Lot 25 = dark grey box plots and Lot 44 = light grey box plots). Data were analyzed for all markers falling into the following recombination intervals: 0.10–0.15; 0.15–0.20; 0.20–0.25; 0.25–0.30; 0.30–0.35; 0.35–0.40, and 0.40–0.45. The box plots are bounded by the lower and upper 25 percentiles of the distribution with the middle bar in the box plot showing the median of the distribution. The lower and upper dots represent the 5th and 95th percentile, respectively. The solid bar represents observed map distance = observed recombination distance. The fine-dotted line represents the Kosambi mapping function correction. The long-dashed line represents the Haldane mapping function correction. Map distances were obtained as raw Θ scores from phase maps (using MAPDIS-V[1]) with the option to ignore adjacent double crossovers in the phase vectors invoked. Mapping module in LINKMFEX software: see Table 10.1.

recombination levels in relation to their estimated map distances. Sex-specific differences in recombination rate (based upon differential chiasmata distributions between the sexes) have been reported among mammalian and amphibian species, which is in contrast to the findings of general homogeneity for sex-specific chiasmata distributions in avian species (Wallace et al. 1997, Pigozzi and Solari 1999). Without the construction of sex-specific linkage maps and the estimation of sex-specific recombination differences in any new species being studied for aquaculture production, the assignment of an accurate mapping function correction remains enigmatic.

The empirical example outlined for the rainbow trout data set of Danzmann and others (2005) indicates that the two most commonly used mapping function adjustments (i.e., Kosambi and Haldane) may not be appropriate for all investigated species. Although the Kosambi function may be appropriate for species that appear to express some degree of interference in their crossover structure (versus lack of interference for the Haldane function), it is apparent, that even the moderately adjusted map distance obtained with the Kosambi function may overestimate the majority of map distance placements. Where the empirical data suggest that interference levels may be moderate to high, mapping functions with a more conservative reestimation of the actual observed recombination fraction (i.e., a modified binomial) could be applied. The modified binomial may be expressed as:

$$\mathrm{md} = \theta - [\tfrac{1}{2}[1 - (1 - 2\theta/x)^{x}]] + \theta, \tag{10.6}$$

where x = the expected number of crossovers in the interval. See Figure 10.3. This mapping function gives slightly lower recombination estimates of the mapping distance compared to the binomial correction, and much lower levels than those expected with the Kosambi function, yet still provides a slightly higher estimate than observed θ values for $r \geq 0.20$.

When making adjustments for observed recombination levels, it is also important to remember that many of the potential false assignments to a given linkage group can be prevented by arbitrarily choosing a stringent LOD level for acceptance of linkage. If LOD level thresholds are set at a sufficiently high enough level then false inclusions will largely be prevented. Linkage thresholds set at LOD = 3.0 have largely been used for constructing linkage maps (log of odds probability = 0.001 for false inclusion) (Ott 1999). However, even this apparently stringent threshold may result in an increase in Type I error when constructing maps and adjusting map distances. For example, for the data shown in Figure 10.4, marker pairs with pair-wise recombination intervals greater than 0.20 and 0.30 would not be considered to be linked in the Map25, and Map44 mapping panels, respectively, at a LOD = 4.0 threshold. However, at a LOD = 3.0 linkage acceptance threshold, marker pairs are accepted as being linked within the next highest recombination interval in Map25 (= 0.20 − 0.25), and Map44 (= 0.30 − 0.35), thus increasing the likelihood that the observed recombination value may be underestimated. This illustrates the importance of setting a LOD threshold value that is stringent enough to minimize the potential of including loosely linked markers (i.e., $\theta \geq 0.25$) for moderately sized mapping panels.

The above example highlights that if a large enough LOD threshold is set prior to constructing a linkage map, then many potentially false linkage pair assignments can be avoided, and that linkage interval assignments falling within a recombination zone requiring extensive map adjustment (i.e., > 0.30) will rarely occur. Furthermore, as marker densities are increased during the progression of a linkage mapping study, the

probabilities are increasingly diminished that new markers will fall into a 'grey zone' of recombination that may require some mapping distance adjustment. Most newly added markers will have recombination values whereby observed θ values are the best estimate of the recombination distance. However, it is recognized that during initial starting map construction, marker densities will be sparse and it may be appropriate to apply a mapping function correction for adjustments of terminal markers added to a linkage group, or long interval adjustments within the central portions of a linkage group. We would, however, strongly recommend that these types of adjustments be avoided by initially setting a high and stringent LOD level for inclusion of newly added markers. This may initially produce a fragmented map with many small, unlinked segments, but this is more desirable than having spurious linkage arrangements generated that can result when markers with missing genotypes are added to the linkage complement.

Linkage Estimation and Map Order Determination

The first step in establishing a linkage map is determining whether any pair of genetic markers that have been genotyped are linked to one another. As mentioned, linkage is characterized as the nonrandom joint segregation of marker alleles following meiosis. Genetic markers that are physically linked along a chromosome will not segregate independently from one another but will tend to maintain their parental allelic phases without disruption throughout the process of meiotic recombination/reduction (i.e., meiosis I) and separation of chromatids (meiosis II). As outlined above, if markers are very close to one another on a chromosome, then only two sets of genotypes will be observed in the progeny (i.e., alternate parental phases). We would term such a phenomenon as complete linkage characterized by the complete absence of recombinants.

To determine whether marker pairs are linked to one another, it is necessary to calculate the recombination fractions and associated two-point LOD scores among all pair-wise marker combinations (see description in the previous section). This will result in the estimation of: $n \bullet (n - 1)/2$ LOD scores, where n = the number of markers genotyped in a given mapping parent. Since linkage maps are the representation of meiotic events that have occurred in a single individual, the most accurate estimation of recombination fractions are obtained by tracking the findings from meiotic events in single parents. Although methods are available for pooling data across multiple parents, most important aquaculture species are extremely fecund. This facilitates the production of large family sizes that are readily amenable to the methods of direct linkage analysis. The discussion that follows assumes that linkage maps are being constructed within a single sibship.

After two point LOD scores have been determined, it is possible to begin the linear ordering of the markers by first determining the total number of markers that fall within a linkage group (i.e., cluster of linked markers). Several software packages are available that will implement such clustering searches. See the last section in the chapter. Criteria for the inclusion of markers within a linkage group should be done using a fairly stringent LOD threshold. Previous research has suggested an LOD threshold of 3.0 for inclusion (Ott 1999), but we would recommend that a higher threshold be chosen. See discussion in the previous section. Initial ordering, when first starting a mapping study, could be accomplished using an LOD = 3.0 threshold to establish a 'template' order when marker densities are low. However, as additional markers are

added, researchers should establish a more stringent LOD threshold to minimize the chances of Type I error assignments to linkage groups (i.e., chance inclusion of markers to a linkage group due to the increased number of pair-wise recombination tests performed). After the linkage group clusters have been established, it is possible to begin map ordering of recombinant markers. Nonrecombinant markers represent a single locus and should therefore be collapsed into a single composite marker prior to ordering. The first step in the process is to establish the nearest neighbor marker positions of all markers within the linkage group. For any source marker, the most likely nearest neighbor to that marker will have the lowest recombination value and highest two-point LOD score value. Although it may appear *a priori* that the lowest recombination value alone should dictate evidence for a linkage affiliation, it must be remembered that recombination estimates are only approximate estimators for linkage associations given the fact that similar recombination values may be found among marker pairs that vary in their degree of missing genotypes. Low recombination values generated when extensive genotypic information is missing for a given marker may generate spurious linkage associations. Therefore, those marker pairs that have the lowest associated recombination value and highest associated LOD score should be regarded as the most likely nearest neighbor combinations when determining initial map orders. In instances where nearest neighbor markers differ by these criteria (i.e., one marker appears to possess the lowest recombination distance, while another marker has a higher LOD score for assignment), preference for nearest neighbor map placement should be given to the marker with the higher LOD score association.

The establishment of proper marker ordering is a nontrivial exercise given the sobering fact that it is possible to generate $(n + x)!/2$, linkage map orders, where n is the total number of unique marker positions genotyped, and x = the total number of zero recombination clusters detected in the linkage group. Thus, if 60 markers were assigned to a particular linkage group, and 12 zero recombination clusters were detected in the linkage group represented by 37 markers, then $(23 + 12)!/2 = 5.16 \times 10^{39}$ map orders would be possible. Marker ordering begins by arbitrarily selecting a source marker and determining the nearest neighbor markers to this marker. Marker clusters with zero recombination from one another are of course excluded from the analysis with the caveat that any single marker chosen to represent this cluster has the most complete genotypic information available.

For any point source marker, the first two nearest neighbors to this marker will have three possible true map orders. For example, if the initial source marker is called '3,' and two new additional markers (i.e., '2' and '4') were identified as being the nearest neighbors to this marker, then the three possible map orders for these markers are 2-3-4, 3-2-4, and 3-4-2. Note that the ordering 4-2-3 and 2-4-3 are identical to the latter two map orders, and simply indicate that marker 3 would either be an upstream marker or downstream marker in the linkage group.

To determine which of these multipoint orderings are correct, the recombination intervals (Θ) between the markers would be compared. If marker 3 is internal within the multipoint cluster then the following relationships will generally hold:

$$(\Theta_{23} + \Theta_{34}) \geq \Theta_{24}$$

$$\Theta_{23} < \Theta_{24}$$

$$\Theta_{34} < \Theta_{24} \tag{10.7}$$

See Figure 10.5.

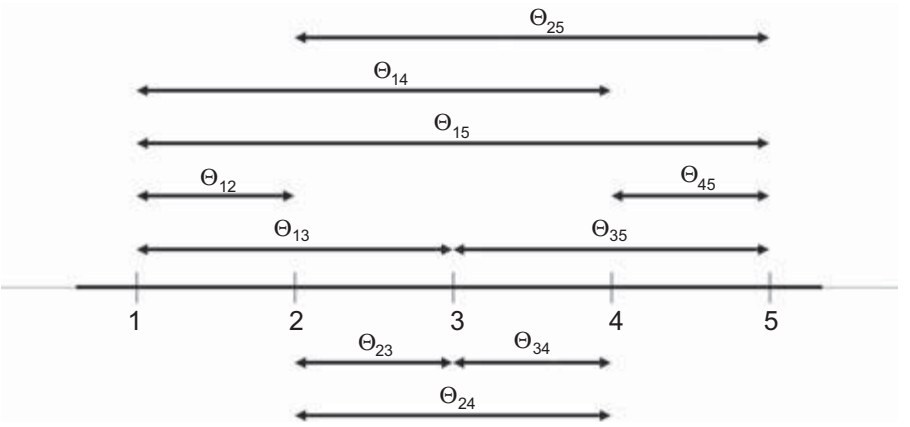


Figure 10.5. Crossover intervals that are involved in multipoint ordering of up to two adjacent marker positions in a linear genetic map.

If one of the following conditions is observed, then support would be obtained that marker 3 is a terminal marker in the multipoint cluster, and that the next nearest neighbor should be added to assess the flanking position.

$$(\Theta_{23} + \Theta_{24}) \geq \Theta_{34} \qquad (\Theta_{24} + \Theta_{34}) \geq \Theta_{23}$$
$$\Theta_{23} < \Theta_{34} \qquad \Theta_{24} < \Theta_{23}$$
$$\Theta_{24} < \Theta_{34} \qquad \Theta_{34} < \Theta_{23} \qquad (10.8)$$

The observation of either $\Theta_{24} < \Theta_{34}$, or $\Theta_{24} < \Theta_{23}$ holding true would suggest that marker 3 is a terminal marker in the cluster.

After the flanking markers in a 3-point cluster are determined, the flanking markers themselves are chosen as the source marker and the procedure described above is repeated. If the order 2-3-4 was ascertained to be the most likely ordering for the multipoint set, and the next nearest neighbor marker placed with marker 2 was determined to be marker 1, then by extension, the same interval tests could be applied to the 1-2-3 cluster ordering (see Figure 10.5), and extended to include marker 4 so that the following recombination criteria would hold.

$$\Theta_{12} + \Theta_{23} + \Theta_{34} \geq \Theta_{14}$$
$$\Theta_{12} + \Theta_{23} \geq \Theta_{13} \qquad \Theta_{23} + \Theta_{34} \geq \Theta_{24}$$
$$\Theta_{13} \leq \Theta_{14} \qquad \Theta_{24} \leq \Theta_{14}, \text{ if } \Theta_{13} \text{ or } \Theta_{24} \neq \text{ or } > 0.50$$
$$\Theta_{12} \leq \Theta_{13} \qquad \Theta_{23} \leq \Theta_{13} \qquad (10.9)$$

Similarly, for the addition of the next nearest marker proximal to marker 4, for example, marker 5, the following recombination distributions would hold:

$$\Theta_{23} + \Theta_{34} + \Theta_{45} \geq \Theta_{25}$$

$$\Theta_{23} + \Theta_{34} \geq \Theta_{24} \qquad \Theta_{34} + \Theta_{45} \geq \Theta_{35}$$

$$\Theta_{24} \leq \Theta_{25} \qquad \Theta_{35} \leq \Theta_{25}, \text{ if } \Theta_{24} \text{ or } \Theta_{35} \neq \text{ or } > 0.50$$

$$\Theta_{23} \leq \Theta_{24} \qquad \Theta_{34} \leq \Theta_{24} \tag{10.10}$$

This procedure is repeated iteratively (i.e., evaluating the accuracy of all four marker intervals) until a complete map order is obtained. By starting with each marker in the linkage group, the likelihood for any given map order can be obtained as: t/N, where t is the number of orders of a specific type obtained over all possible N marker initiated maps.

For any given three marker multipoint test, two terminal markers and one central marker will always be identified. If it is determined for all subsequent marker additions, that the recombination interval for one of the terminal markers is always greater than the entire initial multipoint interval, and if the most likely placement of the nearest neighbor converges on the same marker for both flanking markers, then it is most likely that this marker represents a terminal marker in the assembly. As an example, let us use initial multipoint ordering 2-3-4. If it is found that marker 1 appears to be the nearest neighbor to marker 2 and also to marker 4, and $\Theta_{14} \geq (\Theta_{34} + \Theta_{23})$, and $\Theta_{12} < \Theta_{14}$, then it is most likely that marker 4 represents a terminal position in the linkage group.

Difficulties arise in map ordering when markers with missing genotypic information are included. This situation will apply to most mapping panel datasets. When two markers are compared to one another, and both markers have missing genotypes then their estimated recombination interval may in fact be either somewhat larger or smaller than their true recombination fraction due to chance. Thus, when multipoint ordering of such markers is attempted, ambiguous orders may result dependent upon the sequence of marker additions. To minimize the errors that are inherent when using markers with missing genotypic information, it is recommended that a framework map first be constructed. Such a map would use only those markers with complete genotypic information to establish an initial map order. Following this, markers would be added to the framework map in order of decreasing information content. In other words, the marker(s) that would be added in the next step of map order construction to a framework map would be markers, or the marker, missing the least amount of genotypic information. The last markers, or marker, to be added to the map would be marker(s) with the least amount of genotypic information. For an additional discussion of mapping order methods, see Speed and Zhao (1993).

Crossover Interference

Crossover interference is the phenomenon whereby a single crossover event during meiosis tends to diminish or restrict the probability that an adjacent crossover event

will occur within the same region. Complete positive interference indicates that at most there will be a single crossover event per chromosome arm. Hence, for a metacentric linkage group this would allow two crossover events (one on each arm of the metacentric) to exist within the linkage group. Methods such as gene-centromere mapping have been used to infer the relative lengths and orientations of gene markers along a chromosome arm (Danzmann and Ghabi 2001), and have relied on the phenomenon of chromatid interference to derive the relative order of markers along a chromosome. Basically, the methodology consists of scoring homozygous and heterozygous genotypes in the progeny of a mapping panel following the induction of techniques that suppresses meiosis II (e.g., gynogenesis). The higher the proportion of heterozygous progeny, the higher the probability is that a crossing-over event has occurred between the genetic locus and the centromere. Thus, the observed proportion of the heterozygotes scored in the mapping panel is taken as a measure of the genetic distance between the centromere and marker as a function of $y/2$, where y indicates the proportion of heterozygotes scored. See Danzmann and Ghabi (2001) for a more detailed description of the methodology.

Interpretations of gene-centromere mapping data, however, rely on the premise that crossover events will be randomly distributed throughout the length of a chromosome so that it is highly unlikely to observe the complete transmission of heterozygous progeny following gene-centromere mapping methods unless a single crossover event has occurred between the centromere and the marker in question. If chiasmata junctions are, however, constrained within restricted chromosomal regions (Jansen-Seaman et al. 2004) then more distal markers from the centromere within a linkage group may actually be characterized by mixtures of homozygous and heterozygous genotypes, if a second crossover junction point occurs along a chromosome that is distal to the first and may involve crossing-over among any of the four chromatid arms involved in meiosis I crossovers. For example, Danzmann and Gharbi (2001) (see their Figure 2) outlined how a double crossover event involving only three of the four chromatid strands (with one strand common to both crossover events) would result in the observation of all heterozygous progeny in the terminal telomeric regions of the chromosome. Conversely, double crossover events involving the same two chromatid arms in both crossover events would result in the production of only homozygous progeny for the telomeric marker position. See Figure 10.6. Thus, more distal markers along a chromosome arm may in fact appear more proximal to the centromere than they truly are, using gene-centromere mapping methods.

The most reliable method for inferring phase changes in the linear order of marker genotypes is to construct a linkage phase map for a given chromosome gene map order, then directly examine the incidence of crossover points along the length of the chromosome. For example, in Figure 10.1, mapping progeny 10, 11, and 37 show three different scenarios with respect to phase changes along the chromosome. Progeny 37 is characterized as a complete parental phase genotype, as there are no changes in genotype phase for all the markers in the linkage group. Individual 10 is observed to possess a single crossover phase, while individual 11 possesses a double crossover phase. When interpreting phase maps it is important to consider the position of the centromere in relation to the phase map.

Maps characterized by a high incidence of double crossover events would be expected with metacentric chromosomes as each arm of a metacentric chromosome is expected to exhibit at least one crossover event during meiosis with complete

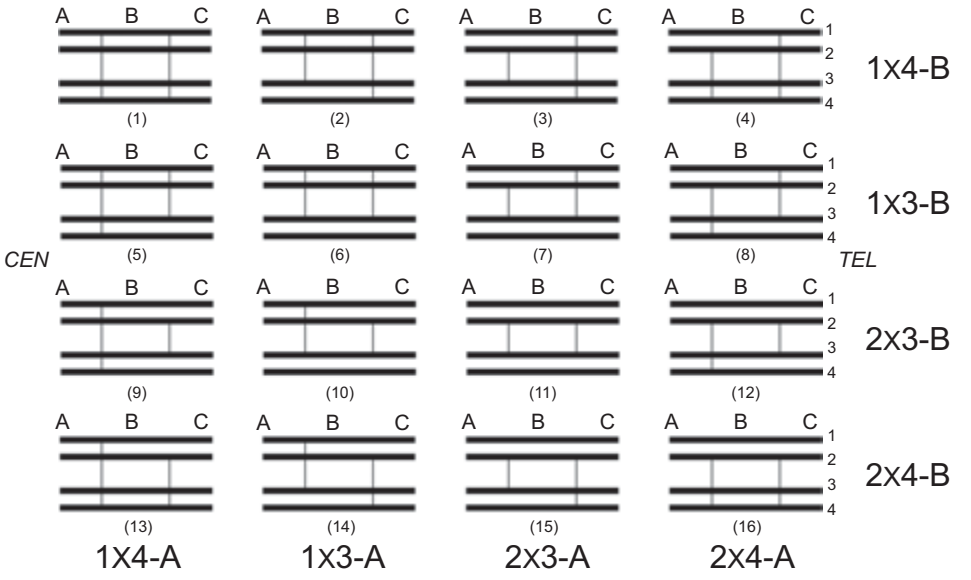


Figure 10.6. Distribution of double crossover events that may influence the resultant genotypes of markers terminally located on a linkage group. All 16 possible crossover types among the four chromatids from a homologous pair of chromosomes that pair at meiosis I are shown. Crossovers in regions proximal to the centromere are depicted as crossover type A, while those closer to the telomere are depicted as crossover type B. For a given marker 'C' in the telomeric region of the chromosome, certain chromatid crossover events may produce a recombinant genotype at marker 'C,' despite undergoing two intercalary crossover events (i.e., all of those combinations with only one strand in common at each crossover point = configurations 2, 4, 5, 7, 10, 12, 13, and 15) (adapted from Danzmann and Gharbi 2001).

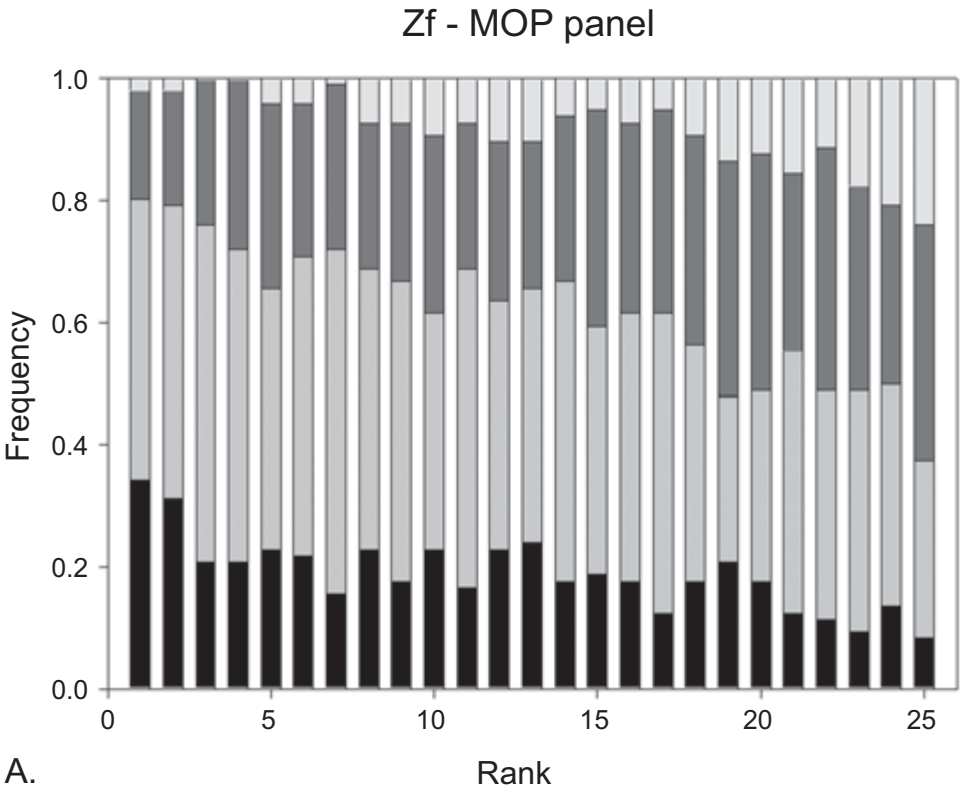
interference. Conversely, the incidence of double crossover events may not be expected to be as high for single-armed or acrocentric type chromosomes. Empirical evidence suggests that more crossover events are restricted toward the telomeric ends of metacentric chromosomes and that acrocentric type chromosomes may have higher overall levels of crossovers per unit physical length of chromosome and may have more proximally distributed (i.e., closer to the centromere) crossover events (Dumas and Britton-Davidian 2002, Froenicke et al. 2002, Kauppi et al. 2004), although species-specific variation in these rates may exist (Jansen-Seaman et al. 2004). Phase established genetic linkage maps have revealed that often more than one crossover event can occur per chromosome arm, and that the best indicator for the likelihood of observing multiple crossover junctions is individual arm length. Genetic map distances are roughly proportional to the physical size of a chromosome. For example, if the ranking in linkage group size of the 25 zebrafish (*Danio rerio*) chromosomes (size in centiMorgans [cM] of recombination distance) is made against the actual physical size of the linkage group in Mb of DNA sequence (data from http://www.ensembl.org for Zv4 physical map data and http://zfin.org for the MOP mapping panel genetic map data), there is a highly significant association between genetic map size and physical chromosome size ($F_{[1,24]} = 10.508$; $P = 0.0036$).

Larger-sized chromosomes, both in terms of genetic map distance and physical base pair length, have a higher frequency of double, and higher orders of crossovers (i.e., > 2 x) events throughout their length. See Figure 10.7. Examination of the data portrayed in Figure 10.7A reveals that the incidence of multiple crossover events can in fact be quite high. For example, in the two largest zebrafish chromosomes according to map distance size (i.e., Dr7 = rank 25, and Dr14 = rank 24), the most conservative observation of crossover frequencies (i.e., excluding adjacent double crossover genotypes) indicates that multiple crossover events are present in greater than 20% of the progeny genotyped in the MOP mapping panel. Both of these linkage groups are now also estimated to be the largest chromosomes in the zebrafish genome according to Mb of DNA assigned to the chromosome (physical map data from Zv6, May 2006 release, www.ensembl.org).

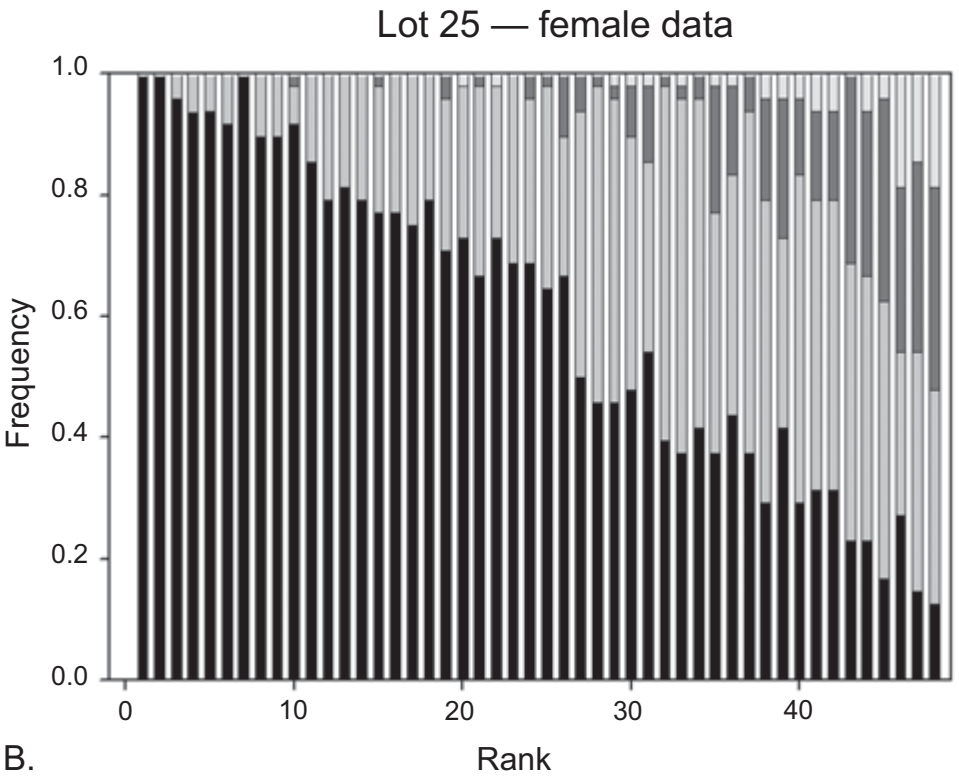
The above discussion suggests that even during the early stages of genetic map construction it may be possible to glean insights into the physical structure of chromosomes upon which a linkage map is being constructed. Those linkage groups showing a higher incidence of double crossover or multiple crossover events could be regarded as representative of the largest chromosomes in the genome. However, it is important ultimately, to relate the actual ordering of the markers in different linkage groups to the physical chromosomes themselves using procedures such as fluorescent *in situ* hybridization (FISH). FISH data serve to anchor the actual location of a given genetic marker to its physical location on a chromosome. For example, recent FISH data relating the genetic maps in rainbow trout to their chromosomal locations (Phillips et al. 2006) has revealed that the fifth largest chromosome in the genome of this species corresponds to a linkage group (Om8) that has been observed to possess an extremely low level of recombination (Danzmann et al. 2005). For example, there are three separate lot 25 female LOD = 4.0 clusters within this linkage group, that are ranked at positions 1, 8, and 15 (Figure 10.7B), and at positions 4 and 24 for the two separate LOD = 4.0 clusters detected in the lot 44 mapping panel (Figure 10.7C). Thus, different chromosomes within the genome of a species may be differentially regulated by underlying biological factors that alter their intrinsic rates of recombination and alter the relationship between chromosome size and recombination rate.

Segregation Distortion, Missing Data, and Scoring Errors

The most serious limitations to the construction of genetic linkage maps arise from the inclusion of falsely genotyped, or incompletely genotyped progeny data. Obviously the former is of far greater concern, but false assignments of markers to a given linkage group because of incomplete genotyping information is also a problem that a researcher must deal with when conducting a linkage analysis study. One of the ways that genotyping, or genotype misinterpretations can be recognized, is if a particular genetic marker is associated with strong deviations from expected 1:1 Mendelian segregation. Some linkage mapping software packages will conduct tests for conformity to Mendelian expectations (i.e., tests for segregation distortion). Any genetic markers that remain unassigned to a linkage group (especially in the later stages of genetic mapping studies) should be reverified by genotyping the marker again. This does not imply that all instances of segregation distortion detected in a mapping study



A.



B.

Figure 10.7. (*Continued*)

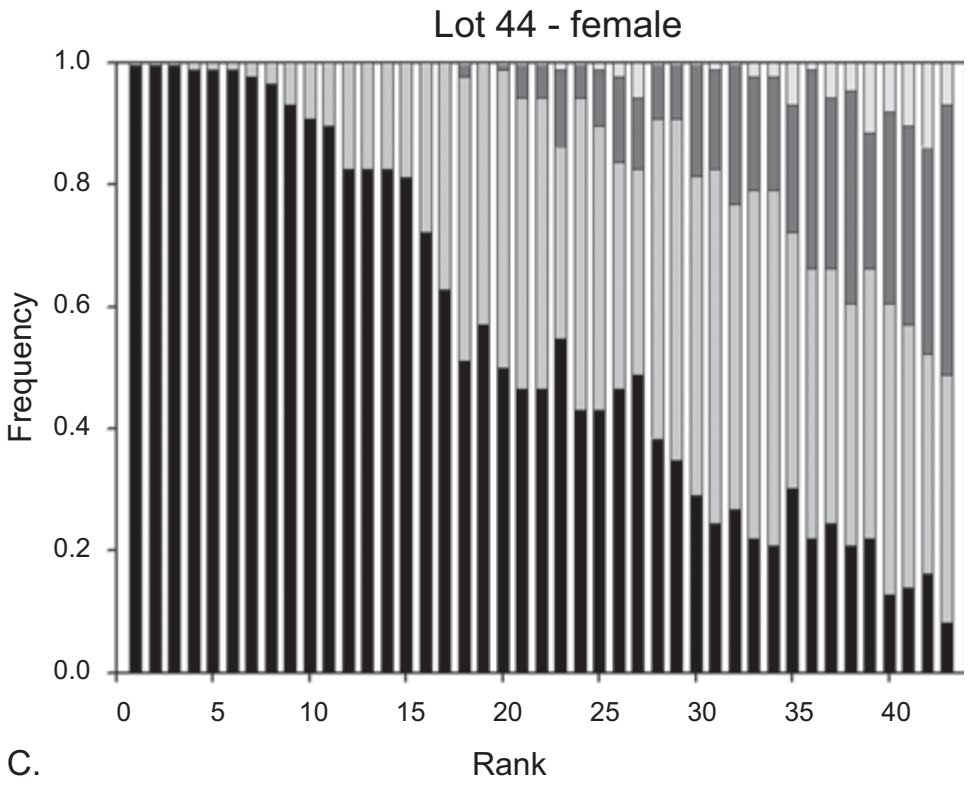


Figure 10.7. Empirical assessment of observed map distance sizes versus frequency of different chromosomal phases in zebrafish MOP mapping panel (A), and rainbow trout female maps for lot 25 (B) and lot 44 (C). Nonrecombinant parental phases are shown at the bottom of each stacked histogram (black bar) followed by the observed frequency of single crossover progeny (medium-grey bar), double crossover progeny (dark-gray bar), and multiple crossover progeny (> = 3) (light-gray bar). The ranking of genetic map sizes are shown in an ascending scale with rank 1 corresponding to the smallest genetic map. All map distances were estimated using the option to ignore adjacent double cross-over positions in the mapping data. Details on the data used to produce the figures can be found at www.uoguelph.ca/~rdanzman/appendices.

are due to false genotyping. In fact most instances of segregation distortion that are detected are likely to be real and can be verified as such by the colinkage of other markers showing varying levels of segregation distortion within a particular linkage group segment. Instances of real segregation distortion may arise if there are lethal or semilethal alleles segregating in a mapping parent under study, that are expressed within the combining genetic background of the alternate parent (i.e., epistatic or dominant/recessive expression states).

Missing genotypic information can also be a problem due to the fact that remaining allele phases can spuriously align with the incorrect linkage group by chance, if the remaining alleles happen to fall within a region of recombination (i.e., within their true linkage group), and this region is currently not genotyped for neighboring markers. This misalignment may also be significantly compounded if the poorly genotyped marker contains one or two genotyping errors. The best way to circumvent (or at least detect) these types of errors is to use multiple mapping panels for your genetic map construction. If, for example, a given marker assigns to a linkage group in one of your

mapping parents, while it is not assigned to this linkage group for all your other mapping parents, it is an indication that it has likely been improperly assigned due to missing and/or inaccurate genotypic information.

Pseudolinkage

This section is intended more as a general interest section for researchers working on species where pseudolinkage affinities have been detected (e.g., salmonid fishes), and thus, may not be of widespread interest. It is, however, included within this chapter on genetic mapping methods applied to aquaculture species, since researchers are currently only beginning linkage mapping studies in a few species. Given the fact that pseudolinkage may be more widespread than currently recognized, we have decided to include a brief overview on the topic.

Pseudolinkage is perhaps one of the most inappropriately labelled terms in the realm of transmission genetics given the fact that it is meant to convey the notion that the linkage arrangements associated with the markers in the linkage group are false. Although it is true that there will be two separate physical linkage clusters contained within each single pseudolinkage grouping, and given the fact the joining of two separate physical linkage groups is a false linkage, it should also be recognized that their association is mediated through a very real biological process brought about by the dynamics of meiotic transmission. Pseudolinkage clusters are often characterized by the end-to-end or telomeric linkages of two separate linkage groups that can appear genetically linked if the only criterion used to construct the linkage arrangement is an LOD specified threshold. This threshold is based upon the maximum recombination distance between a pair of markers and the sample size that was used to obtain the estimate.

Pseudolinkage can be detected by ignoring the true parental linkage phases in the cross. Linkage, as previously outlined, is simply the magnitude of deviation away from an expected level of 50% recombination between a pair of markers. As the recombination level between a pair of markers decreases, the frequency of the parental phases will proportionately increase. Pseudolinkage is characterized by an unexpected excess of recombinant genotypes in comparison to parental phases in the progeny transmission vectors. Thus, as the level of recombinant genotypes increases from the expectation of 50%, the proportion of parental phases will consequently decrease. By ignoring the phase assignments to a given set of markers, and testing all pair-wise combinations of markers for their differential increase/decrease away from an expected 1:1:1:1 ratio in the progeny transmission vectors, it is possible to detect these arrangements. For example, if it is known that the parental linkage phases for a pair of markers with alleles A/a and B/b, are AB and ab, then genotypes Ab and aB would represent recombinant progeny phases. Classical linkage would be identified by the underabundance of these latter two genotypic combinations in the progeny transmission vectors. Pseudolinkage, however, would be characterized by the overabundance of these two genotypic classes in relation to the frequency of AB and ab genotypes.

Pseudolinkage occurs when the meiotic disjunction processes are regulated to prevent the random assortment of chromatids following crossing-over in meiosis I and meiosis II. This could occur if there is some form of meiotic drive that exists during meiosis that can preferentially result in the nonrandom distribution of certain

chromatid combinations in gametes. The most likely mechanism for this would be centromere-specific sequence motifs that may have fairly uniform kinetochore/meiotic spindle binding sites on the nuclear envelope for separations at meiosis I and telomere-specific binding sites that direct or influence separations at meiosis II. These physical binding sites may direct meiotic disjunctions by bringing centromeric and telomeric regions with higher similarity into one region. By the nature of meiotic reductional (meiosis I) and divisional (meiosis II) segregations, two such nodes are expected to exist within each cell for each division (Heyting 1996). Factors that could regulate the affinity of centromeric regions for either node may be their ancestry of transmission through either maternal or paternal somas, and thus chromatids may segregate along divisions of maternal-based or paternal-based centromere regions. There is supportive evidence that such processes occur in all-female based hybridogenetic vertebrate species (Bogart and Licht 1986, Mateos and Vrijenhoek 2002). During meiosis, it is known that all of the chromosomes of maternal origin are preferentially segregated to the surviving first polar body, while paternal chromosomes are sequestered to the disintegrating second polar body (Ragghianti et al. 1995). Other factors that could promote the meiotic occurrence of such mechanisms are recent hybridization events. Maternal and paternal chromosome sets will be more similar to one another due to their shared recent common ancestry in the maternal and paternal populations, and thus may demonstrate pairing affinities during meiosis in the hybrid offspring.

Telomeric nodes may be subject to more rearrangements, however, due to the fact that recombination events are expected to be higher in these regions of the chromosome. Therefore, the affinity of telomeric nodes for each other may be more variable within the genome of a species. This differential affinity of telomeric ends for each other is seen as a key element in the overall probability of detecting or forming pseudolinkage arrangements within species (Wright et al. 1983, Allendorf and Thorgaard 1984, Allendorf and Danzmann 1997).

Software Resources for the Construction of Linkage Maps

The construction of a linkage map is a repetitive and iterative task that is best achieved using dedicated software tools. A valuable resource for such software is the database recently established at Iowa State University (http://www.animalgenome.org/cgi-bin/util/sw_index), which provides summary information on more than 300 computer programs for genetic analysis in a convenient organized and searchable format. Some of the most popular packages with aquaculture geneticists are listed in Table 10.1 along with key features, such as compatible operating systems, supported mapping populations, and licence requirements. Each package typically provides standard tools to compute pair-wise LOD scores, identify linkage groups, estimate marker order, and calculate map distances. Some packages (e.g., LINKMFEX) also include a module to test for segregation distortion, which saves importing the data into a separate piece of software. Another useful feature is the ability to export map graphics for use in other applications (e.g., JOINMAP, MAPMAKER), although this task is generally made easier using software specifically developed for this purpose (e.g., MAPCHART, Voorrips 2002). In practice, the choice of a particular linkage program will be eventually dictated by technical constraints and personal preferences. Each package has its own strengths

Table 10.1. Overview of popular packages for the construction of linkage maps in aquaculture species.

	Platform	Licence	Interface	Populations[1]	Reference
CARTHAGENE	PC, UNIX	freeware	command line, graphic	F2, backcross, RIL, outcross	de Givry et al. 2005[2]
CRIMAP	PC, UNIX	freeware	command line	pedigree	Green et al. 1990[3]
JOINMAP	PC	commercial	graphic	F2, backcross, RIL, DH, outcross	Stam 1993[4]
LINKMFEX	PC	freeware	graphic	outcross	Danzmann and Gharbi 2001[5]
MAPMAKER	PC, MAC, UNIX	freeware	command line	F2, backcross, RIL, DH	Lander et al. 1987[6]
MAPMANAGER	PC, MAC	freeware	graphic	F2, backcross, RIL	Manly and Olson 1999[7]

[1]F2: F2 intercross; RIL: recombinant inbred lines; DH: double haploids
[2]http://www.inra.fr/Internet/Departements/MIA/T/CarthaGene/
[3]http://compgen.rutgers.edu/multimap/crimap/
[4]http://www.kyazma.nl/index.php/mc.JoinMap/
[5]http://www.uoguelph.ca/~rdanzman/software/LINKMFEX/
[6]ftp://ftp-genome.wi.mit.edu/distribution/software/mapmaker3/
[7]http://www.mapmanager.org/mmQT.html

and weaknesses, and we recommend that researchers evaluate several software packages with their own linkage data before choosing one that best meets their needs. However, it is important to acknowledge that, while pair-wise LOD scores and two-point recombination fractions should be identical across packages, substantial differences may be observed in marker ordering and map distance estimates depending on the algorithms implemented in the software.

Overview of Linkage Maps in Aquaculture Species

Genome mapping in aquaculture species has historically lagged behind progress in terrestrial livestock (Tong and Chu 2002). However, considerable advances have been made in the past several years, and the gap is closing at an increasing rate (Table 10.2). Moderate to high-density linkage maps have been developed for all five groups of aquaculture species identified as priority targets at the First Aquaculture Species Genome Mapping Workshop (Alcivar-Warren et al. 1997), including salmonids, tilapia, catfish, shrimp, and oyster. Since then, mapping efforts have been extended to several other species (e.g., European sea bass, Japanese flounder, scallop) and the flurry of mapping studies recently presented at the last International Symposium on Genetics in Aquaculture (Montpellier, June 2006) indicates that primary linkage maps will soon be available for most species of significant importance (e.g., sea bream, Atlantic halibut, flat oyster). Thus far, simple sequence repeats (SSR) and

Table 10.2. Current status of linkage maps in aquaculture species.

	Mapping panel(s)	Marker type(s)[1]	Markers[2]	LGs (n)[3]	Reference[4]
Finfish					
Salmonids					
Rainbow trout (*Oncorhynchus mykiss*)	double haploids	AFLP, SSR, genes	1359	40 (30)	Young et al. 1998 Nichols et al. 2003
	inter-strain backcross	SSR, AFLP, genes	1439	31 (29-32)	Sakamoto et al. 2000 Danzmann et al. 2005
Atlantic salmon (*Salmo salar*)	outcross	AFLP, SSR	527	31–33 (27-29)	Moen et al. 2004
	inter-strain backcross	SSR, genes	64	15 (27-29)	Gibley et al. 2004
Brown trout (*Salmo trutta*)	inter-strain backcross	SSR, genes	302	37 (40)	Gharbi et al. 2006
Arctic char (*Salvelinus alpinus*)	inter-strain backcross	SSR, AFLP, genes	327	46 (39-40)	Woram et al. 2004
Tilapia					
Nile tilapia (*Oreochromis niloticus*)	haploids	AFLP, SSR	174	30 (22)	Kocher et al. 1998
Nile tilapia × blue tilapia × Mozambique tilapia (*O. niloticus* × *O. aureus* × *O. mossambicus*)	interspecific 3-way cross	AFLP, SSR	292	24 (22)	Agresti et al. 2000
Nile tilapia × blue tilapia (*O. niloticus* × *O. aureus*)	interspecific F2 intercross	SSR, genes	552	24 (22)	Lee et al. 2005

(*Continued*)

Table 10.2. (*Continued*)

	Mapping panel(s)	Marker type(s)[1]	Markers[2]	LGs (n)[3]	Reference[4]
Catfish					
Channel catfish (*Ictalurus punctatus*)	outcross	SSR, genes	293	32 (29)	Waldbieser et al. 2001
Channel catfish × blue catfish (*Ictalurus punctatus* × *I. furcatus*)	interspecific backcross	AFLP	506	44 (29)	Liu et al. 2003
Walking catfish (*Clarias macrocephalus*)	haploids	AFLP	146	31 (27)	Poompuang and Nakorn 2004
Other finfish					
Japanese flounder (*Paralichthys olivaceus*)	inter-strain hybrids	AFLP, SSR	463	25–27 (23)	Coimbra et al. 2003
European sea bass (*Dicentrarchus labrax*)	outcross	SSR, genes	174	25 (24)	Chistiakov et al. 2005
Ayu (*Plecoglossus altivelis*)	inter-population hybrids	AFLP, SSR	195	36 (28)	Watanabe et al. 2004
Common carp × Boshi carp (*Cyprinus carpio* × *C. pellegrini*)	interspecific haploids	SSR, genes, RAPD	272	50 (50)	Sun and Liang 2004
Yellowtails (*Seriola quinqueradiata, S. lalandi*)	interspecific hybrids	SSR	200	21–25 (24)	Ohara et al. 2005
Shrimp					
Black tiger shrimp (*Penaeus monodon*)	outcross	AFLP	673	20 (44)	Wilson et al. 2003

Kuruma prawn (*Penaeus japonicus*)	outcross	AFLP	246	44 (43)	Moore et al. 1999
	outcross	AFLP	401	31–43 (43)	Li et al. 2003
White shrimp (*Penaeus vannamei*)	outcross	AFLP	394	47–51 (44)	Pérez et al. 2004
Chinese shrimp (*Penaeus chinensis*)	inter-population hybrids	AFLP	231–241	35-36 (44)	Li et al. 2006
Molluscs					
Eastern oyster (*Crassostrea virginica*)	outcross	AFLP, SSR, genes	133–158	12 (10)	Yu and Guo 2003
Pacific oyster (*Crassostrea gigas*)	inter-line double hybrids	SSR	102	10 (10)	Huber and Hedgecock 2004
	inter-strain backcross	AFLP	349	10–11 (10)	Li and Guo 2004
Zhikong scallop (*Chlamys farreri*)	inter-population hybrids	AFLP	545	19–20 (19)	Li et al. 2005
	inter-population hybrids	AFLP	503	23–25 (19)	Wang et al. 2005
Pacific abalone (*Haliotis discus hannae*)	inter-population hybrids	AFLP, RAPD, SSR	384	19–22 (18)	Liu et al. 2006
Others					
Sea urchin (*Stongylocentrotus nudus* × *S. intermedius*)	interspecific hybrids	AFLP	324-339	23–24 (NA)	Zhou et al. 2006

[1]Marker types are listed in decreasing order of abundance in each map.
[2]Numbers in this column either show the total number of markers in the map or separate numbers for each sex-specific map.
[3]Numbers in this column denote the total number of linkage groups in each map or sex-specific maps, with numbers in brackets indicating the number of expected linkage groups based on the haploid number of chromosomes (n) in the species (NA: not available).
[4]Multiple references indicate successive map versions, with information in columns three to five relating to the latest reference.

amplified fragment length polymorphisms (AFLP) have been the markers of choice for map development, although the growing popularity and availability of single nucleotide polymorphisms (SNP) suggests that future mapping efforts will incorporate a larger proportion of these markers. See Liu and Cordes (2004) for a review.

References

Agresti JJ, S Seki, A Cnaani, S Poompuang, EM Hallerman, N Umiel, G Hulata, GAE Gall, and B May. 2000. Breeding new strains of tilapia: development of an artificial center of origin and linkage map based on AFLP and microsatellite loci. Aquaculture, 185, pp. 43–56.

Alcivar-Warren A, R Dunham, P Gaffney, T Kocher, and G Thorgaard. 1997. First Aquaculture Species Genome Mapping Workshop. Anim. Genet, 28, pp. 451–452.

Allendorf FW and RG Danzmann. 1997. Secondary tetrasomic segregation of MDH-B and preferential pairing of homeologues in rainbow trout. Genetics, 145, pp. 1083–1092.

Allendorf FW and GH Thorgaard. 1984. Tetraploidy and the evolution of salmonid fishes. In: BJ Turner, Ed. Evolutionary genetics of fishes. Plenum Press, New York, pp. 1–46.

Babiak I, S Dobosz, K Goryczko, H Kuzminski, P Brzuzan, and S Ciesielski. 2002. Androgenesis in rainbow trout using cryopreserved spermatozoa: the effect of processing and biological factors. Theriogenology, 57, pp. 1229–1249.

Bogart JP and LE Licht. 1986. Reproduction and the origin of polyploids in hybrid salamanders of the genus *Ambystoma*. Can. J. Genet. Cytol, 28, pp. 605–617.

Chistiakov DA, B Hellemans, CS Haley, AS Law, CS Tsigenopoulos, G Kotoulas, D Bertotto, A Libertini, and FA Volckaert. 2005. A microsatellite linkage map of the European sea bass *Dicentrarchus labrax* L. Genetics, 170, pp. 1821–1826.

Coimbra MRM, K Kobayashi, S Koretsugu, O Hasegawa, E Ohara, and A Ozaki. 2003. A genetic linkage map of the Japanese flounder, *Paralichtys olivaceus*. Aquaculture, 220, pp. 203–218.

Danzmann RG, M Cairney, WS Davidson, MM Ferguson, K Gharbi, R Guyomard, L-E Holm, E Leder, N Okamoto, A Ozaki, CE Rexroad III, T Sakamoto, JB Taggart, and RA Woram. 2005. A comparative analysis of the rainbow trout genome with 2 other species of fish (Arctic charr and Atlantic salmon) within the tetraploid derivative Salmonidae family (subfamily: Salmoninae). Genome, 48, pp. 1037–1051.

Danzmann RG and K Gharbi. 2001. Gene mapping in fishes: a means to an end. Genetica, 111, pp. 3–23.

de Givry S, M Bouchez, P Chabrier, D Milan, and T Schiex. 2005. CARTHAGENE: multipopulation integrated genetic and radiation hybrid mapping. Bioinformatics, 21, pp. 1703–1704.

Dumas D and J Britton-Davidian. 2002. Chromosomal rearrangements and evolution of recombination: comparison of chiasma distribution patterns in standard and Robertsonian populations of the house mouse. Genetics, 162, pp. 1355–1366.

Froenicke L, LK Anderson, J Wienberg, and T Ashley. 2002. Male mouse recombination maps for each autosome identified by chromosome painting. Am. J. Hum. Genet, 71, pp. 1353–1368.

Gharbi K, A Gautier, RG Danzmann, S Gharbi, T Sakamoto, B Hoyheim, JB Taggart, M Cairney, R Powell, F Kreig, N Okamoto, MM Ferguson, L-E Holm, and R Guyomard. 2006. A linkage map for brown trout (*Salmo trutta*): Chromosome homeologies and comparative genome organization with other salmonid fish. Genetics, 172, pp. 405–2419.

Gilbey J, E Verspoor, A McLay, and D Houlihan. 2004. A microsatellite linkage map for Atlantic salmon (*Salmo salar*). Anim. Genet, 35, pp. 98–105.

Green PK, K Fall, and S Crooks. 1990. Documentation for *CRIMAP*, version 2.4.

Heyting C. 1996. Synaptonemal complexes: structure and function. Curr. Opinion Cell Biol, 8, pp. 389–396.

Hubert S and D Hedgecock. 2004. Linkage maps of microsatellite DNA markers for the Pacific oyster *Crassostrea gigas*. Genetics, 168, pp. 351–362.

Jansen-Seaman MI, TS Furey, BA Payseur, L Yontao, KM Roskin, C-F Chen, MA Thomas, D Haussler, and HJ Jacob. 2004. Comparative recombination rates in the rat, mouse, and human genomes. Genome Res, 14, pp. 528–538.

Kaback DB, V Guacci, D Barber, and JW Mahon. 1992. Chromosome size-dependent control of meiotic recombination. Science, 256(5054), pp. 228–232.

Kauppi L, AJ Jeffreys, and S Keeney. 2004. Where the cross-overs are: recombination distributions in mammals. Nat. Rev. Genetics, 5, pp. 413–424.

Kocher TD, WJ Lee, H Sobolewska, D Penman, and B McAndrew. 1998. A genetic linkage map of a cichlid fish, the tilapia (*Oreochromis niloticus*). Genetics, 148, pp. 1225–1232.

Lander ES, P Green, J Abrahamson, A Barlow, MJ Daly, SE Lincoln, and L Newburg. 1987. MAPMAKER: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics, 1, pp. 174–181.

Li L and XM Guo. 2004. AFLP-based genetic linkage maps of the Pacific oyster *Crassostrea gigas* Thunberg. Mar. Biotechnol, 6, pp. 26–36.

Li L, JH Xiang, X Liu, Y Zhang, B Dong, XJ Zhang. 2005. Construction of AFLP-based genetic linkage map for Zhikong scallop, *Chlamys farreri* Jones et Preston and mapping of sex-linked markers. Aquaculture, 245, pp. 63–73.

Li YT, K Byrne, E Miggiano, V Whan, S Moore, S Keys, P Crocos, N Preston, and S Lehnert. 2003. Genetic mapping of the kuruma prawn *Penaeus japonicus* using AFLP markers. Aquaculture, 219, pp. 143–156.

Li Z, J Li, Q Wang, Y He, and P Liu. 2006. AFLP-based genetic linkage map of marine shrimp *Penaeus* (*Fenneropenaeus*) *chinensis*. Aquaculture, 261, pp. 463–472.

Liberman U and S Karlin. 1984. Theoretical models of genetic map functions. Theor. Pop. Biol. 25, pp. 331–346.

Liu X, X Guo, Q Gao, H Zhao, and G Zhang. 2006. A preliminary genetic linkage map of the Pacific abalone *Haliotis discus hannae* Ino. Mar. Biotechnol., 8, pp. 386–397.

Liu Z, A Karsi, P Li, D Cao, and R Dunham. 2003. An AFLP-based genetic linkage map of channel catfish (*Ictalurus punctatus*) constructed by using an interspecific hybrid resource family. Genetics, 165, pp. 687–694.

Liu ZJ and JF Cordes. 2004. DNA marker technologies and their applications in aquaculture genetics. Aquaculture, 238, pp. 1–37.

Manly KF and JM Olson. 1999. Overview of QTL mapping software and introduction to Map Manager QT. Mamm. Genome, 10, pp. 327–334.

Mateos M and RC Vrijenhoek. 2002. Ancient versus reticulate origin of a hemiclonal lineage. Evolution, 56, pp. 985–992.

May B, JE Wright, and M Stoneking. 1980. Joint segregation of biochemical loci in Salmonidae: Results from experiments with *Salvelinus* and review of the literature on other fish species. J. Fish Res. Board Can., 36, pp. 1114–1128.

Moen T, B Hoyheim, H Munck, and L Gomez-Raya. 2004. A linkage map of Atlantic salmon (*Salmo salar*) reveals an uncommonly large difference in recombination rate between the sexes. Anim. Genet., 35, pp. 81–92.

Moore SS, V Whan, GP Davis, K Byrne, DJS Hetzel, and N Preston. 1999. The development and application of genetic markers for the Kuruma prawn *Penaeus japonicus*. Aquaculture, 173, pp. 19–32.

Nichols KM, WP Young, RG Danzmann, BD Robison, C Rexroad, M Noakes, RB Phillips, P Bentzen, I Spies, K Knudsen, FW Allendorf, BM Cunningham, J Brunelli, H Zhang, S Ristow, R Drew, KH Brown, PA Wheeler, and GH Thorgaard. 2003. A consolidated linkage map for rainbow trout (*Oncorhynchus mykiss*). Anim. Genet., 34, pp. 102–115.

Ohara E, T Nishimura, Y Nagakura, T Sakamoto, K Mushiake, and N Okamoto. 2005. Genetic linkage maps of two yellowtails (*Seriola quinqueradiata* and *Seriola lalandi*). Aquaculture, 244, pp. 41–48.

Ott J. 1999. Analysis of human genetic linkage. The Johns Hopkins University Press, Baltimore, MD.

Pandian TJ and S Kirankumar. 2003. Androgenesis and conservation of fishes. Curr. Science, 85, pp. 917–931.

Perez F, C Erazo, M Zhinaula, F Volckaert, and J Calderon. 2004. A sex-specific linkage map of the white shrimp *Penaeus* (*Litopenaeus*) *vannamei* based on AFLP markers. Aquaculture, 242, pp. 105–118.

Phillips RB, KM Nichols, JJ DeKoning, MR Morasch, KA Keatley, C Rexroad III, SA Gahr, RG Danzmann, RE Drew, and GH Thorgaard. 2006. Assignment of rainbow trout linkage groups to specific chromosomes. Genetics, 174, pp. 1661–1670.

Pigozzi MI and AJ Solari. 1999. Equal frequencies of recombination nodules in both sexes of the pigeon suggest a basic difference with eutherian mammals. Genome, 42, pp. 315–321.

Poompuang S and U Na-Nakorn. 2004. A preliminary genetic map of walking catfish (*Clarias macrocephalus*). Aquaculture, 232, pp. 195–203.

Qumsiyeh MB. 1994. Evolution of number and morphology of mammalian chromosomes. J. Hered., 85, pp. 455–465.

Ragghianti M, F Guerrini, S Bucci, G Mancino, H Hotz, T Uzzell, and GD Guex. 1995. Molecular characterization of a centromeric satellite DNA in the hemiclonal hybrid frog *Rana esculenta* and its parental species. Chromosome Res., 3, pp. 497–506.

Sakamoto T, RG Danzmann, K Gharbi, P Howard, A Ozaki, SK Khoo, RA Woram, N kamoto, MM Ferguson, L-E Holm, R Guyomard, and B Hoyheim. 2000. A microsatellite linkage map of rainbow trout (*Oncorhynchus mykiss*) characterized by large sex-specific differences in recombination rates. Genetics, 155, pp. 1331–1345.

Speed TP and H Zhao. 2003. Chromosome Maps. In: Balding DJ, Bishop M, Cannings C, Eds. Handbook of Statistical Genetics, 2nd ed. John Wiley & Sons Ltd. pp. 3–38.

Stam P. 1993. Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. Plant J., 3, pp. 739–744.

Sun XW and LQ Liang. 2004. A genetic linkage map of common carp (*Cyprinus carpio* L.). And mapping of a locus associated with cold tolerance. Aquaculture, 238, pp. 165–172.

Tong J and KH Chu. 2002. Genome mapping in aquatic animals: Progress and future perspectives. Russ. J. Genet., 38, pp. 612–621.

Turney D, T de los Santos, and NM Hollingsworth. 2004. Does chromosome size affect map distance and genetic interference in budding yeast? Genetics, 168, pp. 2421–2424.

Voorrips RE. 2002. MapChart: Software for the graphical presentation of linkage maps and QTLs. J. Hered., 93, pp. 77–78.

Waldbieser GC, BG Bosworth, DJ Nonneman, and WR Wolters. 2001. A microsatellite-based genetic linkage map for channel catfish, *Ictalurus punctatus*. Genetics, 158, pp. 727–734.

Wallace H, BMN Wallace, and GMI Badawy. 1997. Lampbrush chromosomes and chiasmata of sex-reversed crested newts. Chromosoma, 106, pp. 526–533.

Wang LL, LS Song, YQ Chang, W Xu, DJ Ni, and XM Guo. 2005. A preliminary genetic map of Zhikong scallop (*Chlamys farreri* Jones et Preston 1904). Aquaculture Res., 36, pp. 643–653.

Watanabe T, H Fujita, K Yamasaki, S Seki, and N Taniguchi. 2004. Preliminary study on linkage mapping based on microsatellite DNA and AFLP markers using homozygous clonal fish in ayu (*Plecoglossus altivelis*). Mar. Biotechnol., 6, pp. 327–334.

Wilson K, YT Li, V Whan, S Lehnert, K Byrne, S Moore, S Pongsomboon, A Tassanakajon, G Rosenberg, E Ballment, Z Fayazi, J Swan, M Kenway, and J Benzie. 2002. Genetic mapping of the black tiger shrimp *Penaeus monodon* with amplified fragment length polymorphism. Aquaculture, 204, pp. 297–309.

Woram RA, C McGowan, JA Stout, K Gharbi, MM Ferguson, B Hoyheim, EA Davidson, WS Davidson, C Rexroad, and RG Danzmann. 2004. A genetic linkage map for Arctic

char (*Salvelinus alpinus*): evidence for higher recombination rates and segregation distortion in hybrid versus pure strain mapping parents. Genome, 47, pp. 304–315.

Wright JE, K Johnson, A Hollister, and B May. 1983. Meiotic models to explain classical linkage, pseudolinkage, and chromosomal pairing in tetraploid derivative salmonid genomes. Isozymes Curr. Top. Biol. Med. Res., 10, pp. 239–260.

Young WP, PA Wheeler, VH Coryell, P Keim, and GH Thorgaard. 1998. A detailed linkage map of rainbow trout produced using doubled haploids. Genetics, 148, pp. 839–850.

Yu ZN and XM Guo. 2003. Genetic linkage map of the eastern oyster *Crassostrea virginica* Gmelin. Biol. Bull., 204, pp. 327–338.

Zhou Z, Z Bao, Y Dong, S Wang, C He, W Liu, L Wang, and F Zhu. 2006. AFLP linkage map of sea urchin constructed using an interspecific cross between S*trongylocentrotus nudus* (Venus) and *S. intermedius* (Mars). Aquaculture, 259, pp. 56–65.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 11
Detection and Analysis of Quantitative Trait Loci (QTL) for Economic Traits in Aquatic Species

Abraham Korol, Andrey Shirak, Avner Cnaani, and Eric M. Hallerman

Aquaculture is a rapidly growing sector of agriculture, with production steadily increasing at about 10% per year (FAO 2006). Many aquaculture stocks have been selectively bred with the aim of improving traits of economic interest. The selection strategy traditionally has been based on an inheritance model considering individual phenotypes, often with pedigree information, with use of appropriate statistical tools. Recent advances in experimental design and development of molecular genetic markers have made it possible to dissect the genetic variability of complex traits into components attributable to the segregation of quantitative trait loci (QTL). A QTL is defined as a chromosomal segment with an effect on expression of a trait of interest. Future breeding plans for fish will involve integration of classical selective breeding methods with selection also upon genetic markers closely linked to segregating QTLs, an approach termed marker-assisted selection (MAS). See Chapter 12. Steps in the detection of QTLs and application of MAS (Poompuang and Hallerman 1997) are (1) genetic marker development, (2) development of a moderate density linkage map, (3) mapping of QTLs using genetic markers, with analysis of interactions between QTLs and development of a model for inheritance and expression of the trait, and (4) practical application of the results of this research by practice of MAS. Variations upon this sequence of steps are possible. Notably, assessment of the possible effects of candidate genes controlling expression of the targeted quantitative trait may be implemented. Such candidate genes may be identified based on knowledge of the physiological function of the gene, comparative analysis using linkage information from other species, from cases of positional overlap between candidate genes and marked QTLs, linking of the physical map of a species to its linkage map, or identification of synteny with model organisms. Whatever the pathway to discovery, once variation directly affecting expression of a trait is identified, gene-assisted selection (GAS) may be practiced.

A wide spectrum of questions about the nature of quantitative variation can be considered in terms of "genetic architecture," which is of primary importance for genetics, functional genomics, developmental and evolutionary biology, and ecology. These include genomic distribution of QTL effects; relative contributions of additive and nonadditive effects to genetic variation and response to selection; the role of overdominance, epistasis, coadaptation, and pleiotropy in such phenomena as heterosis and speciation; developmental (longitudinal) variation in QTL effects, and vice versa, developmental homeostasis or canalization; and genetic basis of the "reaction norm" (QTL-environmental interactions). Many of these questions are internally

"multivariate" (i.e., their formulation is biologically more relevant when trait complexes rather than separate traits are treated in one framework) (Korol et al. 1995). Statistical considerations (i.e., significance, detection power, and mapping accuracy and resolution) comprise a complementary source of motivation for "multivariate" analysis. Thus, low accuracy of estimated chromosome position of detected QTLs is one of the major obstacles for various applications of QTL mapping. We argue that this is caused partially by using simplified methods of QTL analysis that leave untouched a considerable part of mapping information contained in the data. Although the history of aquaculture genetics for quantitative traits is short compared to that for other animal models, there is no doubt that its fast development will be facilitated by the new genomic revolution; hence, the urgent need and high potential of new methods.

In this chapter we present methodologies for detection and analysis of QTLs for economic traits in aquatic species, including both aquaculture and model species. We set the context with a brief consideration of QTL detection and MAS and present a brief overview of progress in detection of QTL in aquaculture species. We then demonstrate computational aspects of QTL detection by analyzing experimental data for tilapia. It is our hope that by demonstrating the use of statistically powerful tools we will advance the discussion of QTL detection and MAS in aquaculture species.

An Overview of QTL Detection

Detection of QTLs and use of the knowledge in MAS can be conceived as having four stages. To provide context for the discussion that follows, we briefly describe each stage of a QTL detection program.

Genetic Marker Development

The potential value of genetic markers for purposes of genetic improvement through MAS depends upon their being linked to segregating QTLs. A segregating QTL allele of moderate effect can be detected reliably by a marker locus occurring within approximately 20 centiMorgans (cM) along the chromosome (Soller et al. 1976). To detect a reasonable proportion of the QTL segregating in a mapping population, a large number of markers need to be well distributed throughout the genome of interest. A quantitative assessment of the experimental power of marker-based screenings for detection of QTLs (Kashi et al. 1990) showed that the number of QTLs identified was more responsive to the degree of polymorphism of marker loci than to the number of individuals screened. A large collection of highly polymorphic marker loci is, therefore, a precondition for successful use of the linkage-based approach to detecting QTLs. A wide range of DNA marker types has been used in fish genomic research, including amplified fragment length polymorphism (AFLP), randomly amplified polymorphic DNA (RAPD), sequence tagged sites (STS), variable number of tandem repeat (VNTR), simple sequence repeat (SSR) and single nucleotide polymorphism (SNP) markers. See the chapters in Part I, as well as Ferguson and Danzmann (1998) and Liu and Cordes (2004).

Genomic Map Development

Genomes may be mapped at three different levels:

- Genetic maps that represent the linear order of markers along a chromosome
- Physical maps that localize large DNA segments onto the cytological karyotype of the species
- Genome sequencing in which DNA sequences are aligned along the respective chromosomes throughout the genome

Genetic maps are the most common means of displaying chromosomal organization, and typically are built of marker arrangements within various linkage groups. Detailed linkage maps have been constructed for a number of aquatic species, including zebrafish (Postlethwait et al. 1994, Shimoda et al. 1999, Woods et al. 2000), rainbow trout (Young et al. 1998, Sakamoto et al. 2000, Nichols et al. 2003a), Atlantic salmon (Lindner et al. 2000), medaka (Naruse et al. 2000), channel catfish (Waldbieser et al. 2001), Japanese flounder (Coimbra et al. 2003), and tilapia (Kocher et al. 1998, Agresti et al. 2000, McConnell et al. 2000, Lee et al. 2005). Genetic maps are built using data on the frequencies of recombination events among pairs of genetic markers transmitted by a given parent. Markers that are physically near one another on a chromosome have a reduced probability of having their allelic phase disrupted during meiosis, and hence a high frequency of parental haplotypes is transmitted to their progeny. In contrast, markers that are located far from one another on a linkage group may experience a crossover event in most meioses generated (Thorgaard et al. 1983, Danzmann and Gharbi 2001). Data for constructing such marker-linkage maps are generally produced using F_2 intercross or backcross designs; that is, F_1 parents produced by mating two related strains or species are mated between themselves or to the parental strains (Danzmann and Gharbi 2001).

In a number of mapping studies, segregation analysis was performed through manipulation of complete chromosome sets (haploids and polyploids) or of the parent of origin (gynogenesis or androgenesis) (Purdom 1969). For example, early mapping of salmonids focused on allozyme polymorphisms, including the use of half-tetrad analysis in meiotic gynogens (Thorgaard et al. 1983, Allendorf et al. 1986), which allowed mapping of genes with respect to the centromeres. Gynogenesis, a form of parthenogenesis, is a process by which the chromosomal complement of the female is preserved, while at the same time the genetic contribution of the male is eliminated. In artificial gynogenesis, irradiated sperm is used to trigger embryogenesis in eggs, but is not involved in syngamy (Monaco et al. 1984, Chourrout 1984). Meiogynes are generated with temperature or pressure shock, which suppresses the second meiotic division. One pair of sister chromatids is retained to create a single diploid embryo. Mitogynes are generated by applying a later shock to suppress the first mitotic division, thereby restoring diploidy (Nagy et al. 1979, Streisinger et al. 1981, Shirak et al. 1998). Androgens are produced by using radiation to inactivate eggs and fertilizing them with sperm; diploidy is restored by suppressing the first cell division (Stanley 1976, Parsons and Thorgaard 1984, Myers et al. 1995). Distances from genes to centromere can be determined if half-tetrads from a single meiosis can be observed. For example, the meiogynes include two sister chromatids that are products of the first meiotic division. The proportion of heterozygous meiogynes (parental genotype) for a gene provides a measure of the recombination frequency between this gene and a centromere (centromere distance). Most distal loci

remain heterozygous after meiogynogenesis due to crossing over between nonsister chromatids in the reductional meiotic division, while paracentromeric loci give rise to homozygotes (Nagy et al. 1979, Streisinger et al. 1981, Thorgaard et al. 1983, Chourrout 1984). Thus, analysis of meiogyne segregations provides valuable information for mapping the centromeres in the existing linkage map. The mathematical formulas based on hypothetical and experimental models allow conversion of the proportion of heterozygotes into centimorgans units for intervals between two loci and between a locus and a centromere (Anderson and Rhoades 1931 in *Drosophila*, Barratt et al. 1954 in *Neurospora*). Such calculations were demonstrated in QTL mapping in trout and tilapia (Allendorf et al. 1986, Don and Avtalion 1990, Shirak et al. 2006). A significant practical problem in the production of gynogens and androgens is the low yield of viable progeny due to the presence of deleterious alleles that are not often expressed in outbred populations (Purdom 1969, Mair 1993). This explanation is supported by observations of increased viability of gynogenetic individuals in successive generations of carp, trout, and tilapia (Nagy and Csanyi 1982, Allendorf et al. 1986, Shirak et al. 1998).

In addition to linkage maps, a variety of genomic tools, such as expressed sequence tag (EST) libraries, bacterial artificial chromosome (BAC) libraries and BAC contig maps have been developed for various fish species (reviewed by Clark 2003). Three fish species, torafugu (*Takifugu rubripes*), spotted green pufferfish (*Tetraodon nigroviridis*), and zebrafish (*Danio rerio*) are now under the process of large-scale genome sequencing (Crollius et al. 2000, Woods et al. 2000, Aparicio et al. 2002, Baden et al. 2002).

A genomic map is useful for QTL detection because with a collection of evenly spaced markers, genome coverage is achieved at minimum cost and QTLs can be mapped to specific marker intervals or chromosomal locations. The genome mapping efforts mentioned above to a large degree were motivated to support QTL detection and MAS experiments.

Ordering multilocus maps cannot be considered just a technical challenge, especially because some fishes display a peculiar deviation from random segregation of nonhomologous chromosomes referred to as pseudo-linkage (Johnson et al. 1987, Sakamoto et al. 2000, O'Malley et al. 2003, Woram et al. 2003). Although pseudo-linkage is observed in many species (Korol et al. 1994, Korol 2001), it is not well known to the mapping community. Pseudo-linkage may affect not only the quality of the map, but in our context also may result in errors in QTL mapping (Peng et al. 2000, Sivagnanasundaram et al. 2004).

Thus, two major problems should be addressed in multilocus genetic mapping: markers that belong to nonhomologous chromosomes should not be assigned to the same linkage group, whereas markers from the same chromosome should be placed on the genetic map in the same order as they reside in the DNA molecule. Under the conditions of small sample size and considerable deviations of recombination rates between nonsyntenic markers from 50%, the problem of clustering cannot be solved by an arbitrary choice of a certain (constant) threshold value of recombination or significance of LOD. Indeed, in experiments with the foregoing characteristics, recombination values between groups of markers from different chromosomes may be smaller than those between adjacent markers within a chromosome. Moreover, even if the markers are clustered correctly into linkage groups, reliable ordering cannot be guaranteed due to various complications stemming from incorrect scoring of DNA markers, missing data, negative interference, distorted segregation, and dominance of some markers. Therefore, in the methodology developed by Korol's group for building

reliable maps, the results of multilocus ordering should be verified based on a resampling (jackknife) procedure (Mester et al. 2003a).

The core procedure of this approach (implemented in the MultiPoint software package, see http://www.multiqtl.com for details) includes the following stages:

1. Clustering of the total set of markers into linkage groups with a stringent initial threshold level of pairwise recombination frequencies ($rf_0 \sim 0.15$), to prevent joining loci from nonhomologous chromosomes in one linkage group due to pseudolinkage (Korol et al. 1994, Peng et al. 2000, Korol 2001).
2. Replacing groups of tightly linked (nonrecombining) markers by their most informative "delegates" (bin markers) that further comprise the skeleton map.
3. Ordering markers in the obtained linkage groups based on the minimum total map length criterion (Mester et al. 2003a, 2003b). The ordering steps are alternated with verification steps and removal (if necessary) of any problematic markers detected. The approach adopted in this system differs from other existing methodologies by its optimization power, allowing rapid ordering of hundreds of loci followed by resampling-based verification of the obtained multilocus maps to characterize the stability of marker ordering rather than the confidence interval of each marker position (in cM). The obtained estimates of neighborhood stabilities and deviations from monotonic increase in recombination rates from the tested marker to its neighbors allow detection and removal of problematic markers, thereby improving map quality.
4. Merging the obtained linkage groups into larger linkage groups by relaxing the threshold rf_0; returning to stage 3, if necessary.
5. Attaching previously removed markers to their best intervals on the skeleton map.

To illustrate the foregoing scheme, we reanalyzed public-domain data on zebrafish (http://zebrafish.mgh.harvard.edu/). The data set includes scores for about 3,850 markers of a small F_2 mapping population ($n = 44$ individuals); hence, the challenge of getting reliable mapping results and importance of the verification procedure. Only codominant markers were selected for this illustration, still amounting to about 3,700 markers. Having in mind the danger of combining nonsyntenic markers in one linkage group (LG), we started with a rather stringent threshold recombination rate rf_0. It appeared that $rf_0 = 0.1, 0.15, 0.20,$ and 0.25 resulted in 44, 29, 24, and 1 linkage groups, respectively. Thus, $rf_0 = 0.15$ was chosen for step 1 of the foregoing algorithm. Implementation of the steps 2–4 resulted in 25 linkage groups, with skeleton maps including from 22 to 56 markers, map lengths (using Kosambi mapping function) varying from 62 to 132 cM, with maximum gap per map varying from 6 to 24 cM. The number of markers attached to the skeleton maps varied from 67 to 133. In addition, a short linkage group also was obtained (25 cM length, with 7 skeleton markers and 14 attached markers). Figure 11.1 illustrates how the verification steps allow detection and removal of problematic markers.

QTL Detection and Characterization

Many of the physiological traits measured in an organism are quantitative at the phenotypic level. These traits usually are controlled by several (or many) genes and are affected by environmental factors. The chromosomal positions of genes underlying these traits are called quantitative trait loci, or QTLs. QTLs are mapped by linkage

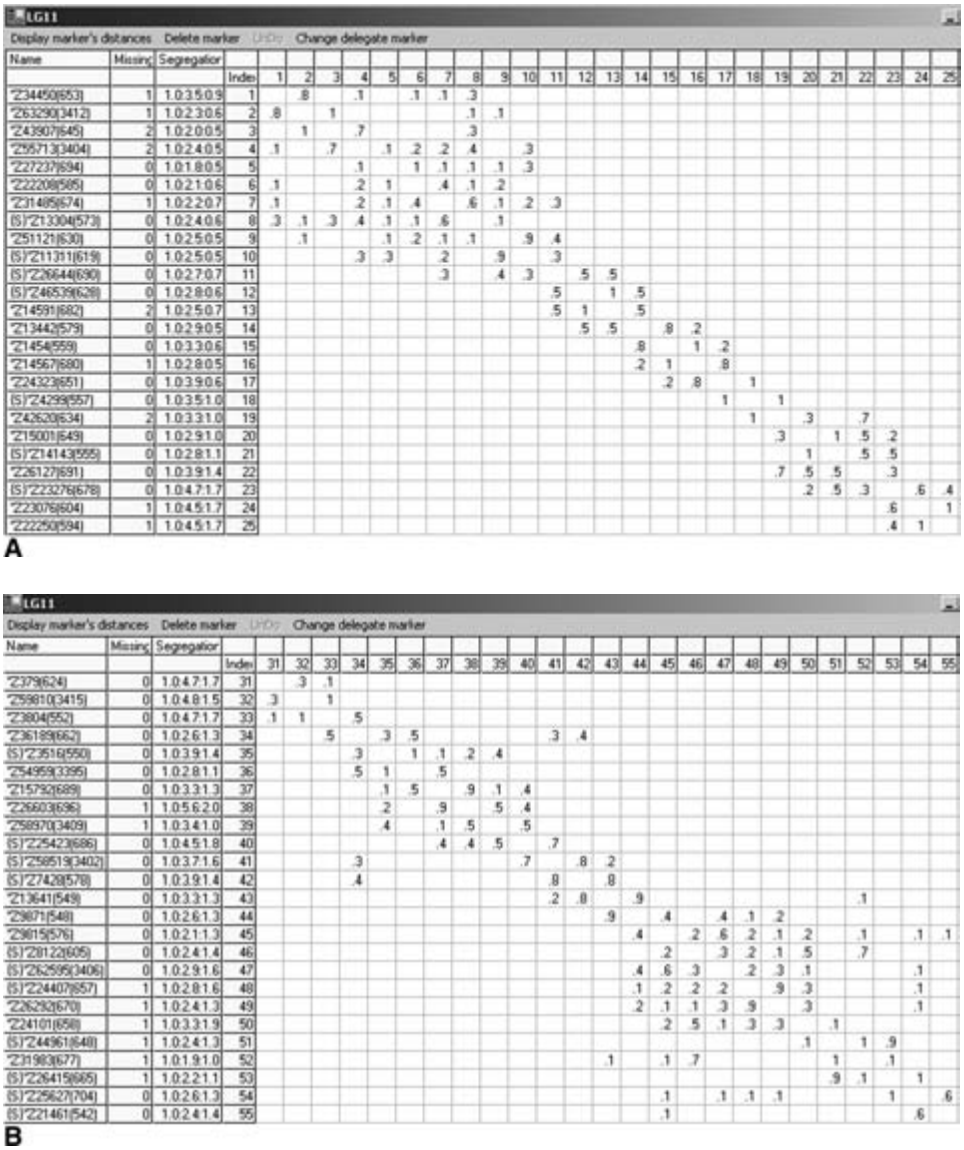
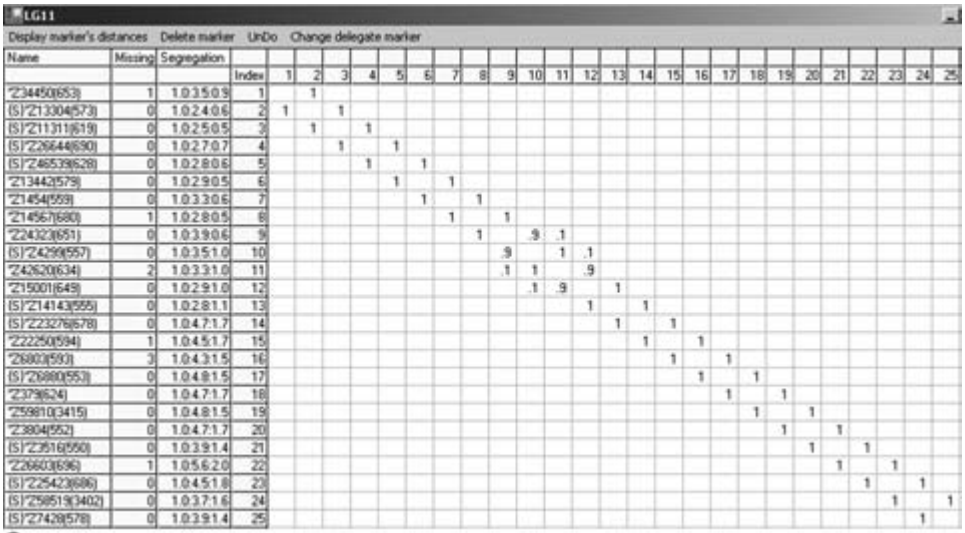


Figure 11.1. Detection of unstable neighborhoods using jackknife resampling (a and b) and improving map stability by removing markers causing the instability (c and d). The example is based on zebrafish data set from Web site http://zebrafish.mgh.harvard.edu/. Parts a and b represent two fragments of a linkage group including 130 markers, whereas parts c and d are the corresponding fragments of the skeleton map. Note the high stability of the skeleton map (for each marker, the frequency of the neighbors on the left hand and right are no less than 0.8, and mostly 0.9–1.0).



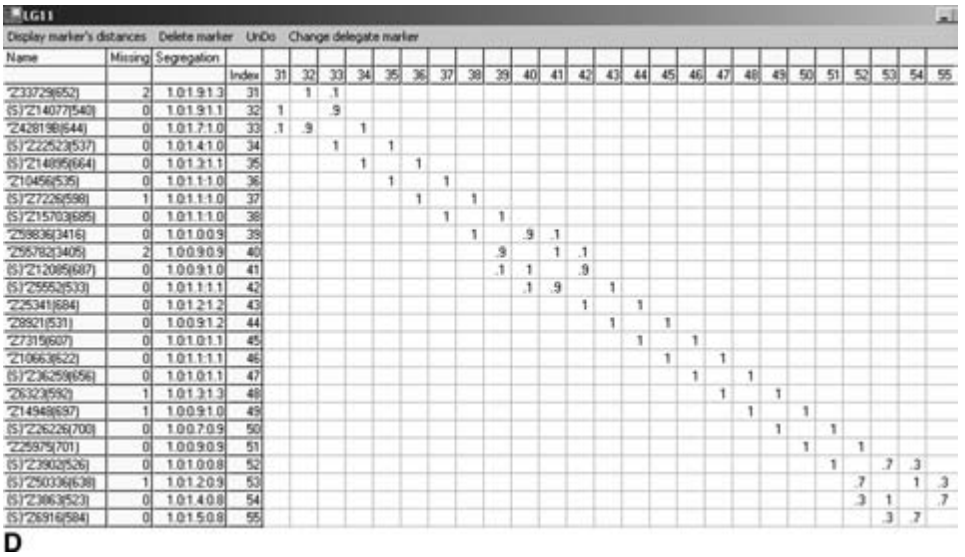


Figure 11.1. (*Continued*)

disequilibrium with molecular markers exhibiting Mendelian segregation. The principle of QTL mapping is straightforward, and was noted more than 80 years ago (Payne 1918, Sax 1923) and recently reviewed by Mackay (2001). If a QTL is linked to a marker locus, there will be a difference in mean values for the quantitative trait among individuals with different genotypes at the marker locus. Consider an autosomal marker locus, M, and a quantitative trait locus, Q, each with two alleles (i.e., M_1, M_2 and Q_1, Q_2). The additive and dominance effects of the QTL may be termed a and d, respectively, and the recombination rate between the marker locus and QTL is c. In the classical F_2 mapping design, by crossing individuals with genotype $M_1M_1Q_1Q_1$ to those with genotype $M_2M_2Q_2Q_2$, and

breeding the F_1 progeny at random, the difference in the mean value of the quantitative trait between homozygous marker classes in the F_2 is $a(1-2c)$. Similarly, the difference between the mean phenotypic value of the homozygous marker classes and the heterozygote is $d(1-2c)^2$. Indeed, let the recombination rate c between marker and QTL be the same in male and female meiosis. Then, at both sides the frequencies of gametes will be $\frac{1}{2}(1-c)\,M_1Q_1$, $\frac{1}{2}\,cM_1Q_2$, $\frac{1}{2}\,cM_2Q_1$, and $\frac{1}{2}(1-c)\,M_2Q_2$. Upon random union of gametes, the groups M_1M_1 and M_2M_2 can be represented as mixtures $\frac{1}{2}[(1-c)^2\,Q_1Q_1{:}2\,c(1-c)\,Q_1Q_2{:}\,c^2\,Q_2Q_2]$ and $\frac{1}{2}[\,c^2\,Q_1Q_1{:}2\,c(1-c)\,Q_1Q_2{:}(1-c)^2\,Q_2Q_2]$, respectively, with mean values $m+a$, $m+d$, and $m-a$. Therefore, the mean values of M_1M_1 and M_2M_2 can easily be found as $m+\frac{1}{2}(1-2c)\,a$ and $m-\frac{1}{2}(1-2c)\,a$, so that the difference between the marker homozygotes is $(1-2c)a$. Similarly, the difference between the mean phenotypic value of the homozygous marker classes and the heterozygote is $d(1-2c)^2$. If the QTL and marker locus are unlinked, $c=0.5$ and the mean value of the quantitative trait will be the same for each of the marker genotypes. The closer the QTL and the marker locus, the larger is the phenotypic difference between the marker genotypes. The maximum difference is observed when the marker genotypes correspond exactly with the QTL (i.e., when the marker and QTL loci are one and the same). Given a population that is genetically variable for the quantitative trait and a polymorphic marker linkage map, a test for differences in trait means between marker genotypes, for each marker, can be done. The marker in a local region exhibiting the greatest difference in the mean value of the trait is thus the one closest to the QTL (Lander and Botstein 1989, Mackay 2001). This suggests the basic principle of a genome scan for QTL. That is, by screening a collection of markers spanning the entire genome, segregation of QTLs throughout the entire genome can be detected. Lander and Botstein (1989) suggested employing interval analysis rather than marker analysis. The difference between these two analytical variants is not important when marker density is high, but not in cases of low density. Indeed, in case of marker analysis the difference between the marker groups decreases proportional to c, whereas in interval analysis it is proportional to c^2, hence higher QTL detection power of interval analysis.

Until two decades ago, the primary limitation for mapping QTL was lack of marker loci. Many subsequent studies have revealed abundant molecular polymorphism at the level of variation of single nucleotides (SNPs), short di-, tri-, or tetra-nucleotide tandem repeats (microsatellites), longer tandem repeats (minisatellites), and mutations at restriction sites (RFLPs and AFLPs) (Mackay 2001). Methods of detection of molecular variation have evolved from RFLP analysis using Southern blots, to high-throughput methods for discovery and genotyping of polymorphisms (Kristensen et al. 2001).

As noted above, QTL mapping is traditionally based on linkage between markers and trait values that occurs within mapping populations or families. The precision with which a QTL can be localized relative to marker loci is proportional to the number of recombination events between the trait locus and markers, which depends on total number of crossovers per chromosome per meiosis, distribution of recombination events along the chromosome, and the sample size. These factors, together with the relatively low individual effects of the QTLs, limit the QTL detection power and, especially, the mapping precision (Ronin et al. 2003). The fact that the average number of crossover exchanges per chromosome is usually low (1–3 exchanges) allows us to screen the chromosomes for presence-absence of QTL effects by using a small to moderate number of markers (e.g., marker density of approximately 20 cM would be enough). However, for fine mapping, the requirements are much more challenging, calling for simultaneous

increase in marker density and sample size, to allow the appearance of the recombinants. Consequently, during the last decade, numerous efforts were devoted in order to improve QTL mapping accuracy and efficiency by moving from *family-based* linkage analysis to *population-based* linkage disequilibrium (LD) analysis. LD analysis uses nonrandom association between QTL and marker alleles at two or more loci within a targeted population. While linkage analysis relies on recombination during 1–3 generations, LD analysis uses recombination events accumulated within the population during hundreds and thousands of generations, hence, its high potential for fine mapping. Besides these historical recombination events, LD is caused by fitness interactions between genes and such nonadaptive processes as population structure, inbreeding, and stochastic effects. In addition to these factors, the "distance" of LD-associations depends on the organism, specific population, and genome regions, and may vary from just a few hundreds of base pairs to hundreds of kilobase pairs. Hence, unlike family-based QTL detection by scanning through the genome linkage maps, LD-based detection needs huge numbers of markers for genome-wise scanning, unless good candidate regions can be suggested. These candidates may be derived from comparative genomics or by using linkage analysis. Therefore, modern fine QTL mapping is a multistep process; an initial genome scan is performed using linkage analysis, followed by higher resolution confirmation studies for detected QTLs, and culminates with LD or association mapping to identify candidate genes (Mackay 2001).

Linkage mapping of QTL in organisms capable of inbreeding begins by choosing parental strains that are genetically variable for the trait of interest. A mapping population then is derived by back-crossing the F_1 progeny to one or both parents, or mating F_1 individuals to create an F_2 population, or constructing recombinant inbred lines (RIL) by breeding F_2 sublines to homozygosity. These methods are very efficient for detecting marker-trait associations, since crosses between inbred lines generate maximum LD between QTL and marker alleles, and ensure that only two QTL alleles segregate, with known linkage phase (Mackay 2001). The choice of method depends on the biology of the organism, and the power of the different methods, given the heritability of the trait of interest (Darvasi 1998).

As the availability of molecular markers increased, guidelines for experimental design and improved statistical methods for mapping QTL were developed. Least-squares (LS) methods test for differences between marker-class means using either ANOVA or regression (Soller et al. 1976). LS methods have the advantage that they easily can be extended to cope with QTL interactions and fixed effects using standard statistical packages, but also have the disadvantage that the assumptions of homogeneity of variances may be violated. Maximum likelihood (ML) (Lander and Botstein 1989) uses full information from the marker-trait distribution, and explicitly accounts for the QTL data being mixtures of distributions (normal distributions usually are assumed). However, ML methods are less versatile, computationally intensive, and require specialized software packages. There is, in fact, little difference in power between LS and ML designs (Haley and Knott 1992).

Two important statistical considerations regarding experimental design and statistical analysis for mapping QTLs are experimental power and significance threshold. In cases where power is low, not all QTL will be detected, leading to overestimation of the detected effects, and poor repeatability of results. The second problem relates to the multiple tests for marker-trait associations in a genome scan analysis. To maintain the conventional experiment-wise significance level of 0.05, a more stringent

significance threshold for each test is needed, based on the number of independent tests. Permutation (Churchill and Doerge 1994) or other resampling methods (Lander and Kruglyak 1995) are widely accepted for providing appropriate significance thresholds. Weller and others (1998) proposed applying the false discovery rate (FDR) for multiple comparisons in QTL analysis.

The number of mapped putative QTLs underestimates the total number of loci that potentially contribute to the genetic variations in the traits. Increasing the sample size would enable mapping of QTLs with smaller effects and separating linked QTLs on the basis of a larger number of recombinant events. The challenge of high-resolution QTL mapping is that individual QTLs are expected to have small effects that are sensitive to the environment. QTLs, as detected in genome scans, are not genetic loci but relatively large chromosomal regions containing one or more loci affecting the trait. The large number of genes in the chromosomal intervals to which QTLs map limits the genetic inferences that can be drawn from analysis of most mapping populations (Mackay 2001).

Association studies of candidate genes can be conducted without *a priori* evidence for linkage (Tabor et al. 2002). Knowing that a given gene is likely to be important in a certain physiological process can mark this gene as a candidate gene for influencing a trait of interest. Once the candidate gene has been located in the genome, characterized and sequenced, its nature of expression and influence on physiological function can be studied (Danzmann and Gharbi 2001).

A practical approach for QTL mapping based on a "multiple" analytical strategy that allows a significant improvement of mapping quality is discussed here. Our proposed general scheme includes joint analysis of multiple-trait complexes (MLT), multiple-environment data (ME), multiple-interval mapping (MIM), and combined analyses MLT × MIM, ME × MIM, and even ME × MLT × MIM. Clearly, multivariate methods do not automatically guarantee an improvement of QTL mapping results. See the discussion in Korol and others (2001). Indeed, combining in one analytical framework multiple traits or a trait scored in multiple environments may lead to both technical obstacles and principal complications. In some situations, too many QTLs spread along a chromosome may underlie the scored trait complex, resulting in a decrease rather than an increase of mapping precision. Nevertheless, despite these constraints, analysis of numerous data sets supports a clear tendency of considerably increasing QTL mapping quality, sometimes dramatically, by moving from simplistic single-trait to more sophisticated multivariate methods. It appears that a more detailed analysis of the same data set may not only increase QTL detection power, but also improve the QTL mapping accuracy (the confidence interval of the estimated QTL position). Thus, with a rather modest sample size (n = 100–200), one may reach fine mapping (2–3 cM or less) by using MLT × MIM or ME × MIM combinations. Such an improvement may have highly important consequences for evolutionary, ecological, and breeding applications of QTL analysis.

Application to Aquatic Species

Against the background of the general overview of theory of QTL detection, we turn now to applications of aquaculture species. We start with a brief overview of progress to date on aquaculture species, and then turn to a prospective view, discussing new multiple analytic approaches that could be applied to future QTL detection and analysis studies.

Retrospective View

Progress in development of genetic markers and maps and detection of QTLs for production-related traits differs among aquaculture species. To show this progress, we briefly consider progress for salmonids and tilapias.

The salmonids include several aquaculturally important species, including Atlantic salmon (*Salmo salar*), rainbow trout (*Oncorhynchus mykiss*), and coho salmon (*O. kisutch*). Major breeding programs have been undertaken for salmonids. Key traits of breeding interest include growth rate, disease and parasite resistance, carcass traits (dress-out, flesh color, low gaping frequency), and upper thermal tolerance. Genetic linkage maps have been developed for rainbow trout (Young et al. 1998) and Atlantic salmon (Hoyheim et al. 1998). BAC physical maps have been constructed for these species (Palti et al. 2004, Ng et al. 2005). The rainbow trout cytogenetic and genetic linkage groups have been integrated (Ruth Phillips, University of Washington-Vancouver, personal communication). The Atlantic salmon and coho salmon linkage groups have been partially assigned to chromosomes, including dimorphic sex chromosomes (Phillips et al. 2005, Artieri et al. 2006). Comparative maps among the salmonids have been developed by Danzmann and others (2005). Because of their high economic value, considerable research effort has addressed QTL detection in salmonids. QTLs have been detected for growth rate, disease and parasite resistance, upper thermal tolerance, developmental rate, precocious male maturation, spawning date, and other traits of selective breeding or evolutionary interest (Table 11.1).

Tilapia (*Oreochromis* sp.) production is especially important in many developing countries, although tilapias have been widely introduced globally because of their hardiness, disease resistance, and ease of culture. Breeding issues for tilapia include growth rate, sex determination, cold tolerance, and disease resistance. The ability to cross *Oreochromis* and *Sarotherodon* species and thereby create viable interspecific hybrids makes them an ideal organism for mapping studies using backcross or F_2 families (Poompuang and Hallerman 1997; Cnaani et al. 2003, 2004b). The most recent version of the developing tilapia linkage map (Lee et al. 2005) consists of 24 linkage groups, while the genome consists of 22 pairs of chromosomes (Martins et al. 2004). The tilapia map spans 1,311 cM and consists of more than 520 microsatellite markers and approximately 30 Type I gene markers. Recently, mapping of 11 markers for genes of the sex-determination pathway and the *Dax1* locus merged two linkage groups, LG 16 and LG 21 (Shirak et al. 2006). QTLs have been detected for growth rate, sex determination, cold tolerance, stress response, disease resistance, survival, and coloration (Table 11.1).

Marker development, genetic mapping, and QTL detection experiments for other aquaculture species—including channel catfish, basses, shrimps, and oysters—have been initiated, although results to date are more limited in scope.

Prospective View

Against the background of general principles and approaches for QTL detection described above, illustrated by reference to case studies in the scientific literature, we now take a prospective view, showing how multiple analytic tools can contribute

Table 11.1. Examples of QTLs detected in aquatic species, with indication of primary source of interest in the trait (SB = selective breeding, E = evolutionary) and genetic architecture where known.

Trait	Genetic architecture	Reference
	Atlantic salmon	
Resistance to infectious salmonid anemia and Aeromonas salmonicida	No apparent effect of genetic background	Grimholt et al. 2003
Resistance to infectious salmonid anemia	Two loci	Moen et al. 2004b
	Rainbow trout	
Growth rate, condition factor, precocious maturation	Possible epistasis among growth rate and precocious male maturation	Martyniuk et al. 2003
Growth rate, spawning date	Possible pleiotropy among body size and spawning date	O'Malley et al. 2003
Growth rate	Two loci, one major	Reid et al. 2005
Condition factor	Four loci, one major	Reid et al. 2005
Resistance to Ceratomyxa shasta	Polygenic	Nichols et al. 2003
Resistance to infectious hepatic necrosis virus	Polygenic	Palti et al. 1999 Palti et al. 2001
	Two or three loci, one with possible pleiotropy on embryo length	Robison et al. 2001, Khoo et al. 2004
	Associations with three linkage groups	Rodriguez et al. 2005
Killer cell-like activity	Single chromosomal region	Zimmerman et al. 2004
Upper thermal temperature	Two loci, additive interaction	Jackson et al. 1998
	Epistasis among paternal alleles depends on genomic background	Danzmann et al. 1999
	Three loci, one major	Perry et al. 2001
Upper thermal tolerance	Two loci	Somorjai et al. 2003
Meristic traits	Different loci associated with each meristic trait; expression may be affected by maternal and environmental factors	Nichols et al. 2004
Pyloric caecae number	Three major loci	Zimmerman et al. 2005
Albinism	Dominant, single-gene trait	Nakamura et al. 2001
Spawning date	Highly polygenic; 13 markers in even linkage groups	Sakamoto et al. 1999
	Coho salmon	
Flesh color	Additive genetic variance with environmental influence; one linked marker	Arenada et al. 2005
	Arctic char	
Growth rate	Association with candidate gene GHRH/PACAP2	Tao and Boulding 2003

Table 11.1. (*Continued*)

Trait	Genetic architecture	Reference
Upper thermal tolerance	Two loci	Somorjai et al. 2003
Growth rate	Two loci, one major	Reid et al. 2005
Condition factor	Four loci, one major	Reid et al. 2005
	Tilapias	
Growth rate	Three loci in three linkage groups	Cnaani et al. 2004
Sex determination	Three loci in three linkage groups	Cnaani et al. 2004
Stress response	Seven loci in five linkage groups	Cnaani et al. 2004
Immune response	Six loci in five linkage groups	Cnaani et al. 2004
Prolactin expression level, Growth under salt challenge	Polymorphism in candidate gene	Streelman and Kocher 2002
Sex ratio, susceptibility to inbreeding	Epistasis among sex determination and viability	Palti et al. 2002
Sex ratio and viability	Three loci with epistasis	Shirak et al. 2002
Sex	Epistasis among XY locus in lg01 and WZ locus in lg03	Lee et al. 2003, 2004
Sex	One locus	Moen et al. 2004a
Cold tolerance	One locus	Cnaani et al. 2003
Growth	One locus	Cnaani et al. 2003
Sex	One locus	Cnaani et al. 2003
Survival	Three loci	Cnaani et al. 2003
Disease resistance	Lysis level codominantly expressed	Shirak et al. 2006 Shirak et al. 2000
Body and peritoneum coloration	Body color—single gene inheritance; peritoneum color—interaction of R and D genes	
Survival, sex, and red body coloration	One locus for each trait	Lee et al. 2005
	Common carp	
Cold tolerance	Four markers associated with trait, one mapped to lg05	Sun and Liang 2004

greater statistical power to QTL detection experiments. In the examples presented below, our main objective is to illustrate some of those analytical techniques for QTL mapping using Cnaani and others' (2004) performance and molecular marker data on F_2 hybrid tilapia. We try to show that employing more sophisticated methods and models allows us to diversify the QTL analysis, thereby improving its quality. Aspects of "quality" that may be improved follow:

1. Significance level of QTL detection (i.e., false positive detection level)
2. QTL detection power (level of false negatives)
3. Accuracy of parameter estimation, especially for the QTL chromosomal position
4. Possibility to test (discriminate) biologically important hypotheses (e.g., additive versus dominant or overdominant effects; linkage versus pleiotropy, additive versus

epistatic effects for linked or unlinked QTLs, QTL-environmental interaction with respect to QTL effects, and the genetic architecture aspects mentioned above

We believe that the examples below demonstrate the advantages of using a wide spectrum of models and analytical techniques in order to support more efficient extraction of the mapping information hidden in the data and thereby reach a better understanding of the genetic architecture of the targeted trait complex and underlying biological process(es). Thus, the examples below are merely illustrations; a more exhausting analysis of this data set will be provided elsewhere. For illustrations, we will concentrate mainly on linkage groups 19 and 23 (LG 19 and 23, respectively) (Lee et al. 2005). All analyses were conducted using the MultiQTL analytic package (http://www.multiqtl.com).

Using the examples below, we demonstrate the spectrum of tools relevant to current and future QTL mapping efforts for aquaculture organisms, despite the limitations of the available data set. In the following analyses, p is statistical significance (upon permutation tests), SD (L) is standard deviation of the estimated QTL position, $P_{0.01}$ is QTL detection power at the significance level $p = 0.01$, *PEV* is the percentage of explained phenotypic variance, and d and h are the (doubled) additive and heterozygous QTL effects, respectively. In displaying the results, we show the difference between the two homozygotes for a QTL affecting trait X with the designation $d = X(QQ) - X(qq)$.

The Mode of QTL Action (Simplifying the Model for QTL Effect)

Generally, for an F_2 population, the QTL can be characterized by additive (d) and heterozygous (h) effects. Various hypotheses about the QTL effect can be considered and compared: pure additive effect ($d \neq 0$ and $h = 0$), dominant ($h = d/2$) and recessive ($h = -d/2$) effects, no additive effect ($d = 0$ and h $\neq 0$), and general effect ($d \neq 0$ and $h \neq 0$). It makes sense to test whether or not one of the foregoing simplifications of the general (d, h) model can be accepted. Indeed, using a simplified model (e.g., $d \neq 0$ and $h = 0$) that does not differ significantly from the general model actually means excluding a nonsignificant (excess) parameter. In many cases, this allows increasing the quality of the results (Table 11.2). The examples provided in Table 11.2 (in the section on "single-trait analysis") show that removing the excessive parameters from the model might considerably improve one or several of the quality parameters (e.g., significance level and precision of QTL location). No less important is that this procedure is helpful in testing the mode of expression of the QTL effect. Thus, for the trait "weight," the effect of the QTL on LG19 is highly significant, but its additive effect (d) is close to zero. The LOD score associated with the simplified model ($d = 0$, h) does not differ significantly from that for the general model (d, h), which means that LG19 carries a QTL with a negative overdominant (heterosis) effect on animal body weight. In another example, on change in ceruloplasmin (ferroxidase) score (Δcerulopl) caused by challenging the animal to air exposure, the obtained relationship between the estimates of parameters for QTL effect on LG19 ($d = 111 \pm 30.4$ and $h = -41.8 \pm 27.8$) is very close to that expected for a recessive QTL effect ($h = -d/2$). This simplified model did not differ from the initial one, and its acceptance resulted in a nearly five-fold improvement of significance (p-value) and considerable narrowing of the confidence interval for the estimated QTL effect and location.

Multiple-Trait analysis

Under small sample size, a QTL with pleiotropic effects on various traits may not be detected if each of its individual effects is low. But even with significant effects, the mapping resolution and location precision based on single-trait analysis remain disappointingly low. These properties could be considerably improved by moving to multiple-trait QTL analysis (Korol et al. 1995, 2001). In addition to pleiotropy, multiple-trait analysis may be helpful in situations when the targeted traits are controlled by tightly linked QTL. The application of multitrait analysis to situations of linked or pleiotropic QTLs could bring about a remarkable improvement of mapping results especially for correlated trait complexes (Korol et al. 2001). Moreover, joint analysis of two correlated traits may improve the efficiency of mapping even if only one of the traits is affected by the analyzed chromosome. See Korol and others (1995). The second trait serves in such a case as a covariate (as in ANOVA with covariates). We illustrate here the advantages of two-trait analysis, although, in principle, many more traits can be analyzed simultaneously. The chosen examples extend those of single-trait analysis, and in turn, are further extended by using genome-wise MIM analysis for the two-trait combinations considered.

For the QTL from LG19 affecting the lysozyme trait, the results of mapping become much more precise when the analysis also includes the correlated trait albumin (the residual correlation of these traits within the QTL groups is $R_{xy} = +0.46$). In particular, in none of 10,000 permutations was the two-trait LOD = 6.07 exceeded (hence, the significance $p < 5 \times 10^{-5}$). In fact, instead of evenly distributed LOD along this linkage group revealed by single-trait analysis, two-trait analysis detected a QTL at the right end of the linkage group. The detection power estimated based on 2,000 bootstrap runs was 97.5% (compared to 83.0% for single-trait analysis), and the standard deviation of the QTL position decreased to just 3.0 cM (from 10.6 cM in single-trait analysis). This may reflect the real situation, although we cannot rule out the possibility that this linkage group includes two QTLs; the modest sample size and availability of only three markers for this linkage group do not allow reliable testing of linked-QTL models. Within the class of single-QTL models for each linkage group, a further improvement was achieved when this trait pair was analyzed using the MIM approach for all available linkage groups (plus separate markers. See Table 11.2 and Figure 11.2). In this case, the LOD score was 11.55, the detection power reached 100%, and the estimated QTL effects on lysozyme and albumin were seven-fold and four-fold greater than their standard deviations.

QTL-E Analysis

Like multiple-trait analysis, joint analysis of a trait scored across several environments may considerably improve the quality of mapping results (Jansen et al. 1995, Korol et al. 1998, Yagil et al. 2006). However, even more important is its suitability for testing corresponding biologically important hypotheses, such as the existence of QTL × E interaction, and if the answer is positive, characterization of the environmental dependence of the corresponding parameters (additive effect, epistasis, and residual variation of the trait). Table 11.3 displays consequent steps of such analysis

Table 11.2. Illustration of improved quality of QTL mapping using two-trait analysis combined with multiple interval mapping.

Lg	Trait(s)	Model	LOD	p	$P_{0.01}$, %	SD(L) cM	d	h	PEV, %
					Single-trait analysis				
lg19	lysozyme1	1. d, h	2.83	0.0076	72.1	10.1	−16.8 ± 3.7	−5.1 ± 3.6	19.1 ± 5.8
		2. h=0	2.58	0.0018	83.0	10.6	−17.2 ± 3.2		15.8 ± 4.9
lg19	Δcerulopl	1. d, h	2.75	0.0058	69.0	4.7	111 ± 30.4	−41.8 ± 27.8	39.2 ± 10.6
		2. h=−d/2	2.70	0.0012	74.9	1.5	101 ± 16.8		34.3 ± 10.0
lg19	weight	1. d, h	4.64	0.0004	91.5	9.7	−3.2 ± 19.0	−45.6 ±12.5	22.5 ± 7.9
		2. d = 0	4.63	0.0003	95.1	7.4		−45.7 ± 10.9	20.5 ± 6.5
					Two-trait analysis				
lg19	lysozyme1-albumin	h1=0 & h2=0	6.07	$<5{\cdot}10^{-5}$	97.5	3.0	−16.7 ± 3.8		17.4 ± 6.3
							0.224 ± 0.098		5.9 ± 4.2
		h1=0, h2= d2/2, +MIM	11.55	$<5{\cdot}10^{-5}$	100	2.4	−15.1 ± 2.0		18.4 ± 4.0
							0.251 ± 0.067		5.1 ± 2.5
							* Rxy = +0.46 /+ 0.65		
lg19	Δcerulopl-globulin	d, h	4.91	0.0010	83.3	4.7	109 ± 18.4		41.4 ± 10.6
		h1=−d1/2 h2=−d2/2		$<5{\cdot}10^{-5}$	99.9	2.6	0.412±0.194		20.6 ± 15.2
		h1=−d1/2	10.07				107 ± 17.1		31.4 ± 6.8
		d2=0, +MIM					Rxy = −0.37 / 0.34	− 0.399 ± 0.111	18.9 ± 8.0
lg23	glucose2-weight	h1=d1/2 d2=0	5.14	0.0002	94.5	5.4	7.54 ± 6.07	−50.5 ± 10.6	4.0 ± 4.4
									28.1 ± 9.1
		h1=−d1/2	11.68	$<5{\cdot}10^{-5}$	100	1.8	5.54 ± 3.28		1.6 ± 1.5
		d2=0, +MIM						−41.5 ± 5.6	18.6 ± 4.0
							Rxy = +0.15 / +0.46		
lg10	glucose2-weight	d1=0, h1=−d1/2	2.28	0.036	54.5	14.5		10.9 ± 8.5	11.0 ± 10.7
							61.2 ± 22.0		34.2 ± 19.2
		d1=0,	6.78	$<5{\cdot}10^{-5}$	97.7	12.5		12.9 ± 4.9	10.1 ± 5.5
		h1=−d1/2, +MIM					46.7 ± 8.3		19.7 ± 5.7
							Rxy = +0.13 / +0.28		

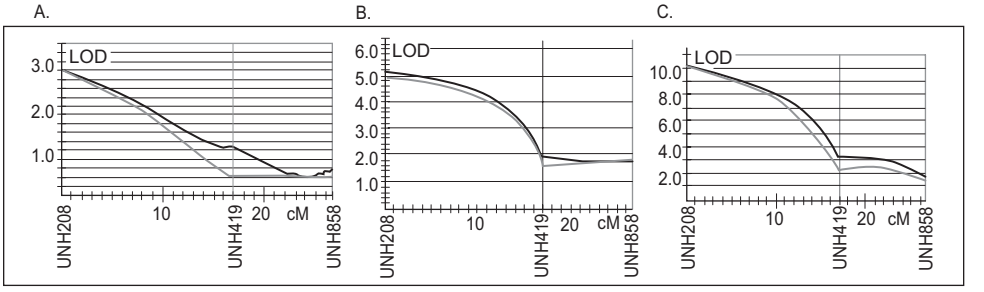


Figure 11.2. Improving the quality of QTL mapping for trait "change in ceruloplasmin" (Δcerulopl) using two-trait analysis combined with MIM. (a) Single-trait analysis of Δcerulopl; (b) two-trait analysis of Δcerulopl and globulin; (c) the same as (b) but combined with MIM. Red—general model, green—simplified model (as shown in Table 11.2).

on some examples for single- and two-trait situations. For the first example, for the glucose QTL in LG23, the general (unconstrained) model proved not to be significant by permutation test ($p = 0.068$). However, despite the total nonsignificance (over two environments), the estimate representing the substitution effect in E2 (after challenging the animals to stress) is sixfold higher than that in E1 (before challenging the animals) and threefold higher than its standard error (25.4 ± 8.1). This calls for testing the hypothesis H_0 ($QTL \times E = 0$ or $E1 = E2$) with respect to parameter d; the result was $P(QTL \times E = 0) = 0.029$, allowing rejection of H_0. Having in mind the nonsignificant estimates of d and h for E1, we can build a new submodel with $d1 = h1 = 0$. With respect to E2, it appears that, given the condition $d1 = 0$ and $h1 = 0$, two models, with additive effect ($h2 = 0$) and recessive effect ($h2 = -d2/2$), do not differ significantly from the general model ($d2, h2$), but the recessive model gives a slightly higher LOD value and more precise parameter estimates. Fitting this model concludes the analysis.

The result obtained on the effect of LG23 on glucose scores in E2 is very similar to the result that could be obtained if data on E2 were analyzed separately. This is due to the fact that, in this example, the QTL × E interaction is displayed in the extreme form that the effect in E1 was zero. An opposite situation is displayed by the second example regarding the effect of LG10 on hematocrit scores. The individual analysis of this trait in either E1 or E2 could not detect any significant effect despite the similarities of LOD graphs. Joint analysis in E1 and E2 indicated that if these regions do affect the trait, the effect can be described using a dominance model for both E1 and E2 (i.e., $h1 = d1/2$ and $h1 = d1/2$). This simplification resulted in an increase of significance and accuracy of the estimates. Joint analysis in E1 and E2 indicated that if this region does affect the trait, there should not be QTL × E interaction (the corresponding test gives $P[QTL \times E = 0] = 0.865$). Thus, using the dominance model with the assumption of 'no QTL × E interaction' gives the final model with a reasonable significance ($p < 0.01$), detection power ($P_{0.01} = 57\%$), and estimation accuracy (d is threefold greater than its standard error).

More complicated situations are displayed for the example of a joint QTL × E analysis of the two-trait combination ceruloplasmin-lysozyme with LG19 markers. Using a general (d, h) model with no restrictions on the effects of either of the two traits in both E1 and E2 environments, a highly significant ($p \approx 0.001$) pleiotropic QTL was

found (with detection power $P_{0.01} = 93.5\%$). Fitting the H_0 model ($QTL \times E = 0$) simultaneously for both additive (d) and heterozygous (h) effects and both traits and testing it against the unrestricted H_1 ($QTL \times E \neq 0$) allows rejection of H_0 at the $p = 0.002$ level of significance. However, it appeared that the two environments differed not only for the parameters characterizing the QTL effects; the residual variance of lysozyme (but not ceruloplasmin) in E2 was much higher than that in E1. Hence, direct comparison of QTL effects in E1 and E2 was not possible, because $d1 \neq d2$ does not exclude the possibility that $d1/\sigma 1 \approx d2/\sigma 2$ may still hold. To manage this problem, the analysis should be conducted on variables normalized to keep $\sigma 1 \approx \sigma 2$. After moving to normalized variables, the difference between $H_0 = (QTL \times E = 0)$ and H_1 ($QTL \times E \neq 0$) becomes even more significant than for the initial data (p was reduced from 0.002 to 0.0006). The results for the normalized model simplified to take into account the dominance-recessiveness of the QTL effects are shown in Table 11.3. The considerable improvement of significance and detection power caused by this simplification is noteworthy ($p < 5 \cdot 10^{-6}$, $P_{0.01} = 99.6$). What remains is to test whether the highly significant deviation from $H_0 = (QTL \times E = 0)$ is caused by the QTL effect on ceruloplasmin (C) or lysozyme (L). This can be achieved by assuming separately $C1 = C2$ and $L1 = L2$. As can be seen from the results in the table, the assumption $C1 = C2$ causes a highly significant (0.0004) reduction of LOD compared to the level characteristic of H_1, whereas $L1 = L2$ is quite compatible with H_1 (significance >0.40).

Epistasis

An important aspect of the genetic architecture of quantitative traits is epistasis. Our next example demonstrates the complicated possibilities of testing epistasis between linked QTLs from LG23, in the context of QTL-E analysis, using data on lysozyme scored before (E1) and after (E2) challenging the animals. Due to the abovementioned observation about higher residual variance in E1, the analysis is conducted on normalized data. In this model, the two linked QTL will be referred to as a and b (Table 11.4). Epistasis in each environment is represented by a vector of epistatic parameters ε ($\varepsilon 1$, $\varepsilon 2$, $\varepsilon 3$, $\varepsilon 4$) (for additive-additive, additive-heterozygous, heterozygous-additive, and heterozygous-heterozygous interactions, respectively).

The first step is to test, using the unconstrained model, whether the H_0 hypothesis (no effect of the linkage group on the trait) and the H_1 hypothesis (only one QTL in the linkage group affects the trait) can be rejected when compared to the H_2 alternative hypothesis (two linked QTL in the linkage group affect the trait). The tests conducted allow rejecting both H_0 and H_1 in favor of H_2. The next step is testing whether epistasis is significant in either E1 or E2. This can be achieved by setting separately $\varepsilon 1 = 0$ and $\varepsilon 2 = 0$, fitting the resulting constrained models and comparing them with the unconstrained model (where $\varepsilon 1 \neq 0$ and $\varepsilon 2 \neq 0$ simultaneously). These comparisons gave $P(\varepsilon 1 = 0) = 0.208$ and $P(\varepsilon 2 = 0) = 0.019$. In other words, epistasis was significant after challenge (in E2), but not before (in E1). The results of 1,000 bootstrap runs of the final fitted model (with the parameters shown in Table 11.4) are presented in Figure 11.3. It is noteworthy that the solution for nearly 75% of runs was on the interval pair 1–4. In general, we can conclude that QTL-environmental

Table 11.3. Illustration of analysis of chosen quantitative traits (glucose, hematocrit, ceruloplasmin, and lysozyme levels) across environments (E 1 and 2), as single and multiple trait analyses.

Lt	Trait	E	Model	LOD	p	$P_{0.01}$[a] %	SD(L) cM	d	h	P{*QTL* × *E*=0}
					Single-trait analysis					
lg23	glucose	1	dl, hl	2.83	0.068	48.7	10.3	3.7 ± 6.2	−0.9 ± 6.8	0.029
		2	d2, h2					25.4 ± 8.1	−6.6 ± 7.0	
		1	h = d =0	2.71	0.026	46.5	6.1			
		2	d2, h2					23.7 ± 4.8	−8.3 ± 5.8	
		1	h1 = d1 = 0	2.60	0.007	59.9	4.5			
		2	h2 = −d2/2					23.7 ± 4.8		
lg10	hemato	1	dl, h1	2.63	0.078	44.5	12.1	6.77 ± 4.15	2.54 ± 2.82	0.865
		2	d2, h2					5.86 ± 2.88	1.84 ± 1.95	
		1	h1 = d1/2	2.53	0.021	51.8	7.9	7.16 ± 3.71		
		2	h2 = d2/2					5.04 ± 2.33		
		1,2	d1 = d2 = d	2.51	0.007	56.9	6.2	5.88 ± 1.71		
					Two-trait analysis					
lg19	cerulopl	C1	d1, h1					−16.7 ± 14.1	11.4 ± 11.6	
	(C)	C2	d2, h2					72.6 ± 23.2	−43.8 ± 17.2	
	lysozyme	L1	d1, h1	8.82	0.001	93.5	5.1	−15.4 ± 4.7	−7.4 ± 3.6	0.002
	(L)	L2	d2, h2					−26.9 ± 19.3	7.1 ± 12.8	(0.0006)
		C1	h1 = −d1/2					−. 517 ± 0.352		
		C2	h2 = −d2/2	8.68	$<5\cdot10^{-6}$	99.6	5.1	1.53 ± 0.52		
		L1	h1 = d1/2					−. 924 ± 0.245		
		L2	h2 = −d2/2					−. 519 ± 0.388		
		C1 = C2	h = −d/2							
		L1	h1 = d1/2	4.22	0.005	81.2	10.7			0.0004
		L2	h2 = −d2/2							
		C1	h1 = −d1/2					−. 521 ± 0.347		
		C2	h2 = −d2/2	7.68	$<5\cdot10^{-6}$	99.1	3.9	1.55 ± 0.42		0.433
		L1 = L2	h = 0					−. 842 ± 0.241		

Table 11.4. Results of tests for epistasis between linked QTL as a part of analysis of QTL-environmental interactions (data on lysozyme).

			p		SD(L) cM for interval			*d*		*h*	
E	Model	LOD	H_2:H_0	H_2:H_1	1	4	$P(\varepsilon_i = 0)$	a	b	a	b
1	d1a, h1a, d1b, h1b, ε1	10.09	0.021	0.009			0.208				
2	d2a, h2a, d2b, h2b, ε2						0.019				
1	d1a, h1a, d1b, h1b,ε1=0							−0.902 ± 0.450	0.607 ± 0.436	0.021 ± 0.383	−0.385 ± 0.293
2	d2a, h2a, d2b,h2b, ε2	8.20	0.011	0.008	8.3	9.1		−0.906 ± 0.691	0.206 ± 0.734	0.961 ± 0.501	−0.599 ± 0.607
1	d1a, h1a=0, d1b, h1b=−d1b/2, ε1=0	8.17	0.002	0.002	8.2	8.2		−0.928 ± 0.287	0.668 ± 0.287	0	−d1b/2
2	d2a, h2a, d2b=0,h2b, ε2							−0.860 ± 0.597	0	−1.03 ± 0.507	−0.638 ± 0.601

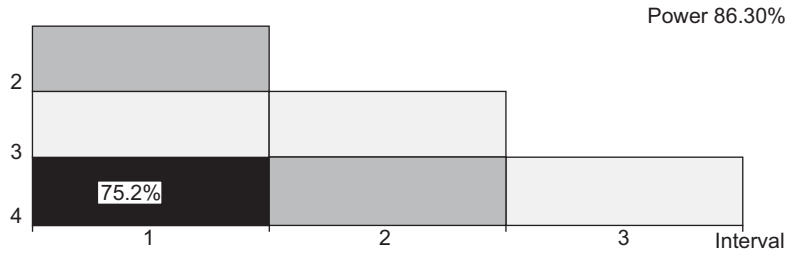


Figure 11.3. QTL-E analysis of a two-QTL model with epistasis for the effect of lg01 on lysozyme. The results of 1,000 bootstrap runs for the simplified model as shown in Table 11.4.

interaction, as a part of genetic architecture and reaction norm, may include significant dependence of epistasis on external conditions.

Marker-assisted Selection

Marker-assisted selection (MAS) is covered in detail in Chapter 12, therefore, here we provide only conceptual introductions. Once QTLs have been detected and their relative quantitative effects have been estimated, how can this knowledge be applied in practical marker-assisted selective breeding? Two modes of application may be anticipated. First, we can select directly upon the family material on which the QTL mapping was done. We anticipate that this mode of MAS will have limited applicability, and mostly for basic research. Second, we can select upon existing commercial broodstocks. We anticipate that this mode of MAS will have more general applicability. The commercial broodstock will be screened for segregation of QTLs of interest and to determine the coupling of marker and QTL alleles specific to families within that broodstock. MAS has been well demonstrated in crop plant systems (Collard et al. 2005), such as corn (Yousef and Juvik 2002) and millet (Serraj et al. 2005). If the gene directly affecting a trait is known (as opposed to a genetic marker linked to that gene on the chromosome), then gene-assisted selection (GAS) can be applied. In agricultural animal systems, MAS is being applied to increase litter size in pigs (Rothschild et al. 1996, Visscher and Haley 1998). GAS is being applied to increase scrapie resistance (DEFRA 2006) and to decrease incidence of spider syndrome in sheep (R. Lewis, Virginia Polytechnic Institute and State University, personal communication). To our knowledge, neither GAS nor MAS have yet been applied to aquaculture species.

A key practical question is whether MAS can accelerate genetic progress to the degree that it is cost-effective. The power of selective breeding plus MAS will have to be demonstrated relative to conventional, phenotype-based breeding alone. The efficacy of MAS depends on three factors, the heritability of the trait, the proportion of genetic variance associated with marker(s), and the selection scheme at issue (Lande and Thompson 1990). The relative efficiency of MAS in relation to conventional selective breeding is highest for low heritability traits when selecting on the basis of an individual-based index combining both genetic marker and phenotypic information. Development of selection indices combining phenotypic and marker information depends upon the relationships among individuals, their breeding values as estimated

using classical animal models, and the phenotypic effects of marked QTLs (Spelman and Garrick 1998). The theory for estimating selection indices has been particularly well established for dairy cattle (Hoeschele 1993, Weller 1997). Since fish families can be large and can be reared in single units, the parameters entering the analysis presumably can be estimated with considerable precision. Still, the technical basis for development of selection indices for fishes needs more theoretical work.

Like classical selective breeding, MAS could be particularly effective for aquatic species. In aquatic species, most traits of economic interest are well suited to MAS:

1. Most traits are not sex-limited.
2. Generation intervals in many species are short.
3. Most species have external fertilization.
4. Many traits of interest are heritable.
5. Most species have large progenies.

The prospects for genetic improvement of aquaculture species are good (Gjedrem 1983, 1985). However, MAS cannot be cost-effectively applied for every trait. High heritability traits might best be improved by classical, phenotype-based selection. Traits for which MAS would be most appropriate include sex-limited traits, traits expressed late in life, carcass traits, and low heritability traits (Poompuang and Hallerman 1997). Given that many key traits, such as growth rate, often have high enough heritability to be improved using classical selective breeding, we anticipate that cost considerations might dictate that MAS will be used to develop resource lines (e.g., disease-resistant lines) as opposed to general production lines. These resource lines might be crossed into production stocks as needed to improve targeted traits for which QTL detection and MAS are cost-effective.

Acknowledgments

Andrey Shirak's research was supported by Research Grant No. IS-3561-04 from BARD, the United States-Israel Binational Agricultural Research and Development Program. Avner Cnaani was supported by a Vaadia-BARD postdoctoral fellowship award (FI-338-2003). Eric Hallerman's research in this area is supported, in part, by the U.S. Department of Agriculture CSREES Hatch Program. We thank Micha Ron and Gidi Hulata (Israel Agricultural Research Organization) for agreeing to use of the tilapia data.

References

Agresti JJ, S Seki, A Cnaani, S Poompuang, EM Hallerman, N Umiel, G Hulata, GAE Gall, and B May. 2000. Breeding new strains of tilapia: development of an artificial center of origin and linkage map based on AFLP and microsatellite loci. Aquaculture, 185, pp. 43–56.

Allendorf FW, JE Seeb, KL Knudsen, GH Thorgaard, and RF Leary. 1986. Gene-centromere mapping of 25 loci in rainbow trout. Journal of Heredity, 77, pp. 307–312.

Anderson EG and MM Rhoades. 1931. The distribution of interference in the X-chromosomes of Drosophila. Papers of the Michigan Academy of Science, 13, pp. 227–239.

Aparicio S, J Chapman, E Stupka, N Putnam, J Chia, P Dehal, A Christoffels, S Rash, S Hoon, A Smit, MDS Gelpke, J Roach, T Oh, IY Ho, M Wong, C Detter, F Verhoef, P Predki, A Tay, S Lucas, P Richardson, SF Smith, MS Clark, YJK Edwards, N Doggett, A Zharkikh, SV Tavtigian, D Pruss, M Barnstead, C Evans, H Baden, J Powell, G Glusman, L Rowen, L Hood, YH Tan, G Elgar, T Hawkins, B Venkatesh, D Rokhsar, and S Brenner. 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science, 297, pp. 1301–1310.

Araneda C, R Neira, and P Iturra. 2005. Identification of a dominant SCAR marker associated with colour traits in coho salmon (Oncorhyncus kisutch). Aquaculture, 247, pp. 67–73.

Artieri CG, LA Mitchell, SH Ng, SE Parisotto, RG Danzmann, B Hoyheim, RB Phillips, M Morasch, BF Koop, and WS Davidson. 2006. Identification of the sex-determining locus of Atlantic salmon (Salmo salar) on chromosome 2. Cytogenetics and Genome Research, 112, pp. 152–159.

Baden J, G Powell, L Glusman, L Rowen, L Hood, YH Tan, G Elgar, T Hawkins, B Venkatesh, D Rokhsar, and S Brenner. 2002. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science, 297, pp. 1301–1310.

Barratt RW, D Newmeyer, DD Perkins, and L Garnjobst. 1954. Map construction in Neurospora crassa. Advances in Genetics, 6, pp. 1–93.

Chourrout D. 1984. Pressure-induced retention of second polar body and suppression of first cleavage in rainbow trout: production of all-triploids, all-tetraploids, and heterozygous and homozygous diploid gynogenetics. Aquaculture, 36, pp. 111–126.

Churchill GA and RW Doerge. 1994. Empirical threshold values for quantitative trait mapping. Genetics, 138, pp. 963–971.

Clark MS. 2003. Genomics and mapping of Teleostei (bony fish). Comparative and Functional Genomics, 4, pp. 182–193.

Cnaani A, EM Hallerman, M Ron, JI Weller, M Indelman, Y Kashi, GAE Gall, and G Hulata. 2003. Detection of a chromosomal region with two quantitative trait loci, affecting cold tolerance and fish size, in an F_2 tilapia hybrid. Aquaculture, 223, pp. 117–128.

Cnaani A, S Tinman, Y Avidar, M Ron, and G Hulata. 2004a. Comparative study of biochemical parameters in response to stress in Oreochromis aureus, O. mossambicus and two strains of O. niloticus. Aquaculture Research, 35, pp. 1434–1440.

Cnaani A, N Zilberman, S Tinman, G Hulata, and M Ron. 2004b. Genome-scan analysis for quantitative trait loci in an F_2 tilapia hybrid. Molecular Genetics and Genomics, 272, pp. 162–172.

Coimbra MRM, K Kobayashi, S Koretsugu, O Hasegawa, E Ohara, A Ozaki, T Sakamoto, K Naruse, and N Okamoto. 2003. A genetic linkage map of the Japanese flounder, Paralichthys olivaceus. Aquaculture, 220, pp. 203–218.

Collard BCY, MZZ Jahufer, JB Brouwer, and ECK Pang. 2005. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: the basic concepts. Euphytica, 142, pp. 169–196.

Crollius HR, O Jaillon, C Dasilva, C Ozouf-Costaz, C Fizames, C Fischer, L Bouneau, A Billault, F Quetier, W Saurin, A Bernot, and J Weissenbach. 2000. Characterization and repeat analysis of the compact genome of the freshwater pufferfish Tetraodon nigroviridis. Genome Research, 10, pp. 939–949.

Danzmann RG, M Cairney, WS Davidson, MM Ferguson, K Gharbi, R Guyomard, E Leder, N Okamoto, A Ozaki, CE Rexroad, T Sakamoto, JB Taggart, and RA Woram. 2005. A comparative analysis of the rainbow trout genome with two other species of fish (Arctic charr and Atlantic salmon) within the tetraploid derivative Salmonidae family (subfamily Salmoninae). Genome, 48, pp. 1037–1051.

Danzmann RG and K Gharbi. 2001. Gene mapping in fishes: a means to an end. Genetica, 111, pp. 3–23.

Danzmann RG, TR Jackson, and MM Ferguson. 1999. Epistasis in allelic expression at upper temperature tolerance QTL in rainbow trout. Aquaculture, 173, pp. 45–58.

Darvasi A. 1998. Experimental strategies for the genetic dissection of complex traits in animal models. Nature Genetics, 18, pp. 19–24.

DEFRA (Department for Environment, Food and Rural Affairs, UK). 2006. National scrapie plan.http://www.defra.gov.uk/animalh/bse/othertses/scrapie/nsp/index.html. Accessed January 19, 2006.

Don J and RR Avtalion. 1988. Production of F_1 and F_2 diploid gynogenetic tilapias and analysis of the "Hertwig curve" obtained using ultraviolet irradiated sperm. Theoretical and Applied Genetics, 76, pp. 253–259.

FAO (Food and Agriculture Organization of the United Nations). 2006. Fisheries global information system. http://www.fao.org/figis. Accessed May 2, 2006.

Ferguson MM and RG Danzmann. 1998. Role of genetic markers in fisheries and aquaculture: useful tools or stamp collecting? Canadian Journal of Fisheries and Aquatic Sciences, 55, pp. 1553–1563.

Gjedrem T. 1983. Genetic variation in quantitative traits and selective breeding in fish and shellfish. Aquaculture, 33, pp. 51–72.

Gjedrem T. 1985. Improvement of productivity through breeding schemes. Geojournal, 10, pp. 233–241.

Grimholt U, S Larsen, R Nordmo, P Midtlyng, S Kjoeglum, A Storset, S Saebo, and RJ Stet. 2003. MHC polymorphism and disease resistance in Atlantic salmon (Salmo salar): facing pathogens with single expressed major histocompatibility class I and class II loci. Immunogenetics, 55, pp. 210–219.

Haley CS and SA Knott. 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity, 69, pp. 315–324.

Hoeschele I. 1993. Incorporation of quantitative trait loci equations into an animal model incorporating genetic marker data. Journal of Dairy Science, 76, pp. 1693–1701.

Hoyheim B, R Danzmann, R Guyomard, L Holm, R Powell, and J Taggart. 1998. Constructing a genetic map of salmonid fishes: Salmap. Plant and Animal Genome 6, p. 165 (abstract).

Jackson TR, MM Ferguson, RG Danzmann, AG Fishback, PE Ihssen, M O'Connell, and TJ Crease. 1998. Identification of two QTL influencing upper thermal tolerance in three rainbow trout (Oncorhynchus mykiss) half-sib families. Heredity, 80, pp. 143–151.

Jansen RC, JM Van Ooijen, P Stam, C Lister, and C Dean. 1995. Genotype-by-environment interaction in genetic mapping of multiple quantitative trait loci. Theoretical and Applied Genetics, 91, pp. 33–37.

Johnson KR, JE Wright Jr, and B May. 1987. Linkage relationships reflecting ancestral tetraploidy in salmonid fish. Genetics, 116, pp. 579–591.

Kashi Y, E Hallerman, and M Soller. 1990. Marker-assisted selection of candidate bulls for progeny testing programmes. Animal Production, 51, pp. 63–74.

Khoo SK, A Ozaki, F Nakamura, T Arakawa, S Ishimoto, R Nikolov, T Sakamoto, T Akatsu, M Mochuzuki, I Denda, and N Okamoto. 2004. Identification of a novel chromosomal region associated with infectious hematopoietic necrosis (IHN) resistance in rainbow trout Oncorhynchus mykiss. Fish Pathology, 39, pp. 95–101.

Kocher TD, WJ Lee, H Sobolewska, D Penman, and B McAndrew. 1998. A genetic linkage map of a cichlid fish, the tilapia (Oreochromis niloticus). Genetics, 148, pp. 1225–1232.

Korol AB. 2001. Recombination. In: Encyclopedia of Biodiversity, vol. 5. Academic Press, San Diego, pp. 53–71.

Korol AB, IA Preygel, and SI Preygel. 1994. Recombination Variability and Evolution. Chapman and Hall, London.

Korol AB, Y Ronin, V Kirzhner. 1995. Interval mapping of quantitative trait loci employing correlated trait complexes. Genetics, 140, pp. 1137–1147.

Korol AB, YI Ronin, and E Nevo. 1998. Approximated analysis of QTL-environmental interaction with no limits on the number of environments. Genetics, 148, pp. 1015–1028.

Korol AB, YI Ronin, AM Itskovich, J Peng, and E Nevo. 2001. Enhanced efficiency of QTL mapping analysis based on multivariate complexes of quantitative traits. Genetics, 157, pp. 1789–1803.

Kristensen VN, D Kelefiotis, T Kristensen, and AL Borresen-Dale. 2001. High-throughput methods for detection of genetic variation. Biotechniques, 30, pp. 318–326.

Lande R and R Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics, 124, pp. 743–756.

Lander E and L Kruglyak. 1995. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nature Genetics, 11, pp. 241–247.

Lander ES and D Botstein. 1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics, 121, pp. 185–199.

Lee B-Y, G Hulata, and TD Kocher. 2004. Two unlinked loci control the sex of blue tilapia (Oreochromis aureus). Heredity, 92, pp. 543–549.

Lee B-Y, WJ Lee, JT Streelman, KL Carleton, AE Howe, G Hulata, A Slettan, JE Stern, Y Terai, and TD Kocher. 2005. A second generation genetic linkage map of tilapia (Oreochromis spp.). Genetics, 170, pp. 237–244.

Lee B-Y, DJ Penman, and TD Kocher. 2003. Identification of the sex-determining region in tilapia (Oreochromis niloticus) using bulked segregant analysis. Animal Genetics, 34, pp. 379–383.

Lindner KR, JE Seeb, C Habicht, KL Knudsen, E Kretschmer, DJ Reedy, P Spruell, and FW Allendorf. 2000. Gene-centromere mapping of 312 loci in pink salmon by half-tetrad analysis. Genome, 43, pp. 538–549.

Liu ZJ and JF Cordes. 2004. DNA marker technologies and their applications in aquaculture genetics. Aquaculture, 238, pp. 1–37.

Mackay TFC. 2001. The genetic architecture of quantitative traits. Annual Review of Genetics, 35, pp. 303–339.

Mair GC. 1993 Chromosome-set manipulation in tilapia: techniques, problems and prospects. Aquaculture, 111, pp. 227–244.

Martins C, C Oliveira, AP Wasko, and JM Wright. 2004. Physical mapping of the Nile tilapia (Oreochromis niloticus) genome by fluorescent in situ hybridization of repetitive DNAs to metaphase chromosomes—A review. Aquaculture, 231, pp. 37–49.

Martyniuk CJ, GML Perry, HK Mogahadam, MM Ferguson, and RG Danzmann. 2003. The genetic architecture of correlations among growth-related traits and male age at maturation in rainbow trout. Journal of Fish Biology, 63, pp. 746–764.

McConnell SKJ, C Beynon, J Leamon, and DOF Skibinski. 2000. Microsatellite marker based genetic linkage maps of Oreochromis aureus and O. niloticus (Cichlidae): extensive linkage group segment homologies revealed. Animal Genetics, 31, pp. 214–218.

Mester D, YI Ronin, D Minkov, E Nevo, and AB Korol. 2003b. Constructing large scale genetic maps using evolutionary strategy algorithm. Genetics, 165, pp. 2269–2282.

Mester DI, YI Ronin, Y Hu, E Nevo, and AB Korol. 2003a. Efficient multipoint mapping: making use of dominant repulsion-phase markers. Theoretical and Applied Genetics, 107, pp. 1102–1112.

Moen T, JJ Agresti, A Cnaani, H Moses, TR Famula, G Hulata, GAE Gall, and B May. 2004a. A genome scan of a four-way tilapia cross supports the existence of a quantitative trait locus for cold tolerance on linkage group 23. Aquaculture Research, 35, pp. 893–904.

Moen T, KT Fjalestad, H Munck, and L Gomez-Raya. 2004b. A multistage testing strategy for detection of quantitative trait loci affecting disease resistance in Atlantic salmon. Genetics, 167, pp. 851–858.

Monaco PJ, EM Rasch, and JS Balsano. 1984. In: BJ Turner, Ed. Evolutionary Genetics of Fishes. Plenum Press, New York. pp. 311–328.

Myers JM, DJ Penman, Y Basavaraju, SF Powell, P Baoprasertukul, KJ Rana, N Bromage, and BJ McAndrew. 1995. Induction of diploid androgenetic and mitotic gynogenetic Nile tilapia (Oreochromis niloticus L.). Theoretical and Applied Genetics, 90, pp. 205–210.

Nagy A and V Csanyi. 1982. Changes of genetic parameters in successive gynogenetic generations. Theoretical and Applied Genetics, 63, pp. 105–110.

Nagy A, K Rajki, L Horvath, and V Csanyi. 1979. Genetic analysis in carp (Cyprinus carpio) using gynogenesis. Heredity, 43, pp. 35–40.

Nakamura K, A Ozaki, T Akutsu, K Iwai, T Sakamoto, G Toshizaki, and N Okamoto. 2001. Genetic mapping of the dominant albino locus in rainbow trout (Oncorhynchus mykiss). Molecular Genetics and Genomics, 265, pp. 687–693.

Naruse K, S Fukamachi, H Mitani, M Kondo, T Matsuoka, S Kondo, N Hanamura, Y Morita, K Hasegawa, R Nishigaki, A Shimada, H Wada, T Kusakabe, N Suzuki, M Kinoshita, A Kanamori, T Terado, H Kimura, M Nonaka, and A Shima. 2000. A detailed linkage map of medaka, Oryzias latipes: comparative genomics and genome evolution. Genetics, 154, pp. 1773–1784.

Ng SHS, CG Artieri, IE Bosdet, R Chiu, RG Danzmann, WS Davidson, MM Ferguson, CD Fjell, B Hoyheim, SJM Jones, PJ de Jong, BF Koop, MI Krzywinski, K Lubieniecki, MA Marra, LA Mitchell, C Mathewson, K Osoegawa, SE Parisotto, RB Phillips, ML Rise, KR von Schalburg, JE Schein, H Shin, A Siddiqui, J Thorsen, N Wye, G Yang, and B Zhu. 2005. A physical map of the genome of Atlantic salmon, Salmo salar. Genomics, 86, pp. 396–404.

Nichols KM, J Bartholomew, and GH Thorgaard. 2003a. Mapping multiple genetic loci associated with Ceratomyxa shasta resistance in Oncorhynchus mykiss. Diseases of Aquatic Organisms, 56, pp. 145–154.

Nichols KM, PA Wheeler, and GH Thorgaard. 2004. Quantitative trait loci analyses for meristic traits in Oncorhynchus mykiss. Environmental Biology of Fishes, 69, pp. 317–331.

Nichols KM, WP Young, RG Danzmann, BD Robison, C Rexroad, M Noakes, RB Phillips, P Bentzen, I Spies, K Knudsen, FW Allendorf, BM Cunningham, J Brunelli, H Zhang, S Ristow, R Drew, KH Brown, PA Wheeler, and GH Thorgaard. 2003b. A consolidated linkage map for rainbow trout (Oncorhynchus mykiss). Animal Genetics, 34, pp. 102–115.

O'Malley KG, T Sakamoto, RG Danzmann, and MM Ferguson. 2003. Quantitative trait loci for spawning date and body weight in rainbow trout: testing for conserved effects across ancestrally duplicated genomes. Journal of Heredity, 94, pp. 273–284.

Ozaki A, T Sakamoto, S Khoo, K Nakamura, MRM Coimbra, T Akutsu, and N Okamoto. 2001. Quantitative trait loci (QTLs) associated with resistance/susceptibility to infectious pancreatic necrosis virus (IPNV) in rainbow trout (Oncorhynchus mykiss). Molecular Genetics and Genomics, 265, pp. 23–31.

Palti Y, SA Gahr, JD Hansen, and CE Rexroad III. 2004. Characterization of a new BAC library for rainbow trout: evidence for multi-locus duplication. Animal Genetics, 35, pp. 130–133.

Palti Y, KM Nichols, KI Waller, JE Parsons, and GH Thorgaard. 2001. Association between DNA polymorphisms tightly linked to MHC class II genes and IHN virus resistance in backcrosses of rainbow trout and cutthroat trout. Aquaculture, 194, pp. 283–289.

Palti Y, JE Parsons, and GH Thorgaard. 1999. Identification of candidate DNA markers associated with IHN virus resistance in backcrosses of rainbow (Oncorhynchus mykiss) and cutthroat trout (O. clarki). Aquaculture, 173, pp. 81–94.

Palti Y, A Shirak, A Cnaani, G Hulata, RR Avtalion, and M Ron. 2002. Detection of genes with deleterious alleles in an inbred line of tilapia (Oreochromis aureus). Aquaculture, 206, pp. 151–164.

Parsons JE and GH Thorgaard. 1984. Induced androgenesis in rainbow trout. Journal of Experimental Zoology, 231, pp. 407–412.

Payne F. 1918. An experiment to test the nature of the variation on which selection acts. Indiana University Studies, 5, pp. 1–45.

Peng JH, AB Korol, T Fahima, MS Röder, YI Ronin, YC Li, and E Nevo. 2000. Molecular genetic maps in wild emmer wheat, Triticum dicoccoides: genome-wide coverage, massive negative interference, and putative quasi-linkage. Genome Research, 10, pp. 1509–1531.

Perry GML, RG Danzmann, MM Ferguson, and JP Gibson. 2001. Quantitative trait loci for upper thermal tolerance in outbred strains of rainbow trout (Oncorhynchus mykiss). Heredity, 86, pp. 333–341.

Phillips RB, MR Morasch, LK Park, KA Naish, and RH Devlin. 2005. Identification of the sex chromosome pair in coho salmon (Oncorhynchus kisutch): lack of conservation of the sex linkage group with chinook salmon (Oncorhynchus tshawytscha). Cytogenetics and Genome Research, 111, pp. 166–70.

Poompuang S and EM Hallerman. 1997. Toward detection of quantitative trait loci and marker-assisted selection in fish. Reviews in Fisheries Science, 5, pp. 253–277.

Postlethwait JH, SL Johnson, CN Midson, WS Talbot, M Gates, EW Ballinger, D Africa, R Andrews, T Carl, JS Eisen, S Horne, CB Kimmel, M Hutchinson, M Johnson, and A Rodriguez. 1994. A genetic-linkage map for the zebrafish. Science, 264, pp. 699–703.

Purdom CE. 1969. Radiation-induced gynogenesis and androgenesis in fish. Heredity, 24, pp. 431–444.

Reid DP, A Szanto, B Glebe, RG Danzmann, and MM Ferguson. 2005. A QTL for body weight and condition factor in Atlantic salmon (Salmo salar): Comparative analysis with rainbow trout (Oncorhynchus mykiss) and Arctic charr (Salvelinus alpinus). Heredity, 94, pp. 166–172.

Robison BD, PA Wheller, K Sundin, P Sikka, and GH Thorgaard. 2001. Composite interval mapping reveals a major locus influencing embryonic development rate in rainbow trout (Oncorhynchus mykiss). Journal of Heredity, 92, pp. 16–22.

Rodriguez MF, S LaPatra, S Williams, T Famula, and B May. 2005. Genetic markers associated with resistance to infectious hematopoietic necrosis in rainbow and steelhead trout (Oncorhynchus mykiss) backcrosses. Aquaculture, 241, pp. 93–115.

Ronin Y, AB Korol, M Shtemberg, E Nevo, and M Soller. 2003. High resolution mapping of quantitative trait loci by selective recombinant genotyping. Genetics, 164, pp. 1657–1666.

Rothschild MF, C Jacobson, DA Vaske, CK Tuggle, L Wang, TH Short, GR Eckhardt, S Sasaki, A Vincent, DG McLaren, O Southwood, H Van der Steen, A Mileham, and G Plastow. 1996. The estrogen receptor is associated with a major gene influencing little size in pigs. Proceedings of the National Academy of Sciences U.S.A., 93, pp. 201–205.

Sakamoto T, RG Danzmann, K Gharbi, P Howard, A Ozaki, SK Khoo, RA Woram, N Okamoto, MM Ferguson, LE Holm, R Guyomard, and B Hoyheim. 2000. A microsatellite linkage map of rainbow trout (Oncorhynchus mykiss) characterized by large sex-specific differences in recombination rates. Genetics, 155, pp. 1331–1345.

Sakamoto T, RG Danzmann, N Okamoto, MM Ferguson, and PE Ihssen. 1999. Linkage analysis of quantitative trait loci associated with spawning time in rainbow trout (Oncorhynchus mykiss). Aquaculture, 173, pp. 33–43.

Sax K. 1923.The association of size differences with seed-coat pattern and pigmentation in Phaseeolus vulgaris. Genetics, 8, pp. 552–560.

Serraj R, CT Hash, SMH Rizvi, A Sharma, RS Yadav, and FR Bidinger. 2005. Recent advances in marker-assisted selection for drought tolerance in millet. Plant Production Science, 8, pp. 334–337.

Shimoda N, EW Knapik, J Ziniti, C Sim, E Yamada, S Kaplan, D Jackson, F de Sauvage, H Jacob, and MC Fishman. 1999. Zebra fish genetic map with 2000 microsatellite markers. Genomics, 58, pp. 219–232.

Shirak A, A Bendersky, G Hulata, M Ron, and RR Avtalion. 2006. Altered self-erythrocyte recognition and destruction in an inbred line of tilapia (Oreochromis aureus). Journal of Immunology, 176, pp. 390–394.

Shirak A, Y Palti, A Cnaani, AB Korol, G Hulata, M Ron, and RR Avtalion. 2002. Association between loci with deleterious alleles and distorted sex ratios in an inbred line of tilapia (Oreochromus aureus). Journal of Heredity, 93, pp. 270–276.

Shirak A, E Seroussi, A Cnaani, AE Howe, R Domokhovsky, N Zilberman, TD Kocher, G Hulata, and M Ron. 2006. Amh and Dmrta2 genes map to tilapia (Oreochromis spp.) linkage group 23 within QTL regions for sex determination. Genetics, 174, pp. 1573–1581.

Shirak A, A Shmarina, Y Kuperman, and RR Avtalion. 2000. Inheritance of body and peritoneum color in hybrids of Oreochromis aureus and red O. niloticus. Israel Journal of Aquaculture—Bamidgeh, 52, pp. 21–29.

Shirak A, J Vartin, and RR Avtalion. 1998. Production of viable diploid mitogynogenetic Oreochromis aureus using the cold shock technique and its optimization through definition of cleavage time. Israel Journal of Aquaculture—Bamidgeh, 50, pp. 140–150.

Sivagnanasundaram S, K Broman, M Liu, and A Petronis. 2004. Quasi-linkage: a confounding factor in linkage analysis of complex diseases? Human Genetics, 114, pp. 588–593.

Soller M, T Brody, and A Genizi. 1976. Power of experimental designs for detection of linkage between marker loci and quantitative loci in crosses between inbred lines. Theoretical and Applied Genetics, 47, pp. 35–39.

Somorjai IML, RG Danzmann, and MM Ferguson. 2003. Distribution of temperature tolerance quantitative trait loci in Arctic charr (Salvelinus alpinus) and inferred homologies in rainbow trout (Oncorhynchus mykiss). Genetics, 165, pp. 1443–1456.

Spelman RJ and DJ Garrick. 1998. Genetic and economic responses for within-family marker-assisted selection in dairy cattle breeding programs. Journal of Dairy Science, 81, pp. 2942–2950.

Stanley JG. 1976. Production of hybrid androgenetic and gynogenetic grass carp and carp. Transactions of the American Fisheries Society, 105, pp. 10–16.

Streelman JT and TD Kocher. 2002. Microsatellite variation associated with prolactin expression and growth of salt-challenged tilapia. Physiological Genomics, 9, pp. 1–4.

Streisinger G, C Walker, N Dower, D Knauber, and F Singer. 1981. Production of clones of homozygous diploid zebrafish (Brachydanio rerio). Nature, 291, pp. 293–296.

Sun X and L Liang. 2004. A genetic linkage map of common carp (*Cyprinus carpio L.*) and mapping of a locus associated with cold tolerance. Aquaculture, 238, pp. 165–172.

Tabor HK, NJ Risch, and RM Myers. 2002. Candidate-gene approaches for studying complex genetic traits: practical considerations. Nature Reviews Genetics, 3, pp. 391–396.

Tao WJ and EG Boulding. 2003. Associations between single nucleotide polymorphisms in candidate genes and growth rate in Arctic charr (Salvelinus alpinus L.). Heredity, 91, pp. 60–69.

Thorgaard GH, FW Allendorf, and KL Knudsen. 1983. Gene-centromere mapping in rainbow trout: high interference over long map distances. Genetics, 103, pp. 771–783.

Visscher P and CS Haley. 1998. Marker-assisted selection in commercial pig breeding programs, pp. 57–76 In Wiseman J, Varley MA, Chadwick JP, Eds. Progress in Pig Science. Nottingham University Press, Nottingham, UK.

Waldbieser GC, BG Bosworth, DJ Nonneman, and WR Wolters. 2001. A microsatellite-based genetic linkage map for channel catfish, Ictalurus punctatus. Genetics, 158, pp. 727–734.

Weller JI. 1997. Introduction to QTL detection and marker-assisted selection, pp. 259–275 In Miller RH, Pursel VG, Norman HD, Eds. Beltsville Symposia in Agricultural Research 20: Biotechnology's Role in the Genetic Improvement of Farm Animals. American Society for Animal Science, Savoy, IL.

Weller JI, JZ Song, DW Heyen, HA Lewin, and M Ron. 1998. A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. Genetics, 150, pp. 1699–1706.

Woods IG, PD Kelly, F Chu, P Ngo-Hazelett, YL Yan, H Huang, JH Postlethwait, and WS Talbot. 2000. A comparative map of the zebrafish genome. Genome Research, 10, pp. 1903–1914.

Woram RA, K Gharbi, T Sakamoto, B Hoyheim, L-E Holm, K Naish, C McGowan, MM Ferguson, RB Phillips, J Stein, R Guyomard, M Cairney, JB Taggart, R Powell, W Davidson,

and RG Danzmann. 2003. Comparative genome analysis of the primary sex-determining locus in salmonid fishes. Genome Research, 13, pp. 272–280.

Yagil C, M Sapojnikov, A Wechsler, AB Korol, and Y Yagil. 2006. Genetic dissection of proteinuria in the Sabra rat. Physiological Genomics, 25, pp. 121–133.

Young WP, PA Wheeler, VH Coryell, P Keim, and GH Thorgaard. 1998. A detailed linkage map of rainbow trout produced using doubled haploids. Genetics, 148, pp. 839–850.

Yousef GG and JA Juvik. 2002. Enhancement of seedling emergence in sweet corn by marker-assisted backcrossing of beneficial QTL. Crop Science, 42, pp. 96–104.

Zimmerman AM, JP Evenhuis, GH Thorgaard, and S Ristow. 2004. A single major chromosomal region controls natural killer cell-like activity in rainbow trout. Immunogenetics, 55, pp. 825–835.

Zimmerman AM, PA Wheeler, SS Ristow, and GH Thorgaard. 2005. Composite interval mapping reveals three QTL associated with pyloric caeca number in rainbow trout, *Oncorhynchus mykiss*. Aquaculture, 247, pp. 85–95.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 12
Marker-Assisted Selection for Aquaculture Species

Max F. Rothschild and Anatoly Ruvinsky

Introduction

Genetic improvement in fish and other aquaculture species is a relatively new development. Excellent reviews on the progress in many species have been published (e.g., Gjedrem 2000, Benzie 1998, Hulata 2001). While the earliest genetic improvement has used standard methods of selection and cross breeding, developments in molecular genetics have now allowed for the progression of molecular markers for parentage control and species identification (e.g., Heath et al. 1995, Norris et al. 2000, Martin and Soleto 2003). Also see Chapter 8 of this book. A number of mapping projects (Alcivar-Warren et al. 1997) have also led to the development of quantitative trait loci (QTL) projects designed to identify regions or genes and markers associated with specific traits in aquaculture species. Once identified these markers could then be used in marker-assisted selection (MAS). Hulata (2001) presented a review of developments that took place prior to the beginning of this decade.

The purpose of this chapter is to provide the reader with some background and theory related to MAS and information as it pertains to some other species essential for food production and to present information on recent progress in the field of aquaculture.

Definition and Theory of Marker-assisted Selection (MAS)

Traditional animal improvement has relied primarily on population selection based on phenotypic characters or traits. The idea of using genetic markers for selection, both by themselves and with other phenotypic data, termed marker-assisted selection (MAS) was first considered several decades ago when the first molecular markers, restriction fragment length polymorphisms (RFLP) were described in commercially important species (Soller and Beckmann 1982). Since then tremendous progress has been achieved in developing new types of markers, gene mapping, quantitative trait loci (QTL) studies, and in theoretical investigations of potential MAS outcomes. MAS can be best used for traits with low heritability, as a means to improve accuracy of selection, to reduce generation interval by early selection before maturity, and finally to select for traits that are observed in one sex only (Lande and Thompson 1990).

Despite a great deal of promise only a handful of cases demonstrating practical usefulness of MAS in reducing frequencies of recessive alleles causing genetic diseases, determining simple Mendelian traits and improvement in a few species have been reported so far (Dentine 1999, Dekkers 2004). In livestock, commercial

implementation of MAS related to improvement of quantitative traits has been employed for removal of deleterious major genes, growth rate, meat quality, disease resistance, and reproductive traits in pigs and in other species such as cattle where markers are used routinely for improvement of protein percentage in milk and marbling and tenderness in beef cattle. Dekkers (2004) has summarized the progress and use of markers as "the current attitude toward MAS is . . . cautious optimism." Before discussing the state of progress in aquaculture species, the basic principles and the major limitations of MAS will be briefly considered.

Quantitative traits are usually determined by a number of genes, some of them interacting, and influenced by environmental conditions. As a result, when heritability is low, effectiveness of selection based on phenotypic measurements alone may be poor (Lynch and Walsh 1998). However, the development of an approach allowing direct identification and subsequent selection of the most valuable genotypes is an attractive proposition. Such methods could make selection of both quantitative and qualitative traits more efficient. Molecular markers, either the causative ones or those located within the gene or close to that gene of interest, would be the best potential tools for such genotype-oriented selection.

Finding either the causative genetic mutation or effective molecular markers has in general proved to be difficult and requires a high level of genome knowledge for a species. There are two major reasons behind this. First of all, location and identity of genes affecting essential traits is usually not known. Only for the most intensively studied agricultural species, like cattle, pigs, and chickens, has such information become partially available, as genomic projects still are expensive and demanding. Hopefully the situation will steadily improve in the years to come.

The second obstacle, when the marker is not the causative mutation, is the ongoing recombination process, which constantly changes phase between the allele in question and a marker (Figure 12.1). Three possible approaches can reduce the intensity of this problem:

- Finding a marker within a gene which is in strong linkage disequilibrium (LD) with the causative mutation
- Choosing a marker tightly linked to the gene in question and in strong LD
- Flanking the gene by two closely linked markers

Each of these approaches has its own advantages and disadvantages.

Identification and sequencing of some genes affecting qualitative traits, like disease susceptibility, opens the door for searching molecular markers, which could be causative mutations in the gene and which could be directly selected against the deleterious allele. This approach is also known as gene-assisted selection (GAS). Malignant

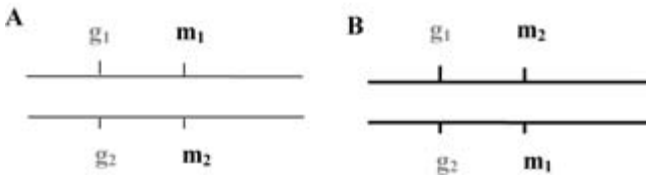


Figure 12.1. Coupling (A) and repulsion (B) phases of the preferred allele (g_1) and the marker (m_1).

hyperthermia syndrome (MHS) in pigs may serve as an example here (MacLennan et al. 1990). Commercial populations of pigs can and have been purged from highly undesirable allele causing MHS in a relatively short period of time. Unfortunately, the number of such examples is not large; similar situations are even much rarer among quantitative traits. Obviously GAS does not carry negative effects of recombination because a causative mutation is located within the gene in question and this is a clear advantage. In most of the selection schemes GAS outperforms MAS and particularly in the short-term (Villanueva et al. 2002). There is a strong expectation that in the near future significant progress will be achieved in identifying links between genes and phenotypic traits, a number of "phenome" projects are under way (Williams 2006).

An alternative approach is choosing closely located markers (<1 centiMorgans [cM]). This in turn requires dense linkage maps that are not available today except for chickens and cattle where draft genome sequences exist. Lack of such maps also adds strict limitations on QTL or association mapping for nearly all aquaculture species at least for some time. More distantly located single markers (>5 cM) are less reliable due to recombination events and could produce in the long run more harm than benefits. A third possibility, flanking markers, is more advantageous (Figure 12.2).

As the probability of double recombination between two linked markers (<5 cM) is low, there is a very small risk (~0.0025) to "lose" control over the preferred allele with a large selective value using only marker information. It is also essential that obtaining homozygotes for a haplotype carrying the preferred allele becomes much easier (Figure 12.3). After QTL in a species are verified, MAS programs become feasible (Evans et al. 2003).

The development of genome sequencing in aquaculture species will eventually lead to dense genetic maps and alleviate these problems. Thus, as soon as knowledge about genes or QTL affecting essential traits becomes available and haplotypes carrying preferred alleles are flanked by molecular markers or causative mutations are found within the genes responsible for particular traits, MAS can be implemented. Theoretical evaluation of potential effects from MAS implementation initially was quite optimistic, ranging from 9% to 64% of gain when compared with non-MAS strategy (Meuwissen and Goddard 1996). Studies and limited practice during the last decade has indicated that these estimates are likely the upper limits, which might be theoretically achieved under specific sets of circumstances favoring MAS. However, there are numerous factors limiting practical gains of MAS.

Requirements and Limits of MAS Implementation

Not all traits are equally suitable for MAS implementation. The basic economic requirements are clear in that profits from the introduction of MAS into breeding

Figure 12.2. A haplotype framed by two closely linked (<5 cM) markers like m_1 and m_2 is likely to carry the preferred allele (g), which can not be tested directly. "Loosing" the preferred allele from the haplotype is unlikely due to a low probability of double recombination events on m_1–m_2 intervals.

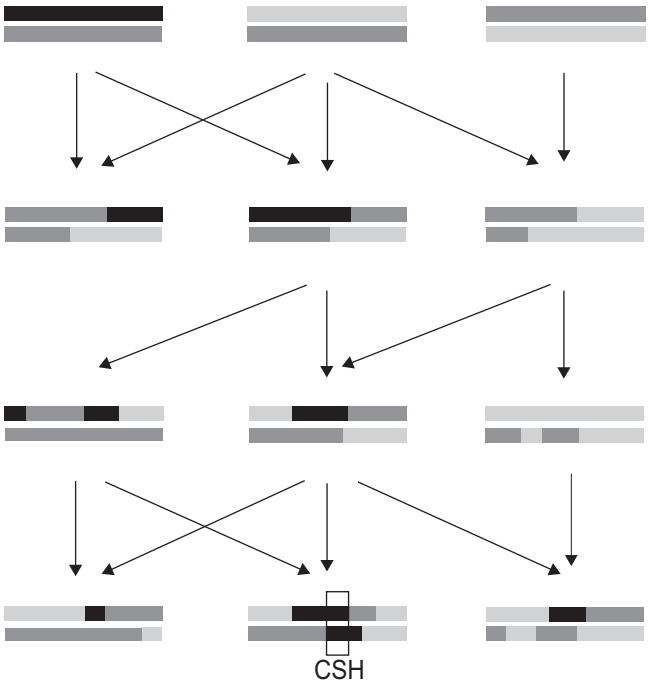


Figure 12.3. Transmission of haplotypes through generations. Homozygotization of a certain chromosome fragment is possible (CSH—*c*hromosome *s*egment *h*omozygosity). Usage of markers framing such a fragment increases opportunities for homozygotization, if it is beneficial for breeding purposes. Adapted from Hayes et al. 2005.

practices must exceed investments in the development of MAS technology. Assuming that this technology is based on relatively complex and expensive methods of sampling, DNA extraction, marker identification, and analysis on a mass scale, costs are initially quite high. Additional costs of MAS, while relatively small may also conflict with commercial use in an industry. There are several criteria affecting potential benefits of MAS, which are briefly considered below.

Timing of Trait Recording

In a variety of situations, traits are not known or cannot be recorded prior to the required selection decisions. Meat or fish quality could serve as a good example. Markers might be very valuable in improving genotypes and "hidden" phenotypes of sires and dams in relation to the meat or fish flesh quality. Another example is resistance or susceptibility of animals to a certain disease or parasite, which might occur only at the time of exposure. Here again MAS could be useful in promoting the most resistant genotypes. It should be mentioned that in some species (including numerous marine species) mass selection at the time of exposure could be an alternative option particularly when there is a major locus determining resistance. Notter and Cockett (2005) provided an additional illustration of potential usefulness relative to time of recording.

Genetic improvement in traits associated with seasonal breeding in sheep is difficult because these traits are generally not expressed until late in life and are usually recorded only in females. Detection of relevant QTL and their use in MAS could therefore substantially enhance selection response. The melatonin receptor 1a gene is polymorphic in many sheep breeds and appears to influence a number of seasonal reproductive responses. A variety of clock genes have been identified in laboratory mammals and have been shown to influence biological rhythms. Thus, the various clock genes represent potentially important candidate genes that may be involved in control of seasonal breeding.

Correlation Between Genotype and Phenotype

Those traits, where correlation between genotype and phenotype is high, are certainly not among the best candidates for MAS unless they cannot be measured directly on the animal as in sex-limited traits. If the trait can be assessed before the selection decision and phenotype is determined by genotype in a significant degree (heritability ~30–40% or above) there is no strong need for additional marker information. The selective objectives can be achieved by using traditional or modern methods of selection (Kinghorn 1997). However, in cases such as reproduction or disease resistance, where heritability is low, individual marker information can be very advantageous.

Age-related Selection

Young animals, which have yet to be involved in progeny-test schemes, present a better opportunity for MAS to be used than older individuals who might have a significant number of offspring with measured traits. As estimated breeding values (EBV) of such young individuals can not be predicted with a reasonable confidence, MAS using markers for the desired traits can be quite useful to preselect animals for further testing or to speed generation interval.

QTL or Major Gene is of Large Effect

Choosing an application of MAS, one has to give proper consideration to the size of the effect caused by a QTL or major gene. Clearly the preference should be given to those QTL, whose effect on the trait varying from moderate to large (~20–40% of the phenotypic variation). Individual smaller-sized QTL (<5%) are not expected to produce a significant benefit from MAS implementation. Stochastic computer simulations have shown that some extra gains are expected even when a trait is controlled by numerous loci of additive small effect distributed along many chromosomes and MAS is practiced (Villanueva et al. 2005). MAS has been compared with other schemes where genetic evaluations were performed using standard BLUP. When the density of markers was high enough, there has been an increase in the accuracy of selection with MAS, and this has led to extra gains (5–11%) when compared with standard BLUP.

However, the commercial viability of such MAS application is not clear since such studies have not considered the level of investment and maintenance costs and practical complications of molecular marker analysis.

Frequency of Preferred Allele or Haplotype

If the preferred allele is rather rare in a population, more gains are expected from MAS. But as is often the case, the "effect" may be poorly estimated. On the contrary, if there is a high frequency of the preferred allele, this does not leave the opportunity for significant improvement and MAS may not be economically viable. Computer modelling made by Schulman and Dentine (2005) shows that with a QTL of moderate size and initial allele frequencies of the favorable allele of 0.05, the response with MAS was 6% higher than with traditional selection in the sires selected after their progeny test.

Short-term Response to MAS

Once a desirable allele or haplotype is chosen and other preconditions are fulfilled, MAS can progress rapidly until fixation of the QTL or the major gene is achieved. It may take a few generations depending on intensity of the process to reach the high level of fixation desired in the population. Further gains in improving a particular trait using this marker become very limited and thus economically not viable. It simply means that in some cases MAS technology has a limited time, during which it is economically justifiable. Because the cost of development of such technology is usually not small it should be considered relative to the potential gains. This could be another impediment to MAS development and its practical implementation at least using currently available molecular technology.

Linkage Disequilibrium

It is quite possible that two or more loci affecting a phenotype are closely linked. In some instances, unfavorable haplotypes carrying alleles with the opposite effect on the breeding value might occur in a population. Such undesirable linkage disequilibrium can be better handled by MAS, which will facilitate identification of rare and desirable recombinant haplotypes. These recombinant haplotypes might present a new opportunity for selection and could be very beneficial. However, the initial investments in gene and QTL mapping, which are unavoidable, might be significant.

Marketing

For companies selling improved breeding stocks, active use of MAS can be viewed also as a marketing tool and may serve as evidence of the high genetic quality of their

product to some customers. GeneSTAR® is a very suitable example of such situations because there is no alternative way to assess potential high meat quality in young bulls (Genetic Solutions 2006). Buyers are willing to pay premium prices for animals with presumed superior qualities.

Long-term Response

Villanueva and others (2002) compared the benefits of MAS with schemes where selection is directly on the QTL (Gene Assisted Selection [GAS]) and with schemes where only phenotypic information was considered. The optimization of the additive genetic contributions (BLUP methodology) has a significant positive impact on genetic response but the use of markers leads to moderate additional short-term gains. Optimized selection schemes with phenotypic information only did nearly as well as standard truncation GAS in the short-term. The maximum accumulated benefit from MAS over conventional selection (BLUP) was less than half of the maximum benefit achieved from GAS, even with very low recombination rates between the markers and the QTL. The authors have shown that the prior information about the QTL effects can substantially increase genetic gain, and, when the accuracy of the priors is high enough, the responses from MAS are practically as high as those obtained with direct selection on the QTL.

A general conclusion can be drawn that MAS can increase the effectiveness of the selection process but this is not always the case (for easily measured traits with high heritability) and comes at a cost, which must be taken into consideration by organizations planning commercial implementation of MAS.

Parentage Testing, Species Identification, and Marker-assisted Introgression

Molecular markers can also be used for parentage testing and as a tool for marker-assisted introgression of desirable alleles and haplotypes (see Chapter 8). In both situations this may increase effectiveness and intensity of selection. Furthermore, markers can be used for species or population identification (see Chapter 9). This has considerable value because occasionally lower priced aquatic species are sold as higher priced ones and consumer fraud can be lessened with occasional marker testing.

Parentage Testing

Knowledge of parental-progeny relationships in a selected population could be very beneficial (see Chapter 8). Unfortunately, it cannot be perfectly known or recorded in some populations. Microsatellites and other molecular markers have proved to be an ample opportunity for a posterior parentage testing. Computer programs like CERVUS (http://helios.bto.ed.ac.uk/evolgen/cervus/cervus.html) provide powerful facilitation of parentage testing. It has been demonstrated that 12 MS loci with an

average number of alleles are quite sufficient to reconstruct parent-offspring relationships with a probability close to 1 (Sunduimijid 2006, personal communication). (Also see Chapter 8.) Using this approach, pedigree information can be obtained for any species, including marine organisms. For instance, an integrated pedigree database was developed at the Oceanic Institute, Hawaii, (http://www.oceanicinstitute.org/nav.php?loc=Research&page=Shrimp_Department) for a shrimp breeding program. This database provides the Oceanic Institute (OI) with a powerful tool for management of their shrimp-breeding program and also generates useful information about shrimp genetics. It is likely that similar parentage testing and performance databases will be developed for other species thus incorporating molecular markers in breeding programs.

Marker-assisted Introgression and Transgenic Technology

So far, the majority of examples of marker-assisted introgression have come from plant breeding due to the significant costs and time constraints for animal species. Very limited examples from animal species exist. At this stage it is difficult to predict whether or when a similar approach will be implemented in any aquaculture species-breeding program.

Transgenic technology on the contrary does not require too much time and in principle allows modification of many critically important traits like disease and parasite resistance, growth intensity, and quality of the final product. There are successful examples of transgenesis in many animal species including fish (Future Fish 2003, Tafalla et al. 2006) and crustaceans (Lu and Sun 2005). The public acceptance of transgenics is an entirely different matter, and MAS clearly would be preferred.

Species Identification

Authenticity of species, especially relative to labeling and marketing claims is commercially quite important. Several countries have legislation that requires accurate labeling. A number of these issues and methods using DNA markers have been considered and were reviewed at length by Martin and Soleto (2003).

MAS Developments in Aquaculture

Genetic Maps in Aquaculture Species

The essential condition for MAS is development of useful resource families, genetic linkage and physical maps, and large numbers of polymorphic genetic markers (Liu and Cordes 2004). In recent years, there has been clear progress in this direction. See Chapter 10. Genetic and physical maps of fish species include those for arctic char (Woram et al. 2004), salmon (Gilbey et al. 2004, Moen et al. 2004), rainbow trout (Young et al. 1998, Sakamoto et al. 2000, Nichols et al. 2003), gilthead seabream (Senger et al. 2006),

catfish (Liu et al. 2003), and tilapia (McConnell et al. 2000) for example. Other species include Pacific oyster (Hubert and Hedgecock 2004, Li and Guo 2004), eastern oyster (Yu and Guo 2003), abalone (Liu et al. 2006), black tiger shrimp (Wilson et al. 2002, 2004; Maneeruttanarungroj et al. 2006), and Pacific white shrimp (Alcivar-Warren et al. 2006). These maps vary in complexity with most consisting initially of microsatellites, amplified fragment length polymorphisms (AFLP), and more recently some single nucleotide polymorphisms (SNP). Fine mapping is being applied only to some species at this time.

QTL and Association Studies

Maps and markers are but the first step. Families in which crosses are produced and many traits measured are also required to obtain quantitative trait loci (QTL) information. See also Chapter 11. Just as mapping activity has increased, so have the efforts to find QTL and useful candidate genes.

A significant number of these QTL and association studies are connected with aquaculture species that have been farmed for a long time. For instance, in rainbow trout such QTL investigations were aimed for thermal tolerance (Jackson et al. 1998; Danzman et al. 1999; Perry et al. 2001, 2005), spawning time (Sakamoto et al. 1999, O'Malley et al. 2003), embryonic development (Robison et al. 2001), and more recently disease resistance. Use of markers to improve disease resistance offers considerable genetic advantage for aquaculture producers. Ozaki and others (2001) identified QTL associated with resistance to infectious pancreatic necrosis (IPN), which is a well-known acute viral disease in rainbow trout. Two putative QTL affecting disease resistances were detected on chromosomes A (IPN RS-1) and C (IPN RS-2). It was suggested that these markers have great potential for use in MAS for IPN resistance. More recently, Zimmerman and others (2004) found a chromosomal region controlling natural killer (NK) cell-like activity in rainbow trout using a genetic map of more than 500 markers including AFLPs and microsatellites. The single major QTL associated with NK-like activity is not linked to the fragmented MHC class I and MHC class II regions and the two QTL previously found to be associated with resistance to IPN virus in rainbow trout.

A large collaborative research group has been exploring QTL in tilapia, another farm-raised fish (Agresti et al. 2000). These investigations have led to discovery of QTL affecting a number of traits including body color (Howe and Kocher 2003) and sex determination (Shirak et al. 2002, Lee and Kocher 2003). More recently Cnaani and others (2004) initiated a search for QTL for innate immunity, response to stress, biochemical blood parameters, and body size in an F2 population derived from an interspecific tilapia hybrid. Despite the small family size and a limited number of markers, 35 significant marker-trait associations, involving 26 markers in 16 linkage groups, were found. Many of these were confirmed in a second experiment. The portion of variance explained by each QTL was on average 11%, with a maximum of 29%, and they found that the average additive effect of each QTL was 0.2 standard deviation units for stress response traits and fish size, with a maximum of 0.33. Some of these results confirm previous studies. More recently, Shirak and others (2006) reported QTL for other disease-related parameters.

Salmon, another extensively raised fish, have also been well investigated. Using 91 microsatellite loci and three full-sib families, Reid and others (2005) examined QTL affecting body weight and condition factor. A total of 10 suggestive and significant QTL for body weight and condition factor were identified. The largest QTL effects for body weight and for condition factor accounted for over 20% of the phenotypic trait variation, respectively. The authors noted that three of the QTL for body weight occur on linkage groups where similar effects have been detected on the homologous regions in either rainbow trout or arctic char.

Species investigated include arctic char and catfish. QTL have been reported in arctic char for temperature tolerance (Somorjai et al. 2003). In catfish, genetic markers have been associated with feed efficiency and performance traits (Karsi et al. 2000, Karsi and Waldbeiser 2005), and immune response (Karsi et al. 2005).

In shellfish, species QTL and association studies are considerably more limited and have for the most part been confined to shrimp and prawns. Glenn and others (2004) found SNPs in the cathepsin L gene and found suggestive associations with growth rate in both Pacific white shrimp and black tiger shrimp. Rocha and others (2005) working in the framework of the BioZEST ATP-research project, attempted to identify putative associations between DNA-markers and shrimp (*Litopenaeus vannamei*) production traits. Parents from 80 full-sib families were genotyped for batteries of SNP markers, and effects of these marker-genotypes on all production traits recorded were evaluated statistically. Two markers were found to be associated with statistically significant effects on an array of production traits, including harvest weight (HWT), test daily gain (TDG), biomass yield, grow-out survival, nursery, stocking and brood-stock weights, and several shrimp carcass and meat quality traits. Marker-effects were remarkably consistent across raceways and matched known phenotypic correlations among traits. Estimated average additive effects of allele substitutions for HWT (mean = 22.2 grams [g]) ranged from 2.2 to 4.5 g for marker 1, and from 1.5 to 1.8 g for marker 2. More recently, Yu and others (2006) examined three genes associated with molting but found no association with growth rate in Pacific white shrimp.

Applications and Limitations in Aquaculture

The growing numbers of markers and the development of useful maps and candidate genes and markers in some species have for the first time allowed commercial aquaculture ventures to consider using markers in selection programs. These results are difficult to find since most research and application in companies is well guarded.

Examples, however, can be found. These include the use of government labs to use DNA markers for identification of fish species as part of regulatory investigations. Furthermore, some companies are now using markers for parentage and selection decisions. One such example is Landcatch Natural Selection (http://www.swim-back.com/), which is a breeding company that is attempting to apply the latest methods in selective breeding technologies to its aquaculture operations worldwide. They use markers for traceability to parental stocks and some selection for trait improvement in salmon and advertise developments and expertise for markers in several other species. Another example is AKVAFORSK Genetic Center (http://www.afgc.no/), which has been

conducting genetic improvement with Atlantic salmon, rainbow trout, Atlantic cod, turbot, sea bass, and several other species. This center has also begun to employ DNA markers for genetic improvement. SyAqua, a shrimp genetic company, has also been working in the area of marker discovery with an eventual eye to using such markers for genetic improvement.

These company activities are in their infancy. Obvious limitations include poorly developed genetic maps and knowledge of the relationships between the markers and traits of economic importance. Also considerable investments are required to proceed with the use of MAS. These limitations have led to strategies like that employed initially by Sygen International (now Genus, Inc.), parent company of SyAqua, to perform marker discovery across many species. It is expected that financial investment will be the major limiting factor even after university and government scientists develop good maps and markers.

Conclusion

The developments of modern aquaculture and farm-raised aquatic species have necessitated the use of all currently available means for genetic improvement. Originally this included only selection and in some cases crossing of strains to produce better products. With the advent of molecular biology and the development of new genetic markers and genetic maps, scientists began to identify QTL regions and genes associated with traits of importance to the many aquaculture species. More recently, ESTs and initial sequencing projects, SNP discovery efforts, and development of linkage and physical maps have opened new opportunities for aquaculture genetics.

MAS is now beginning to be practiced in the industry. In certain instances, MAS has a clear advantage for genetic improvement, parentage control, and species identification. As the value of aquaculture products increase, it is expected that investment in MAS will also increase. The outcomes from such efforts are likely to be more efficiently grown products and increases in disease resistance due to identification of DNA markers associated with disease. It is anticipated that more companies will be employing MAS strategies to improve breeding stock and that its applications in aquaculture, now in its infancy, will grow rapidly.

References

Agresti JJ, S Seki, A Cnaani, S Poompuang, EM Hallerman, N Umiel, G Hulata, GAE Gall, and B May. 2000. Breeding new strains of tilapia: Development of an artificial center of origin and linkage map based on AFLP and microsatellite origin. Aquaculture, 185, pp. 43–56.

Alcivar-Warren A, RA Dunham, P Gaffney, TD Kocher, and GH Thorgaard. 1997. First aquaculture species genome mapping workshop. Anim Genetics, 28, pp. 451–452.

Alcivar-Warren A, D Meehan-Meola, Y Wang, X Guo, L Zhou, J Xiang, S Moss, S Arce, W Warren, Z Xu, and K Bell. 2006. Isolation and mapping of telomeric pentanucleotide (TAACC) (n) repeats of the Pacific Whiteleg shrimp, *Penaeus vannamei*, using fluorescence in situ hybridisation. Mar Biotechnology, 8, pp. 467–480.

Benzie JAH. 1998. Genetic improvement of prawns. Proc 6th World Cong Genet Appl Livest Prod, 27, pp. 103–110.

Cnaani A, N Zilberman, S Tinman, G Hulata, and M Ron. 2004. Genome-scan analysis for quantitative trait loci in an F2 tilapia hybrid. Mol Genet Genomics, 272, pp. 162–172.

Danzmann RG, TR Jackson, and M Ferguson. 1999. Epistasis in allelic expression at upper temperature tolerance QTL in rainbow trout. Aquaculture, 173, pp. 45–58.

Dekkers JCM. 2004. Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. J Anim Sci, 82(E Suppl), pp. E313–E328.

Dentine MR. 1999. Chapter 17: Marker-assisted selection. In: The genetics of cattle, R Fries, A Ruvinsky, Ed. pp. 497–510. CAB International, UK.

Evans GJ, E Giuffra, A Sanchez, S Kerje, G Davalos, O Vidal, S Illan, JL Noguera, L Varona, I Velander, OI Southwood, DJ de Koning, CS Haley, GS Plastow, and L Andersson. 2003. Identification of quantitative trait loci for production traits in commercial pig populations. Genetics, 164, pp. 621–627.

Future fish: Issues in science and regulation of transgenic fish. 2003. PEW Initiation on Food and Biotechnology, Washington DC. Accessed April 7, 2006, http://pewagbiotech.org/research/fish/fish.pdf.

Genetic Solutions. 2006. http://www.geneticsolutions.com.au/content/homeasp?name=GenSol_Home.

Gilbey J, E Verspoor, A McLay, and D Houlihan. 2004. A microsatellite linkage map for Atlantic salmon (*Salmo salar*). Anim Genet, 35, pp. 98–105.

Gjedrem T. 2000. Genetic improvement of cold water fish species. Aquacult Res, 31, pp. 25–33.

Glenn KL, L Grapes, T Suwanasopee, DL Harris, Y Li, K Wilson, and MF Rothschild. 2004. SNP analysis of *alpha-AMY* and *CTSL* genes in *Litopenaeus vannamei* and *Penaeus monodon* shrimp. Anim Genet, 36, pp. 235–236.

Hayes BJ, BP Kinghorn, and A Ruvinsky. 2005. Chapter 20: Genome scanning for quantitative trait loci. In: Mammalian Genomics, A Ruvinsky, JA Marshall Graves, Eds. pp. 507–537. CAB International, UK.

Heath DD, RH Devlin, TJ Hilbish, and GK Iwama. 1995. Multilocus DNA fingerprints in seven species of salmonids. Can J Zool, 73, pp. 600–606.

Howe A and T Kocher. 2003. Comparative mapping of QTL for red color in tilapia. PAG XI Abstracts, http://wwwintl-pagorg.

Hubert S and D Hedgecock. 2004. Linkage maps of microsatellite DNA markers for the Pacific oyster. Genetics, 168, pp. 351–62.

Hulata G. 2001. Genetic manipulations in aquaculture: A review of stock improvement by classical and modern technologies. Genetica, 111, pp. 155–173.

Jackson TR, MM Ferguson, RG Danzmann, AG Fishback, PE Ihssen, M O'Connell, and TJ Crease. 1998. Identification of two QTL influencing upper temperature tolerance in three rainbow trout (*Oncorhynnchus mykiss*) half sib families. Heredity, 80, pp. 143–151.

Karsi A, P Li, S Kim, R Dunham, and ZJ Liu. 2000. Performance traits linked DNA markers and marker assisted selection. PAG VIII Abstracts, http://wwwintl-pagorg.

Karsi A and GC Waldbeiser. 2005. Linkage mapping of the channel catfish proopiomelanocortin (POMC) gene. Anim Genet, 36, pp. 171–173.

Karsi A, WR Wolters, and GC Waldbeiser. 2005. Assignment of immune-related genes to the channel catfish, Ictalurus punctatus, genetic map. Anim Genet, 36, pp. 502–506.

Kinghorn BP. 1997. Genetic improvement of sheep. In: The Genetics of Sheep, L Piper, A Ruvinsky, Eds. pp. 565–591. CAB International, UK.

Lande R and R Thompson. 1990. Efficiency of marker-assisted selection in improvement of quantitative traits. Genetics, 124, pp. 743–756.

Lee B and T Kocher. 2003. Comparative mapping of sex determining genes in tilapia. PAG XI Abstracts, http://wwwintl-pagorg.

Li L and X Guo. 2004. AFLP based genetic linkage maps of the Pacific oyster *Crassostrea gigas Thunberg.* Mar Biotechnology, 6, pp. 26–36.

Liu X, X Liu, X Guo, Q Gao, H Zhao, and G Zhang. 2006. A preliminary genetic linkage map of the Pacific abalone, *Haliotis discus hannai Ino.* Mar Biotechnology, 8(4), pp. 386–397.

Liu Z, A Karsi, P Li, D Cao, and R Dunham. 2003. An AFLP-based genetic linkage map of channel catfish (*Ictalurus punctatus*) constructed by using an interspecific hybrid resource family. Genetics, 165, pp. 687–694.

Liu ZJ and JF Cordes. 2004. DNA marker technologies and their applications in aquaculture genetics. Aquaculture, 238, pp. 1–37.

Lu Y and PS Sun. 2005. Viral resistance in shrimp that express an antisense Taura syndrome virus coat protein gene. Antiviral Res, 67, pp. 141–146.

Lynch M and B Walsh. 1998. Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland, MA, USA.

MacLennan DH, C Duff, F Zorzato, J Fujii, M Phillips, RG Korneluk, W Frodis, BA Britt, and RG Worton. 1990. Ryanodine receptor gene is a candidate for predisposition to malignant hyperthermia. Nature, 343, pp. 559–561.

Maneeruttanarungroj C, S Pongsomboon, S Wuthisuthimethavee, S Klinbunga, KJ Wilson, J Swan, Y Li, V Whan, K-H Chu, CP Li, J Tong, K Glenn, M Rothschild, D Jerry, and A Tassanakajon. 2006. Development of polymorphic expressed sequence tag-derived microsatellites for the extension of the genetic linkage map of the black tiger shrimp (*Penaeus monodon*). Anim Genet, 37, pp. 363–368.

Marshall KL. 2002. Marker assisted selection in the Australian sheepmeat industry, University of New England, Armidale, Australia, 186 pp.

Martin RIP and CG Soleto. 2003. Authenticity of species in meat and seafood products. Proceedings Association International Congress on Authenticity of Species in Meat and Seafood Products, Spain, 281 pp.

McConnell SK, C Beynon, J Leamon, and DO Skibinski. 2000. Microsatellite marker based genetic linkage maps of *Oreochromis aureus* and *O niloticus (Cichlidae)*: Extensive linkage group segment homologies revealed. Anim Genet, 31, pp. 214–218.

Meuwissen THE and ME Goddard. 1996. The use of marker haplotypes in animal breeding schemes. Genet Sel Evol, 28, pp. 161–176.

Moen T, B Hoyheim, H Munck, and L Gomez-Raya. 2004. A linkage map of Atlantic salmon (*Salmo salar*) reveals an uncommonly large difference in recombination rate between the sexes. Anim Genet, 35, pp. 81–92.

Nichols KM, WP Young, RG Danzmann, BD Robison, C Rexroad, M Noakes, RB Phillips, P Bentzen, I Spies, K Knudsen, FW Allendorf, BM Cunningham, J Brunelli, H Zhang, S Ristow, R Drew, KH Brown, PA Wheeler, and GH Thorgaard. 2003. A consolidated linkage map for rainbow trout (*Oncorhynchus mykiss*). Anim Genet, 34, pp. 102–115.

Norris AT, DG Bradley, and EP Cunningham. 2000. Parentage and relatedness determination in farmed Atlantic salmon (*Salmo salar*) using microsatellite markers. Aquaculture, 182, pp. 73–83.

Notter DR and NE Cockett. 2005. Opportunities for detection and use of QTL influencing seasonal reproduction in sheep: A review. Genet Sel Evol, 37(Suppl 1), pp. S39–S53.

O'Malley KG, T Sakamoto, RG Danzmann, and MM Ferguson. 2003. Quantitative trait loci for spawning date and body weight in rainbow trout: Testing for conserved effects across ancestrally duplicated chromosomes. J Hered, 94, pp. 273–284.

Ozaki A, T Sakamoto, S Khoo, K Nakamura, MR Coimbra, T Akutsu, and N Okamoto. 2001. Quantitative trait loci (QTLs) associated with resistance/susceptibility to infectious pancreatic necrosis virus (IPNV) in rainbow trout (*Oncorhynchus mykiss*). Mol Genet Genomics, 265, pp. 23–31.

Perry GM, RG Danzmann, MM Ferguson, and JP Gibson. 2001. Quantitative trait loci for upper thermal tolerance in outbred strains of rainbow trout (*Oncorhynchus mykiss*). Heredity, 86, pp. 333–341.

Perry GM, MM Ferguson, T Sakamoto, and RG Danzmann. 2005. Sex-linked quantitative trait loci for thermotolerance and length in the rainbow trout. J Hered, 96, pp. 97–107.

Reid DP, A Szanto, B Glebe, RG Danzmann, and MM Fergueson. 2005. QTL for body weight and condition factor in Atlantic salmon (*Salmo salar*): Comparative analysis with rainbow trout (*Oncorhynchus mykiss*) and Arctic char (*Salvelinus alpinus*). Heredity, 94, pp. 166–172.

Robison BD, PA Wheeler, K Sundin, P Sikka, and GH Thorgaard. 2001. Composite interval mapping reveals a major locus influencing embryonic development rate in rainbow trout (*Oncorhynchus mykiss*). J Hered, 92, pp. 16–22.

Rocha JL, D Ciobanu, J Magrin, A Mileham, C Otoshi, S Moss, MF Rothschild, B Kinghorn, and H Van der Steen. 2005. Quantitative effects of DNA markers on shrimp growth. Aquaculture America, January 17–20, New Orleans, LA, USA, 69 pp.

Sakamoto T, RG Danzmann, K Gharbi, P Howard, A Ozaki, SK Khoo, PA Woram, N Okamoto, MM Ferguson, LE Holm, R Guyomard, and B Hoyheim. 2000. A microsatellite linkage map of rainbow trout (*Oncorhynchus mykiss*) characterized by large sex-specific differences in recombination rates. Genetics, 155, pp. 1331–1345.

Sakamoto T, RG Danzmann, N Okamoto, MM Ferguson, and PE Ihssen. 1999. Linkage analysis of quantitative trait loci associated with spawning time in rainbow trout (*Oncorhynchus mykiss*). Aquaculture, 173, pp. 33–43.

Schulman NF and MR Dentine. 2005. Linkage disequilibrium and selection response in two-stage marker-assisted selection of dairy cattle over several generations. J Anim Breed Genet, 122, pp. 110–116.

Senger F, C Priat, C Hitte, E Sarropoulou, R Franch, R Geisler, L Bargelloni, D Power, and F Galibert. 2006. The first radiation hybrid map of a perch-like fish: the gilthead seabream (*Sparus aurata L*). Genomics, 87, pp. 793–800.

Shirak A, A Bendersky, G Hulata, M Ron, and RR Avtalion. 2006. Altered self-erythrocyte recognition and destruction in an inbred line of tilapia (*Oreochromis aureus*). J Immunol, 176, pp. 390–394.

Shirak A, Y Palti, A Cnaani, A Korol, G Hulata, M Ron, and RR Avtalion. 2002. Association between loci with deleterious alleles and distorted sex ratios in an inbred line of tilapia (*Oreochromis aureus*). J Heredity, 93, pp. 270–276.

Soller M and JS Beckmann. 1982. Restriction fragments length polymorphisms and genetic improvement. Proc 2nd World Cong Genet Appl Livest Prod, 6, pp. 396–404.

Somorjai IM, RG Danzmann, and MM Ferguson. 2003. Distribution of temperature tolerance quantitative trait loci in Arctic char (*Salvelinus alpinus*) and inferred homologies in rainbow trout (*Oncorhynchus mykiss*). Genetics, 165, pp. 1443–1456.

Sunduimijid B. 2006. Parentage test in Angora goats using microsatellite markers and Cervus program (personal communication).

Tafalla C, A Estepa, and JM Coll. 2006. Fish transposons and their potential use in aquaculture. J Biotechnol, 123, pp. 397–412.

Villanueva B, R Pong-Wong, J Fernandez, and MA Toro. 2005. Benefits from marker-assisted selection under an additive polygenic genetic model. J Anim Sci, 8, pp. 1747–1752.

Villanueva B, R Pong-Wong, and JA Woolliams. 2002. Marker assisted selection with optimised contributions of the candidates to selection. Genet Sel Evol, 34, pp. 679–703.

Williams RW. 2006. Expression genetics and the phenotype revolution. Mamm Genome, 17, pp. 496–502.

Wilson K, Y Li, V Whan, S Lehnert, K Byrne, S Moore, S Pongsomboon, A Tassanakajon, G Rosenberg, E Ballment, Z Fayazi, J Swan, M Kenway, and J Benzie. 2002. Genetic mapping of the black tiger shrimp *Penaeus monodon* with amplified fragment length polymorphisms. Aquaculture, 204, pp. 297–309.

Wilson KJ, CP Li, JG Tong, KH Chu, S Pongsomboon, A Tassanakajon, K Glenn, M Rothschild, D Harris, and Y Li. 2004. Construction of a framework map for the black tiger shrimp *Penaeus*

monodon using genetic markers. World Aquaculture Society Conference, Waikiki, Hawaii, Book of Abstracts, p. 643.

Woram RA, C McGowan, JA Stout, K Gharbi, MM Ferguson, B Hoyheim, EA Davidson, WS Davidson, C Rexroad, and RG Danzmann. 2004. A genetic linkage map for Arctic char (*Salvelinus alpinus*): Evidence for higher recombination rates and segregation distortion in hybrid versus pure strain mapping parents. Genome, 47, pp. 304–315.

Young WP, PA Wheeler, VH Coryell, P Keim, and GH Thorgaard. 1998. A detailed linkage map of rainbow trout produced using doubled haploids. Genetics, 148(2), pp. 839–850.

Yu M, Y Cheng, and MF Rothschild. 2006. SNP analysis of molting related genes in *Penaeus monodon* and *Litopenaeus vannamei* shrimp. Arch Tierz Dummerstorf, 4, pp. 411–412.

Yu Z and X Guo. 2003. Genetic linkage map of the eastern oyster *Crassostrea virginica.* Gmelin Bio Bull, 204, pp. 327–338.

Zimmerman AM, JP Evenhuis, GH Thorgaard, and SS Ristow. 2004. A single major chromosomal region controls natural killer cell-like activity in rainbow trout. Immunogenetics, 55, pp. 825–835.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 13
Construction of Large-insert Bacterial Clone Libraries and Their Applications

Limei He, Chunguang Du, Yaning Li, Chantel Scheuring, and Hong-Bin Zhang

Large-insert bacterial clone (LBC) libraries, including bacterial artificial chromosome (BAC), bacteriophage P1-derived artificial chromosome (PAC), plant-transformation-competent binary BAC (BIBAC), conventional large-insert plasmid-based bacterial clone (PBC), and transformation-competent artificial chromosome (TAC), have been proven to be essential and desirable resources for modern genomics, genetics, and biological research of all organisms, including plants, animals, and microbes. LBC libraries have been widely used in many aspects of these studies, including genome physical mapping, large-scale genome sequencing, chromosome walking for positional cloning of genes and quantitative trait loci (QTL), genome or chromosome analysis by BAC microarray-based comparative genomic hybridization, and long-range genome analysis. In this chapter, we introduce the large-insert DNA library cloning systems developed to date, the state-of-the-art technologies for megabase-sized DNA isolation and LBC library construction, and the applications of LBC libraries in the research of genomics, molecular genetics, and molecular biology.

Large-insert DNA Libraries as Tool for Genomics Research

Genomic DNA libraries are resources essential for all areas of genomics and molecular biology research; therefore, the technologies of constructing libraries from genomic DNA of a species have been developed alongside the advances in different areas of molecular research. To facilitate cloning, maintaining and propagation of a species DNA of interest in a readily manipulated organism, that is, host (e.g., bacteria, bacteriophage, or yeast), for research purposes, several vector systems have been developed and used in construction of genomic DNA libraries (Wu et al. 2004b). These vector systems include plasmids, cosmids, bacteriophages λ and P1, yeast artificial chromosome (YAC), bacteriophage PAC, plant-transformation-competent BIBAC, and TAC. The lengths of DNA fragments that can be cloned in the vectors have been increased from several kilobase (kb) pairs to 1,000 kb pairs. The molecular research has evolved from the characterization of a single gene or a few genes or genomic loci at most in classical molecular genetics or molecular biology to the characterization of large numbers of genes or all genes at the whole genome level in genomics.

Large-insert DNA Libraries

DNA libraries that have average insert sizes of 100 kb or larger (hereafter, defined as large-insert DNA libraries) have become resources essential for many aspects of modern genomics and molecular research, and revolutionized the manpower of manipulating genomes of all organisms (Zhang et al. 1996, Zhang and Wu 2001, Ren et al. 2005). In comparison with the conventional plasmid- (approximately 10 kb), cosmid- (up to 40 kb), or λ phage- (up to 40 kb) based DNA cloning systems, large-insert DNA cloning systems are capable of cloning and stably maintaining DNA fragments of up to 1,000 kb in host cells. The significantly increased cloning capacity of large-insert DNA libraries has offered several advantages over the conventional small-insert DNA libraries. Large-insert libraries reduce the number of clones needed for a complete DNA library of a genome (see below for definition of a complete DNA library), thus allowing its individual clones to be arrayed in microplates, bar-coded, and maintained for further analysis (Zhang et al. 1996). Therefore, large-insert DNA libraries, for the first time, have become real "libraries," with each clone having a unique numbered position or "call number" that is defined by library name, plate number, row letter, and column number of the clone within a plate, so the research community could efficiently communicate and share the information about every individual clone of the libraries. Nevertheless, this is not the case for conventional small-insert libraries. The conventional small-insert DNA libraries may contain every clone of interest, but the clones usually have to be stored in bulk due to the large number of clones needed for a complete library. Because a small-insert library is stored in bulk, the clones of interest have to be identified and isolated from the library by using a long iterative library screening procedure. This is time-consuming and makes it difficult to communicate each clone of the small-insert library in the research community. Importantly, the availability of large-insert DNA libraries has enabled many genomic studies that are difficult or impossible to perform using conventional small-insert DNA libraries. Examples of these studies include, but are not limited to, genome physical mapping, large-scale genome sequencing, chromosome walking for positional cloning of genes and QTLs, genome, or chromosome analysis by BAC microarray-based comparative genomic hybridization, and long-range genome analysis.

However, large-insert DNA libraries, when arrayed in microplates, have also gained several disadvantages relative to conventional small-insert DNA libraries. First, a large amount of freezer space is needed to maintain and archive large-insert, arrayed DNA libraries because they are arrayed in microplates, whereas conventional small-insert DNA libraries are each usually stored in bulk in one or a few microtubes. Second, large robotic workstations are essential to manipulate large-insert, arrayed DNA libraries, such as high-density spotting onto nylon membrane for library screening, library duplication, and library re-arraying, whereas these are unnecessary for manipulating small-insert, bulked DNA libraries. Third, well-trained, experienced technicians are needed to maintain, archive, and manipulate large-insert DNA libraries to prevent the libraries from physical or biological contaminations and accidental loss. These requirements have not only increased the cost of archiving large-insert DNA libraries, but also make it impractical for every research group to have such large, expensive robotic workstations. Large-insert DNA library resource

centers, such as the Texas A&M University GENE*finder* Genomic Resources Center (http://hbz.tamu.edu), are a desirable approach to minimizing the limitations of large-insert DNA libraries.

Two systems have been developed to construct large-insert DNA libraries according to their host organisms. One is yeast artificial chromosome (YAC) (Burke et al. 1987), in which large DNA fragments are cloned in a YAC vector and the YAC recombinant DNA construct is hosted in a yeast strain. The other is large-insert bacterial clone (LBC) (Wu et al. 2004b, Ren et al. 2005), in which large DNA fragments are cloned in a plasmid-based vector and the LBC recombinant DNA construct is hosted in a bacterial strain. Based on modifications of plasmid vectors used for library construction, LBCs are classified into bacterial artificial chromosomes (BAC) (Shizuya et al. 1992), bacteriophage P1-derived artificial chromosomes (PAC) (Ioannou et al. 1994), plant-transformation-competent binary BACs (BIBAC) (Hamilton et al. 1996), large-insert conventional plasmid-based bacterial clones (PBC) (Tao and Zhang 1998), and transformation-competent artificial chromosomes (TAC) (Liu et al. 1999). Based on functions of the cloning vectors, LBCs are further categorized into generally genomic DNA cloning vectors such as BAC and PAC and transformation-competent binary vectors such as BIBAC, binary PBC, and TAC.

Yeast Artificial Chromosomes (YAC)

The YAC system was first reported in 1987 (Burke et al. 1987) to construct large-insert genomic DNA libraries. YACs are linear recombinant DNA molecules, each having all the elements of a native yeast chromosome, including one centromere and two telomeres derived from yeast chromosomes (Figure 13.1). YACs are capable of cloning DNA fragments of up to 1,000 kb. It was due to this feature that YACs had revolutionized research of large genomes in the late 1980s and early 1990s. YAC libraries were constructed for a number of species and used in genome physical mapping of human (Chumakov et al. 1995), mouse (Nusbaum et al. 1999), rice (Kurata et al. 1997, Saji et al. 2001), and *Arabidopsis* (Canilleri et al. 1998, Schmidt et al. 1995, Zachgo et al. 1996), but their utility was limited due to several of their significant disadvantages. The first disadvantage is their high level of chimeric clones, ranging from 10–50% of YACs. A chimeric clone contains an insert that is derived from the ligation of two or more noncontiguous genomic DNA fragments and thus, tends to mislead chromosome walking in positional cloning, physical mapping, and genome sequence assembly when they are used in genome research. The second disadvantage of YACs is their instability in host cells. This results in loss of library fidelity during storage, thus reducing the feasibility of their long-term storage and use. The third disadvantage is complicated isolation of YAC DNA, in which yeast spheroplasts are prepared, embedded in low-melting-point agarose and lysed in the agarose, followed by DNA purification and fractionation by pulsed-field gel electrophoresis and YAC DNA excision from the agarose gel. Not only is the YAC DNA isolation procedure tedious, but the isolated DNA is also easily contaminated with yeast chromosomal DNA, which limits large-scale use of YAC libraries in research of genomics, genetics, and biology (Figure 13.1).

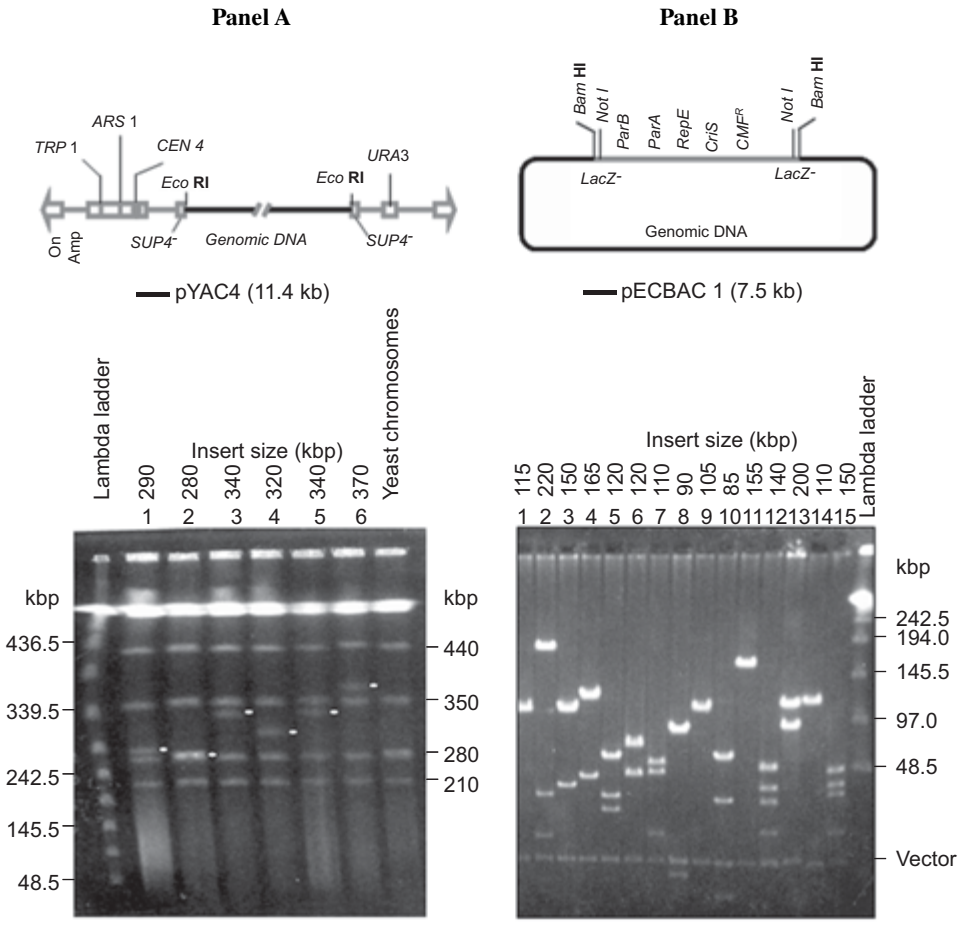


Figure 13.1. Large-insert bacterial clones (LBC) versus Yeast artificial chromosomes (YAC) fractionated on pulsed-field gels (from Wu et al. 2004b). The top figure of panel A shows the construct of a YAC and the lower figure shows YACs indicated by bullets that co-migrated with the three smallest chromosomes (210, 280, and 350 kb, respectively) of their host yeast, making them difficult to purify from the yeast DNA. The top figure of panel B shows the construct of an LBC, and the lower figure shows LBCs. The LBC DNA is readily purified and has no or little contamination with the host bacterial chromosome DNA. By digestion with *Not* I, the LBC insert DNA was released from the cloning vector. The YACs have an insert size range from 280 to 370 kb, and the BACs have an insert size range from 85 to 220 kb.

Large-insert Bacterial Clones (LBC)

The difficulties of YAC analysis and its use in genome research promoted the research of developing user-friendly systems for large-insert DNA library construction. The LBCs were first reported as BACs in 1992 (Shizuya et al. 1992) to construct large-insert DNA libraries. LBCs, including BACs, PACs, BIBACs, PBCs, and TACs, are all plasmid-based, circular DNA constructs (Figure 13.1). LBCs have several advantages over YACs (Figure 13.1). Up to 300 kb of DNA could be cloned in LBCs

and stably maintained in the bacterial host cells (Shizuya et al. 1992, Ioannou et al. 1994, Tao and Zhang 1998, Hamilton et al. 1996, Liu et al. 1999). Although these insert sizes are smaller than those of YACs, they are much larger than those of conventional cosmid and bacteriophage clones, thus being better suited for genome-scale research (Zhang et al. 1996, Ren et al. 2005). Second, LBCs are stable in the bacterial host cells (Shizuya et al. 1992, Ioannou et al. 1994, Tao and Zhang 1998, Hamilton et al. 1996, Liu et al. 1999), which is essential for long-term storage and the use of their libraries in genomics research. Finally, LBC DNA is easily purified from their bacterial host cells, which is routinely done by using the conventional plasmid DNA isolation (alkaline) procedure. This is extremely important for use of the LBC libraries in genomics research as the DNA of a large number of clones can be quickly and simultaneously purified and analyzed using a robot. It is because of these advantages of LBCs over YACs that LBCs have now become the system of choice for construction of large-insert DNA libraries.

It should be pointed out that the terms BAC, PAC, BIBAC, PBC, and TAC have been commonly used to refer to large-insert clones cloned in bacteria. The term LBC is used here to represent the clones of BAC, PAC, BIBAC, PBC, or TAC because the vectors of all of them have the same essential structure, based on either the F plasmid (BAC, BIBAC, PAC, and TAC) or the P1 plasmid (PBC and TAC), and do not include any sequences from the bacterial host genome. Furthermore, it was found that conventional plasmid-based vectors such as plasmids and cosmids previously used to construct small-insert DNA libraries for various biological research purposes can stably clone and maintain up to 300-kb DNA fragments as BACs and PACs (Tao and Zhang 1998). It was also observed that just as with single-copy BAC vectors (Shizuya et al. 1992), plasmid vectors of multiple copies are stable in the *Escherichia coli* strain DH10B that is widely used in BAC and PAC library construction. These findings imply that the conventional plasmid-based vectors, such as plasmids and cosmids, essentially have the same DNA cloning capacity as BAC, PAC, BIBAC, or TAC, suggesting that many plasmid-based vectors traditionally used for small-insert cloning could be used directly or after modification as vectors for large-insert LBC library construction. Since multiple-copy vectors, such as pCLD04541 and pSLJ1711 (Tao and Zhang 1998), are also capable of stably cloning and maintaining large DNA fragments, use of multiple-copy vectors for LBC library construction will significantly facilitate large-scale cloned DNA isolation and analysis.

In comparison, the LBC library construction system uses the electroporation technology (defined as the physical method by Wu et al. 2004b) to transform recombinant DNA constructs into the *E. coli* strain DH10B host, whereas the conventional small-insert plasmid or cosmid library construction system used the $CaCl_2$/heat shock method (defined as the chemical method by Wu et al. 2004b), or *in vitro* phage particle packaging method (defined as the biological method by Wu et al. 2004b) to transform recombinant DNA constructs. The electroporation technology allows the introducing of large-insert DNA plasmid and cosmid clones into bacterial host cells (Shizuya et al. 1992, Tao and Zhang 1998) with an extremely high transformation efficiency compared to the $CaCl_2$/heat shock and *in vitro* phage particle packaging methods previously used in the conventional small-insert plasmid or cosmid library construction. The *E. coli* strain DH10B and its derivatives can stably maintain and propagate existing plasmid vectors with large foreign DNA fragments (Tao and Zhang 1998). All LBC genomic DNA libraries constructed to date are hosted in DH10B or its derivatives. The key features of this strain include mutations that block

recombination (*recA1*), restriction of foreign DNA by endogenous restriction endonucleases (*hsd*/*RMS*), and restriction of DNA containing methylated DNA (5'-methylcytosine or methyladenine residues, and 5'-hydroxymethylcytosine) (*mcrA, mcrB, mcrC,* and *mrr*). Therefore, the electroporation technology and the *E. coli* strain DH10B have contributed to the advent of the LBC cloning technology.

General LBC cloning vectors, such as BAC and PAC, and binary vectors, such as BIBAC, TAC, and PBC, were developed to meet different research purposes. The general LBC cloning vectors can be used only to clone large DNA fragments in bacterial host cells, whereas the LBC binary vectors not only can be used to clone large DNA fragments, but also can be directly transformed into plant cells via *Agrobacterium (*Hamilton et al. 1996, 1999; Liu et al. 1999, 2002; He et al. 2003). In general, BIBACs are more difficult to construct than BACs and are not well-suited for shotgun sequencing due to their larger vector size resulting from the addition of a T-DNA cassette for plant transformation; however, once constructed, BIBACs have the potential to streamline functional analysis and use of the genomics research results for plant genetic improvement.

Construction of Large-insert Bacterial Clone Libraries

Construction of LBC libraries generally includes the following steps:

- Isolation of high-quality DNA from targeted species
- Preparation of clonable DNA fragments from the source DNA
- Preparation of cloning vectors
- Ligation of the source DNA fragments into a cloning vector
- Transformation of the ligated insert/vector recombinant DNA constructs into host cells
- Library characterization and assembly

Figure 13.2 shows a general flow chart of construction of an arrayed LBC library (Ren et al. 2005). This procedure is essentially the same as that used for construction of conventional small-insert DNA libraries. However, for LBC library construction, the source DNA must be of high quality (large size and high purity); the DNA fragments to be cloned must be handled with care to prevent them from physical shearing due to their large size; and the insert/vector recombinants must be transformed into host cells by electroporation that has been proven to enable transformation of LBCs into bacterial cells. Because the stability of LBCs is a major concern for LBC library construction and that host strain has been shown to be a major factor affecting the stability of LBCs (Tao and Zhang 1998), use of a proper bacterial host strain is crucial to construction of LBC libraries. As indicated above, all of the LBC libraries developed to date are hosted in *E. coli* DH10B or its derivatives.

Most of LBC libraries are constructed for the whole genome of targeted species; however, the strategies and associated techniques have been developed to construct an LBC library for a specific genome region (Fu and Dooner 2000). Although whole-genome LBC libraries are essential for comprehensive research of the targeted genome, a significant amount of resources and effort are needed to construct and array the library. This is especially true for species having large genomes. The strategy of genome region-specific cloning has provided a method of constructing a genomic region-specific library, thus facilitating the analysis of a particular genomic region.

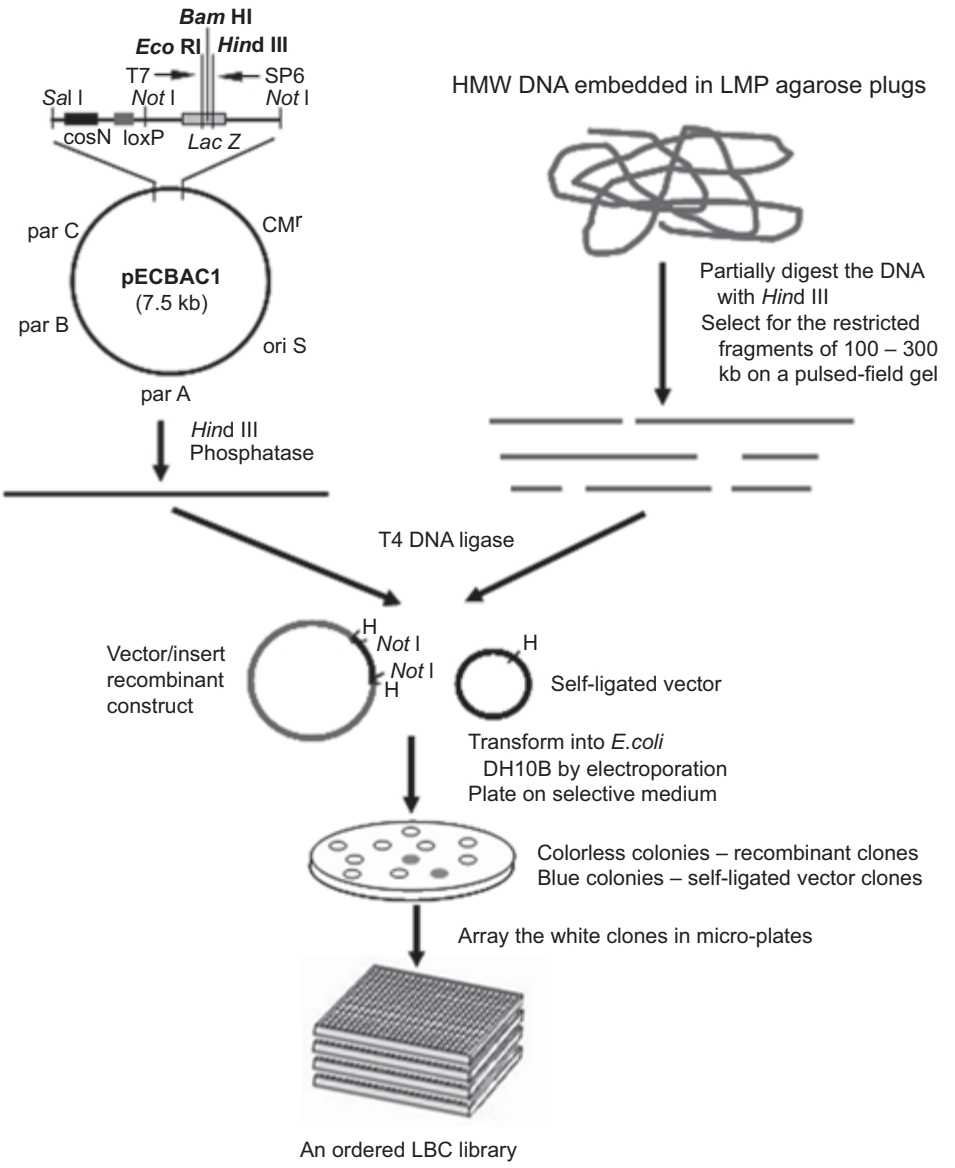


Figure 13.2. A general flow chart for construction of an arrayed large-insert bacterial clone (LBC) library (Ren et al. 2005). There are three cloning sites—*Bam* HI, *Hin*d III, and *Eco* RI, in the pECBAC1 vector (Frijters et al. 1997). In this particular procedure, *Hin*d III is used to generate clonable DNA fragments.

The insert sizes of LBC libraries are significant for their utility in genome research. What is the optimal average insert size for genome research? Generally speaking, the larger the insert sizes of LBC libraries, the better they are suited for genome research. This is because an increase in the average insert size of a library reduces the number of clones to be analyzed and results in the construction of a better quality genomic

physical map, and reduces the steps required for chromosome walking of a particular region. However, using the current techniques it is difficult, if not impossible, to construct a LBC library with an average insert size of 300 kb or larger. Most of the LBC libraries constructed and used to date have average insert sizes ranging from 100 to 200 kb (http://hbz.tamu.edu or http://hbz7.tamu.edu; http://bacpac.chori.org/; http://www.genome.arizona.edu/; http://www.genome.clemson.edu/groups/bac/). Zhang and others (1996) studied the relationships between the average insert sizes of LBC libraries and the probability of completing a 1,000-kb chromosome walk, and Ren and others (2005) studied the relationships between the average insert sizes of LBC libraries and physical map construction. Both showed that an average insert size of 160 kb or larger is desirable for efficient chromosome walking (Zhang et al. 1996) and genome physical mapping (Ren et al. 2005). The utility of the library for chromosome walking and genome physical mapping decreases significantly as the average insert size of a library drops below 160 kb. However, there is relatively little corresponding increase in the efficiency of the library for chromosome walking or physical mapping by increasing the average insert size above 160 kb.

Another factor that needs to be considered before constructing a LBC library is how many clones should be constructed for the library. This depends on the research purposes for which the library is used. Traditionally, the number of clones needed for a genomic DNA library is calculated based on the following formula (Clarke and Carbon 1976) presented by Zhang and others (1996):

$$N = \ln(1 - P)/\ln(1 - I/GS)$$

where N is the number of clones needed for a DNA library, P is the probability of isolating at least one clone from the library using a single-copy sequence, I is the average insert size of the library, and GS is the haploid genome size of targeted species. A complete genomic DNA library is defined as the one from which the probability of isolating at least one clone of interest using a single-copy sequence is greater than 99%. Nevertheless, the most common parameter of describing the number of clones needed for an LBC library today is haploid genome coverage or equivalents, which is calculated by dividing the total length of the clones contained in an LBC library by the haploid genome size of the targeted species. The total length of the clones is the product of the average insert size of the library multiplied by the number of clones contained in the library. The relationship between the probability (P) that a given clone is present in an LBC library and the library genome coverage or equivalents can be estimated by the following formula (Wu et al. 2004b, Ren et al. 2005):

$$P = 1 - e^{-n}$$

where n is the genome coverage or equivalents of an LBC library. For general genome research purposes such as library screening for clones containing a particular gene or regulatory sequence and positional cloning, an LBC library with 5 × genome coverage should be sufficient. A 5 × genome coverage library is equivalent to an LBC library having a >99% probability of isolating at least one clone using a single-copy sequence. However, if the library is to be used for whole genome physical mapping,

approximately 7–10 × genome coverage of clones are needed to construct a quality LBC-based physical map (Xu et al. 2004, 2005; Ren et al. 2005).

Construction of Whole-genome LBC Libraries

Preparation of High-molecular-weight or Megabase-sized DNA

High-molecular-weight (HMW) or megabase-sized, high-quality (high purity and readily digestible) DNA is essential for quality LBC library construction. DNA fragments that are at least fourfold as large as the desired library insert size are needed for construction of a genomic DNA library using the enzymatic method (see below). Since physical shearing is the major problem, megabase-sized DNA must be protected during isolation. Protoplasts (plants), cells (animals), or nuclei (plants and animals) are isolated and embedded in low-melting-point (LMP) agarose. The cells, nuclei, or protoplasts are lysed and DNA purified in the LMP agarose. Although several methods have been developed to prepare megabase-sized DNA from different species, the nuclei method developed by Zhang and others (1995) has become the method of choice for preparation of megabase-sized DNA from different organisms because it is widely applicable, simple, economical, and user-friendly. This method was developed initially for preparation of megabase-sized nuclear DNA from plants. Recently, it has been used for preparation of megabase-sized DNA from animals, insects, and microbes (Xu et al. 2005, H-B Zhang unpublished). Therefore, the nuclei method developed by Zhang and others (1995) and further modified by Zhang and associates (Zhang 2000, Wu et al. 2004b, Ren et al. 2005) is presented here to prepare high-quality megabase-sized nuclear DNA from a variety of different organisms, including aquaculture species.

Materials and Reagents

Materials. Plant leaves or whole seedlings, animal and marine animal muscle tissues or whole blood cells, insect larvae or pupa, or whole microbe cells can be used to isolate megabase-sized DNA using the nuclei method (Zhang et al. 1995, Zhang 2000, Wu et al. 2004b, Ren et al. 2005). The tissues can be fresh or frozen in liquid nitrogen and stored at −80°C before use. The animal blood or microbe cells should be pelleted before use or frozen.

Reagents

1. 10 × homogenization buffer (HB) base: 0.1 M Tris base, 0.8 M KCl, 0.1 M EDTA, 10 mM spermidine, and 10 mM spermine. The HB base is adjusted to pH 9.4–9.5 with NaOH and stored at 4°C.
2. 1 × HB: A suitable amount of sucrose is mixed with a suitable volume of 10 × HB base. The final concentration of sucrose is 0.5 M and HB base is 1 ×. The resultant 1 × HB is stored at 4°C.
3. Nuclei isolation buffer: It is prepared just before use by adding Triton X-l00 at 0.5% (v/v) and β mercaptoethanol at 0.15% (v/v) to 1 × HB with thorough mixing.
4. Lysis buffer: 0.5 M EDTA, pH 9.0–9.3, 1% sodium lauryl sarcosine, and 0.3 milligrams per milliliter (mg/ml) proteinase K. The lysis buffer is made just before use

by mixing equal volumes of a 1.0 M EDTA, pH 9.0–9.3 stock with a 2% sodium lauryl sarcosine stock, followed by adding proteinase K powder.

Preparation of Intact Nuclei

1. Grind 10–100 g of the fresh or frozen tissues into fine powder in liquid nitrogen with a mortar and pestle (usually it takes 20–40 minutes). Keep adding liquid nitrogen to the mortar to prevent the tissue from thawing during the grinding process. Immediately transfer the powder into an ice-cold 1,000-ml beaker containing the nuclei isolation buffer at a ratio of 10-ml/g tissue.
2. Gently swirl the contents with a magnetic stir bar for approximately 10 minutes on ice until the tissue powder has thawed in the nuclei isolation buffer. Filter into ice-cold 250-ml centrifuge bottles through two layers of cheesecloth and one layer of Miracloth (Calbiochem, USA) by gently squeezing with gloved hands.
3. Pellet the homogenate by centrifugation with a fixed-angle rotor at 2,000 g at 4°C for 20 minutes.
4. Discard the supernatant and add approximately 1 ml of ice-cold nuclei isolation buffer to each bottle.
5. Gently resuspend the pellet with assistance of a child's paintbrush that has been presoaked in ice-cold nuclei isolation buffer. Combine the resuspended nuclei from all bottles into a 40-ml centrifuge tube and fill the tube with ice-cold nuclei isolation buffer.
6. Pellet the nuclei by centrifugation at 2,000 g, 4°C for 15 minutes in a swinging bucket centrifuge or fixed-angle rotor.
7. Wash the pellet one to three additional times by resuspension in ice-cold nuclei isolation buffer using a presoaked paintbrush, followed by centrifugation at 2,000 g, 4°C for 15 minutes.
   Note: This step is necessary to minimize the contamination of cytoplast organelles and metabolic substances in the nuclei. It is especially needed in cases in which polyphenolic substances or polysaccharides are abundant in the tissues (Zhao et al. 1994, Hess 2005).
8. After the final wash, resuspend the pelleted nuclei in a small amount (about 1 ml) of ice-cold 1 × HB, count the nuclei, if possible, under a phase contrast microscope, bring to a proper concentration of nuclei/ml with addition of the 1 × HB, and store on ice.
   Note: Alternatively, the concentration of nuclei can be estimated empirically. A concentration of nuclei that is just transparent under light is estimated to be 5–10 × 10^7 nuclei/ml. The proper concentration of nuclei varies, depending on the genome sizes of the species. In general, 5–10 μg DNA/100-μl plug are suitable for LBC library construction.

Embedding the Nuclei in Low-Melting-Point (LMP) Agarose Plugs

1. Prepare a proper volume (as needed) of 1% LMP agarose in 1 × HB, and maintain in a 45°C water bath before use.
2. Prewarm the nuclei suspension to 45°C in a water bath (about 5 minutes), add an equal volume of the prewarmed (45°C) 1% LMP agarose to the prewarmed nuclei suspension and mix gently but thoroughly.
3. Aliquot the mixture into ice-cold 100-μl plug molds (BioRad, USA) on ice with a cut-off pipette tip. When the agarose is completely solidified, transfer the plugs into 5–10 volumes of lysis buffer.

Nuclei Lysis and DNA Purification

1. Incubate the agarose plugs in the lysis buffer for 24–48 hours at 50°C with gentle rotation in a hybridization oven.
2. Incubate the plugs once in 10–20 volumes of ice-cold TE (10 mM Tris-HCl, pH 8.0, 1 mM EDTA pH 8.0) and then three times in 10–20 volumes of ice-cold TE plus 0.1 mM phenylmethyl sulfonyl fluoride (PMSF) on ice, for 1 hour each incubation to remove contaminants.
   Note: PMSF is highly toxic and should be handled in a fume hood.
3. Further wash the plugs by incubating three times in 10–20 volumes of ice-cold TE on ice, 1 hour for each incubation.
4. Store the plugs in TE at 4°C before use. At this stage the plugs can be stored for several months without significant degradation.

Preparation of LBC Cloning Vectors

A number of vectors have been developed for construction of LBC libraries including BAC, BIBAC, TAC, and PBC vectors that have been widely used in LBC library construction. Furthermore, since conventional plasmid-based vectors, as demonstrated by Tao and Zhang (1998), could be used as vectors for LBC library construction, other vectors can be used for LBC library construction according to research needs.

Preparation of cloning vectors is critical for LBC library construction. The vector to be used for LBC library construction must be very pure, completely digested (but not overdigested), and completely dephosphorylated (>95% recombinant clones in a ligation test). Several methods can be used to purify vector DNA and dephosphorylate digested vectors. Introduced here is the procedure that has been routinely used in the laboratory of H-B Zhang at Texas A&M University to prepare different cloning vectors, including pBeloBAC11 (Kim et al. 1996), pECBAC1 (Frijter et al. 1997), pCLD04541 (Jones et al. 1992, Tao and Zhang 1998), pSLJ1711 (Jones et al. 1992, Tao and Zhang 1998), BIBAC2 (Hamilton et al. 1996), pYLTAC7 (Liu et al. 1999), and pBACe3.6 (Frengen et al. 1999). Using the vectors prepared by this procedure, numerous plant, animal, insect, and microbe LBC libraries have been successfully generated (see http://hbz.tamu.edu).

Vector DNA Isolation and Purification

1. Streak the stock cells of a cloning vector on a Luria Broth (LB) agar plate plus appropriate antibiotics and grow at 37°C overnight to obtain single colonies. For the vector cells with blue and white (*Lac*Z) selection such as pBeloBAC11, pECBAC1, pCLD04541, and pSLJ1171, use the LB agar plate containing appropriate antibiotics, 60 μg/ml 5-bromo-4-chloro-3-indolyl-beta-D-galactoside (X-gal) and 14 μg/ml IPTG (isopropylthiogalactoside). For the vector cells with *Sac*B selection such as BIBAC2 and pBACe3.6, use the LB agar plate containing appropriate antibiotics and 5% (w/v) sucrose.
2. Select a single blue colony (*Lac*Z) or a growing colony (*Sac*B) and inoculate into 1,500 ml LB or Terrific Broth (TB) liquid medium containing appropriate antibiotics. Alternatively, directly inoculate 5 μl of a single colony freezer glycerol stock into the 1,500 ml medium. Grow overnight (18–20 hours) at 37°C, 250 rpm.

3. Harvest the culture cells in 250-ml centrifuge bottles and centrifuge at 9,820 g, 4°C for 10 minutes.
4. Discard the supernatant, and completely resuspend the bacterial cell pellet in each bottle (from 200-ml culture) in 10 ml of Solution I (50 mM glucose, 10 mM EDTA, pH 8.0, 25 mM Tris-HCl, pH 8.0) and incubate on ice for 10 minutes.
5. Slowly add 20 ml of freshly prepared Solution II (0.2 N NaOH, 1% sodium dodecyl sulfate [SDS]) per bottle, while the bottle is being gently shaken on ice. The solution should become clear and viscous immediately. Incubate on ice for 5–10 minutes.
6. Add 15 ml of ice-cold Solution III (5 M KOAc, pH 4.8–5.3). Gently invert and swirl to mix the contents. A white precipitate should form immediately. Incubate on ice for 5–10 minutes. The precipitate that forms at 0°C consists of chromosomal DNA, high-molecular-weight RNA, and potassium/SDS/protein/membrane complexes.
7. Centrifuge the bacterial lysate at 3,840 g, 4°C for 10 minutes, and filter the supernatants in all bottles through four layers of cheesecloth into a clean flask to collect any of the precipitate.
8. Transfer the supernatant into fresh centrifuge bottles, add 0.6 volume of isopropanol, mix well, and incubate at room temperature for 10 minutes.
9. Recover the DNA by centrifugation at 15,300 g at room temperature for 10 minutes.
10. Decant the supernatant carefully, and invert the open bottle to allow the last drops of supernatant to drain away. Carefully rinse the pellet and the wall of the bottle with 70% ethanol at room temperature. Drain off the ethanol and place the inverted, open bottle on paper towels for 15–30 minutes at room temperature to allow all traces of ethanol to evaporate.
11. Completely dissolve the DNA pellet (actually the DNA and RNA pellet) in 15 ml of TE. Incubation of the contents at 65°C for 10–20 minutes would be helpful to dissolve the DNA and RNA pellet in TE.
12. Transfer the DNA solution into a graduated cylinder of known weight, add 10 mg/ml ethidium bromide (EB) stock to the DNA solution at a rate of 0.8 ml EB per 10 ml DNA solution, and measure the total volume of the DNA/EB solution. To every ml of DNA/EB solution, add 1.0 g of solid CsCl. Mix the solution gently until the CsCl salt is completely dissolved.
13. Weigh the DNA/CsCl/EB solution in the cylinder and adjust the density of the solution to 1.50–1.60 g/ml (usually 1.56 g/ml).
14. If too much furry scum forms, centrifuge the solution at 8,820 g at room temperature for 5 minutes. The furry scum that floats to the top consists of complexes formed between the EB and bacterial proteins. (Usually, this step is unnecessary.)
15. Transfer the DNA/CsCl/EB solution into ultracentrifugation tubes using a Pasteur pipette, balance the tubes (<0.1 g between the paired tubes) and seal or cap.
16. Centrifuge the density gradients at 45,000 rpm for 16 hours (VTi65 rotor), 45,000 rpm for 48 hours (Ti50), 60,000 rpm for 24 hours (Ti65), or 60,000 rpm for 24 hours (Ti70.1) at 20°C.
    Note: Two bands of DNA, located in the center of the gradient, should be visible under long-wave ultraviolet (UV) light. The upper band, which usually contains less material, consists of linear bacterial chromosomal DNA and nicked circular

plasmid DNA; and the lower band consists of closed circular plasmid DNA. The deep-red pellet on the bottom of the tube consists of EB/RNA complexes. Therefore, the lower band of closed circular plasmid DNA should be collected. Wear a facemask, gloves, and lab coat to conduct steps 16 and 17. Exposure to UV light is harmful to skin and eyes.

17. Remove the cap or insert a 21-gauge needle into the top of the tube to allow air to enter and collect the lower circular plasmid DNA band with an 18-gauge needle.
18. Extract the plasmid DNA solution with an equal volume of H_2O-saturated isoamyl alcohol for 4–5 times to completely remove the EB in the solution (as indicated by colorless solution).
19. Dilute the DNA/CsCl solution with three to four volumes of double distilled H_2O and precipitate the DNA by adding one volume of isopropanol and incubating at 4°C for 15 minutes, followed by centrifugation at 13,800 g, 4°C for 30 minutes.
20. Discard the supernatant, rinse the DNA pellet in 70% ethanol by centrifugation at 13,800 g, 4°C for 10 minutes, air dry, dissolve the DNA in 0.5 ml TE, and measure the concentration of the DNA.
    Note: Handle the DNA pellet carefully during the 70% ethanol wash because the DNA pellet in 70% ethanol is loose and could be easily lost. To minimize this probability, always centrifuge the tubes at the same orientation as Step 19.

Vector DNA Digestion

1. Set up the digestion of vector DNA as below and incubate at 37°C for 2 hours.

H_2O	238 μl		
Vector DNA	100 μl	(5–10 μg)	
10 × Reaction buffer	40 μl		
40 mM Spermidine (Spd)	20 μl		
10 U/μl *Bam* HI, *Hin*d III or *Eco* RI	2 μl		
	400 μl		(13.1)

2. Transfer the reaction onto ice and check a small aliquot of the digestion on a 1% agarose gel to ensure that the digestion is complete. Do not add gel-loading dye to the entire sample.
3. Extract the digestion with an equal volume of saturated phenol/chloroform (1:1), spin at 12,000 g, room temperature for 5 minutes, and transfer the aqueous phase (top layer) into a new tube.
   Note: It is crucial to success to use fresh phenol for this step.
4. Precipitate the DNA by adding 1/10 volume of 3 M NaAc, pH 5.2, and one volume of isopropanol, and incubating at -80°C for 10 minutes, followed by centrifugation at 12,000 g for 15 minutes.
5. Discard the supernatant, wash the pellet carefully with 70% ethanol, air dry, dissolve in 100 μl H_2O, and measure the concentration of the DNA by comparison with samples of lambda DNA of known concentrations on a 1% agarose gel.

Dephosphorylation of Linearized Vector DNA

1. Set up the dephosphorylation reaction as below and incubate at 37°C for 30 minutes.

H_2O	256 μl
Digested vector DNA	100 (10 μg) μl
10 × CIAP reaction buffer	40 μl
1 U/μl CIAP (Invitrogen, USA)*	3.89 μl
	400 μl

(13.2)

   Note: The amount of calf intestinal alkaline phosphatase (CIAP) used depends on the size of the vector. For pBeloBAC11 or pECBAC1 (7.5 kb), add 3.89 U CIAP / 10 μg DNA; for BIBAC2 (23.5 kb), add 1.23 U CIAP/10 μg DNA; and for pCLD04541 (27.6 kb), add 1.0 U/10 μg DNA.
2. Stop the reaction immediately by adding 4 μl 0.5 M EDTA, pH 8.0, 20 μl 10% SDS and 40 μl 1 mg/ml proteinase K in cold TE to the tube. Incubate at 56°C for 30 minutes.
3. Cool down to room temperature, and extract once with an equal volume of saturated phenol, once with an equal volume of saturated phenol/chloroform/iso-amyl alcohol (25:24:1) (see above), and then transfer the aqueous phase (top layer) into a new tube.
   Note: It is crucial to success to use fresh phenol for this step.
4. Precipitate the DNA by adding 1/10 volume of 3 M NaAc, pH 7.0, and one volume of isopropanol, and incubating at -80°C for 10 minutes, followed by centrifugation at 12,000 g for 15 minutes.
5. Discard the supernatant, wash the pellet carefully with 70% ethanol, air dry, dissolve in 200 μl H_2O, and measure the concentration of the DNA using lambda DNA of known concentration on a 1% agarose gel.
6. Adjust the concentration of the DNA to 10 ng/μl for pBeloBAC11 or pECBAC1, and to 40 ng/μl for BIBAC2 or pCLD04541, aliquot the DNA and store in a −20°C freezer. The vector DNA in the 20°C freezer is good for cloning for 3–6 months.
7. Optional: Check the dephosphorylation of the vector using the conventional ligation test. When the vector DNA is ligated to digested lambda DNA at a molar rate of 3 vector:1 digested lambda DNA and transformed into *E. coli* DH10 cells by electroporation (see below), the percentage of recombinant (white) clones should be greater than 95%.

Preparation of Clonable DNA Fragments

Generation of clonable DNA fragments is an essential step for DNA library construction. Large DNA fragments can be generated by either physical shearing or partial digestion of megabase-sized DNA with a restriction enzyme that cuts relatively frequently within a genome. However, almost all of the existing YAC and LBC libraries were developed from the DNA fragments generated by the partial digestion (or enzymatic) method whereas only a few of the LBC libraries were constructed by the physical shearing method (e.g., see http://bacpac.chori.org/). By comparison, the physical

shearing method is independent of the distribution and frequency of restriction sites of a particular restriction enzyme. Therefore, the fragments generated with this method are most randomly distributed in the genome and the library constructed from such fragments is most representative of the total genome. However, the sheared DNA fragments have to be repaired by blunting at the fragment ends, and/or modified by adding restriction site-containing synthetic linkers prior to cloning. For the best results, these subsequent steps are often performed with naked DNA, which would damage the large DNA fragments and lead to a low yield of clonable DNA fragments. Because of this, the partial digestion method has been widely used to fragment megabase DNA for LBC library construction. The major advantage of the enzymatic method is that the partially restricted fragments can be directly cloned (i.e., no further enzymatic modification of the restricted fragment ends is necessary). For LBC library construction, at least four methods have been used to generate large DNA fragments:

1. varying the concentration of the restriction enzyme with a fixed digestion time
2. varying the time of digestion with a fixed amount of the restriction enzyme
3. varying the concentration of a restriction enzyme cofactor (e.g., Mg^{2+}) with a fixed amount of the enzyme
4. varying the ratio of the restriction enzyme to the corresponding methylase

It would be worth noting that because the restriction sites of a particular enzyme are often distributed along the genome unevenly, the distribution of the resultant library clones along the genome are often biased and thus, the representation of the library is often lower, even though the clones of the same genome coverage are in the library as those constructed with the physical shearing method. To minimize the problem associated with the enzymatic method, two or more separate enzymes are recommended for construction of complementary LBC libraries.

Here, we will generate the clonable DNA fragments by varying the concentration of the restriction enzyme with a fixed digestion time because it is the one that has been widely used in LBC library construction. Since it is unknown how much of the restriction enzyme is used to generate the optimum number of clonable DNA fragments with desirable sizes, a set of pilot digestions is often needed to determine the optimal amount of restriction enzyme before large-scale partial digestion is carried out for LBC library construction.

Determination of Optimal Partial Digestion Condition

1. Cut three 100-μl DNA LMP agarose plugs into 27 slices of approximately equal size using a microscope glass slide cover, with nine slices/100-μl plug, and transfer them into a 50-ml Falcon tube.
2. Add 8,730 μl of the incubation buffer to the 50-ml Falcon tube (for nine reactions) and incubate on ice for 30 minutes.

Incubation Buffer (For 9 reactions)

H_2O	$867 \times 9 \times 2 = 15{,}606$ μl
$10 \times$ enzyme buffer	$100 \times 9 \times 2 = 1{,}800$ μl
1 M spermidine	$2 \times 9 \times 2 = 36$ μl
1 M DTT	$1 \times 9 \times 2 = 18$ μl

(13.3)

3. After the first 30-minute incubation, remove the incubation buffer from the Falcon tube, add the remaining 8,730 μl to the tube, and incubate on ice for another 30 minutes.
4. Transfer the plug slices into nine 1.5-ml microtubes, three slices per tube, add 170 μl of the digestion buffer per tube, and incubate on ice for 10 minutes.

Digestion Buffer (For 10 reactions)

H_2O	$137 \times 10 = 1{,}370$ μl
$10 \times$ enzyme buffer	$17 \times 10 = 170$ μl
1 M spermidine	$0.34 \times 10 = 3.4$ μl
1 M DTT	$0.17 \times 10 = 1.7$ μl
10 mg/ml BSA	$10 \times 10 = 100$ μl

(13.4)

5. Add the restriction enzyme (*Bam* HI or *Eco* RI) to be used for the LBC library construction in a volume of 2–10 μl per tube to give a series of amounts of 0.0, 0.2, 0.4, 0.8, 1.2, 1.6, 2.0, 2.4, and 4.8 U per reaction (tube).
   Note: The enzyme should be diluted with the digestion buffer.
6. Mix the reactions and incubate the tubes on ice for 100 minutes.
7. Transfer the tubes into a 37°C water bath and incubate at 37°C for 8 minutes.
8. Stop the reaction by immediately transferring the tubes onto ice and adding 1/10 volume of 0.5 M EDTA, pH 8.0 to each tube.
9. Analyze the partial digestions by pulsed-field gel electrophoresis (Bio-Rad, USA) under the following conditions: 1% agarose in 0.5 × TBE (5 × stock: 0.45 M Tris base, 0.45 M boric acid, 10 mM EDTA, pH 8.3), 12.5°C (cooler settings), 80 (pump settings), 120° included angle, 6 V/cm, initial pulse time of 50 seconds and final pulse time of 50 seconds for 18–24 hours.
10. Stain the gel and photograph. The concentration of the enzyme under which most of the partially restricted fragments fall in a range from 100–300 kb on the gel is selected for large-scale partial digestion for LBC library construction.

Large-scale Partial Digestion of Megabase-sized DNA

1. Cut ten 100-μl DNA LMP agarose plugs into 90 slices using a microscope glass slide cover, with nine slices per 100-μl plug, and transfer them into a 50-ml Falcon tube.
2. Add 29,100 μl of the incubation buffer made as above to the 50-ml Falcon tube (for 30 reactions) and incubate on ice for 30 minutes.
3. After the first 30-minute incubation, remove the incubation buffer from the Falcon tube, add another fresh 29,100 μl of the incubation buffer to the tube, and incubate on ice for another 30 minutes.
4. Transfer the plug slices into thirty 1.5-ml microtubes, three slices per tube, add 170 μl of the digestion buffer made as above per tube, and incubate on ice for 10 minutes.
5. Add the restriction enzyme (*Bam* HI, *Hin*d III or *Eco* RI) to be used for the LBC library construction to each 1.5-ml microtube at the optimal concentration determined above in a volume of 2–10 μl per tube.
6. Mix the reaction and incubate the tubes on ice for 100 minutes.

7. Transfer the tubes into a 37°C water bath and incubate at 37°C for 8 minutes.
8. Stop the reactions by immediately transferring the tubes onto ice and adding 1/10 volume of 0.5 M EDTA, pH 8.0 to each tube.

Size Selection for Restricted DNA Fragments Ranging from 100–300 kb

The First Size Selection.

1. Prepare a 1% agarose gel with a sample trough of suitable size for the DNA-restricted plug slices in 0.5 × TBE and allow the gel to completely solidify. Reserve 2–3 ml of the agarose gel at 60°C for sealing the plugs into the gel.
2. Prechill the gel running buffer (0.5 × TBE) in the electrophoresis chamber.
3. Load the plug slices into the leading edge of the sample trough and also load one to two lanes with very thin slices of lambda ladder PFG markers (New England Biolabs, USA) on either side of the sample DNA trough. Seal the DNA samples and markers by carefully covering them with the molten agarose reserved from Step 1.
4. Size-select the partial digestion by pulsed-field gel electrophoresis under the following conditions:
   - Block 1: 11.5°C (cooler settings), 80 (pump settings), 120° angle, 6 V/cm, initial pulse time of 90 seconds and final pulse time of 90 seconds for 13 hours.
   - Block 2: 11.5°C (cooler settings), 80 (pump settings), 120° included angle, 4 V/cm, initial pulse time of 5 seconds and final pulse time of 5 seconds for 5 hours.
5. Excise the gel zones containing the lambda ladder markers and the two outer 5-mm edges of the DNA sample trough, stain, and mark the 100–300 kb zone with a razor blade on each side. During this time, the remaining gel that contains the DNA sample is maintained at 4°C.
6. Place the stained gel pieces beside the unstained sample region of the gel. Locate the zone ranging from 100–300 kb on the unstained gel using the marks made on the above stained gel, excise with a razor blade (usually 1.0 cm wide) and divide the gel zone into two sections (0.5 cm per section) horizontally: 100–200 kb and 200–300 kb.
7. Electroelute the selected DNA in each gel section (slice) into a length of Spectro/Por 7 (Spectrum Labs, USA) dialysis tubing (cutoff = 14 KD; 1/4 inch in diameter) as follows (DNA electroelution):
   a. Cut the dialysis tube into a piece 5–10 cm long, depending on the size of the gel section to be electroeluted, rinse with ice-cold dd. H_2O a few times and then with ice-cold 0.5 × TBE.
   b. Place the gel section containing the size-selected DNA fragments into the tube, close one end of the tube with a membrane tubing closure (Spectrum Labs, USA), fill the tube with ice-cold 0.5 × TBE to completely submerge the gel slice in the buffer (usually 100–300 μl), remove all bubbles in the tube, and close the other end of the tube.
   c. Submerge the dialysis tubing in 0.5 × TBE in the electrophoresis chamber oriented horizontally and electroelute the DNA in the gel section by pulsed-field gel electrophoresis under the following conditions: 12.5°C (cooler settings), 80 (pump settings), 120° included angle, 6 V/cm, initial pulse time of 35 seconds and final pulse time of 35 seconds for 4 hours.

d. After electrophoresis, turn the dialysis tube 180° and continue to run the tube for 1 minute to allow the eluted DNA to be released from the dialysis tube wall.
e. Carefully collect the DNA in the dialysis tube with a cut-off tip into a 1.5-ml microtube.
   Note: This DNA can be directly used for ligation after dialysis against 0.5 × TE (see below), or go to the second size-selection, as needed.

The Second Size Selection (Optional).

1. Prepare a new 1% agarose gel with sample troughs of suitable sizes for DNA samples collected from the first size selection in 0.5 × TBE and allow the gel to completely solidify. Reserve 1–2 ml of the agarose gel at 60°C for later use in gel loading.
2. Prechill the gel running buffer (0.5 × TBE) in the electrophoresis chamber.
3. Load one to two lanes of thin slices of lambda ladder PFG markers on either side of the sample and seal in the wells with the molten agarose reserved from Step 1.
4. Submerge the gel in the chilled 0.5 × TBE in the buffer chamber, add 1/10 volume of the loading dye to the DNA collected from the first size selection, and load in the gel with a cut-off tip.
5. Run the gel under the following conditions:
   - 12.5°C (cooler settings)
   - 80 (pump settings)
   - 120° included angle, 4 V/cm, initial pulse time of 5 seconds and final pulse time of 5 seconds for 6–8 hours
6. Excise the gel zones containing the lambda ladder markers and the 2- to 5-mm edge of the DNA sample troughs, stain, and mark the compressed DNA (>100 kb) on the gel with a razor blade.
7. Locate the compressed DNA zone on the unstained gel using its corresponding gel mark as reference, excise with a razor blade and perform the electroelution of the DNA as the above first size selection. The DNA eluted from the gel slice can be directly dialyzed in the same dialysis tubing against ice-cold 0.5 × TE.

Ligation of the Size-selected DNA Fragments to the LBC Cloning Vector

1. Dialyze the DNA eluted from the gel slice at least three times in the same tube against one liter of ice-cold 0.5 × TE on ice, for at least 1 hour each time. (It is okay to dialyze the DNA overnight, but it is preferable to conduct ligation immediately after three dialyses.)
2. Carefully collect the DNA sample from the dialysis tube with a cut-off tip into a 1.5-ml microtube.
3. Estimate the concentration of the DNA on a 1% agarose gel using known concentrations of lambda DNA as standards.
4. Set up the ligations using the following criteria:
   - The molar ratio of vector:DNA = 2–10:1.
   - The final insert DNA concentration of the ligation reaction is 1–2 ng/μl.
   - The amount of T4 DNA ligase is 1–2 U/50 μl ligation reaction.
     Note: Do not mix the reaction with a pipette tip-mix the reaction by gently inverting the tube several times. The vector:insert molar ratio would affect the percentage of recombinant clones containing inserts. In general, a smaller vector:insert ratio results in a higher percentage of recombinant clones.

5. Spin briefly to collect the reaction on the bottom of the tube, transfer 10 μl of the ligation reaction mixture into a 0.5-ml tube containing 200–400 ng lambda/*Hin*d III DNA in a volume of 1–2 μl, mix and spin down. The lambda/*Hin*d III DNA is used as an indicator of the DNA insert/vector ligation reaction (ligation check).
6. Incubate the ligation reactions along with the ligation check reactions at 16°C for 8–12 hours.

Transformation of the Ligated DNA into *E. coli* DH10B by Electroporation

Electroporation is a popular technology currently used to deliver vector/insert recombinant DNA constructs into library host cells. In electroporation, the cell membrane is exposed to high-intensity electric field pulses and thus temporarily destabilized in specific regions of the cell. During this period, the cell is highly permeable to exogenous molecules present in the surrounding media. The DNA molecules are thus introduced into the host cells.

1. Prepare agar plates containing 25 g LB base, 15 g agar, appropriate antibiotics, 75 μl 200 mg/ml IPTG, and 3 ml 20 mg/ml X-gal per liter.
2. Check the ligation reactions by running the ligation checks on a 1% agarose gel using the lambda/*Hin*d III DNA as a control. The disappearance of the expected lambda *Hin*d III bands and the formation of a larger single band of the ligation check indicate that the corresponding ligation reaction works.
3. Transform the ligated DNA into electrocompetent *E. coli* Strain DH10B cells (Invitrogen, USA) by electroporation using the Cell Porator and Voltage Booster System (BRL Gibco, USA) at the following settings:

Cell Porator		Voltage Booster	
Voltage:	375 volts	Resistance:	4K ohms
Capacitance:	330 μF		
Impedance:	Low ohms		
Charge rate:	Fast		

4. Collect the cells from the cuvette with a pipette into 0.5–1.0 ml SOC medium (2% Bacto tryptone, 0.5% Bacto yeast extract, 10 mM NaCl, 2.5 mM KCl, 20 mM $MgSO_4/MgCl_2$, 20 mM glucose, pH 7.0) and incubate at 37°C with shaking at 200 rpm for 1 hour to allow the cells to develop resistance to antibiotics.
5. Plate the cells (0.5–1 ml) on the LB agar plate containing antibiotics, IPTG, and X-Gal (see Step 1) and incubate at 37°C for 36–48 hours to allow the colony colors to fully develop.
   Note: The colorless (white) clones for LacZ gene vectors are putative recombinants (LBCs).

LBC DNA Isolation and Analysis

It is essential to isolate DNA from the potential recombinant clones growing on the agar plates and estimate their insert sizes before they are arrayed into microplates for LBC library assembly. Generally, at least 100 random clones are analyzed, of which >95% should have inserts. The insert size of a clone is calculated by adding all insert fragments of the clone appearing on a pulsed-field gel.

1. Inoculate a single colony in 5 ml of LB plus suitable antibiotics (for the BACs in pBeloBAC11 and in pECBAC1, use 12.5 μg/ml chloramphenicol; for the binary BACs in pCLD04541, use 15 μg/ml tetracycline) and grow at 37°C with shaking at 250 rpm for 16 to 20 hours.
2. Centrifuge the entire overnight culture at 1,500 g in a desktop centrifuge for 10 minutes, pour out the supernatant, resuspend the pellet in the remaining culture medium (about 0.05–0.1 ml) with a vortex, and transfer into 1.5-ml microtubes.
3. Add 0.2 ml of Solution 1 (50 mM glucose, 10 mM EDTA, pH 8.0, 25 mM Tris. HCl, pH 8.0) and incubate on ice for 5 minutes.
4. Add 0.4 ml of Solution 2 (0.2 N NaOH, 1% SDS), mix gently, and incubate on ice for 5 minutes.
5. Add 0.3 ml of Solution 3 (3 M KOAc, pH 4.8–5.3), mix gently, and incubate at 4°C for 30 minutes.
6. Centrifuge in a microcentrifuge at 12,000 g for 5 minutes to collect the precipitate.
7. Transfer 0.75 ml of the supernatant to a new 1.5-ml microtube, add 0.45 ml isopropanol, mix, and centrifuge at 12,000 g in a microcentrifuge for 5 minutes to pellet DNA.
8. Remove the supernatant and wash the pellet with 70% ethanol by adding 1.0 ml 70% ethanol to each tube, followed by centrifuging at 12,000 g for 2 minutes and discarding the ethanol supernatant.
   Note: Care must be taken when discarding the ethanol supernatant because the DNA pellet is easily dislodged from the tube wall.
9. Dry the DNA pellet on bench by inverting the tube (15–30 minutes), add 40 μl TE, and dissolve the DNA. Incubation of the pellet at 65°C for 10 minutes helps to dissolve the DNA in TE.
10. Take 5–10 μl of the DNA solution, digest with 1–2 U *Not* I for 1–3 hours at 37°C in a reaction volume of 40 μl as follows:

H_2O	25.8 μl
LBC DNA	8 μl
10 × *Not*I buffer	4 μl
40 mM spermidine	2 μl
10 units/μl *Not*I	0.2 μl
	40 μl (13.5)

11. Stop the reaction by adding 1/10 volume (4 μl) of 10 × gel loading dye, heat the digested DNA at 65°C for 10 minutes (optional), and run on an 1% agarose by pulsed-field gel electrophoresis in 0.5 × TBE under the following conditions:
    - 12.5°C (cooler settings)
    - initial pulse time of 5 seconds
    - final pulse time of 15 seconds
    - 120° included angle
    - 6V/cm
    - 16 hours
12. Stain the gel for 30 minutes, destain for 30 minutes in water, as needed, photograph the gel for your documentation, and estimate both the percentage of clones having no inserts and the insert size of each clone (Figure 13.1).

LBC Library Assembly

Once a ligation that yields a suitable number of clones per transformation (at least 100 recombinant DNA clones/transformation are needed in order to assemble a LBC library from the ligation) and with desirable insert sizes is identified, it is transformed into the host competent cells (DH10B) on a large scale and plated on a proper selective medium as described above. The recombinant clones on the plates are used to assemble the targeted, arrayed LBC library.

1. Array the clones with desirable insert sizes in 384-well microplates, with each well containing 50 μl of the freezing medium plus appropriate antibiotics (see below).

   10 × Freezer buffer (for cell storage): 100 ml

Ingredient	Amount	Final Concentration
K_2HPO_4	6.27 g	360 mM
KH_2PO_4	1.80 g	132 mM
Na citrate	0.50 g	17 mM
$MgSO_4\ 7H_2O$	0.10 g	4 mM
$(NH_4)_2SO_4$	0.90 g	68 mM
Glycerol	44 ml	44% (v/v)
H_2O	to 100 ml	

   Adjustment: Sterilize the buffer by autoclaving and store at room temperature. (It is normal if precipitation occurs during storage.)

   Cell freezer storage medium: 100 ml

Ingredient	Amount
LB broth	90 ml
10 × freezer buffer	10 ml

   Adjustment: Add appropriate antibiotics to the medium before use.

2. Grow at 37°C overnight until the medium in the wells becomes turbid and store the clones at −80°C. The clones can be maintained in the medium at −80°C for 5 or more years without significant problems. To prevent the library from contamination and accident loss, duplicate the library into three copies, one being used as a working copy, one as a master copy, and one as a backup copy.

Construction of LBC Libraries Containing Genome Fragments of Interest

Construction of LBC libraries for a particular genomic region of interest is technically the same as the construction of whole-genome LBC libraries for the steps of preparation of megabase-size DNA from targeted species, preparation of cloning vectors, ligation of the source DNA fragments into a cloning vector, transformation of the ligated insert/vector recombinant DNA constructs into host cells, and library assembly. The major difference between the two is the preparation of clonable DNA fragments from the targeted genomic region for genomic region-specific or enriched library construction. Moreover, a specifically designed vector containing a cloning site of a rare cutting enzyme, such as pNOBAC 1(*Not* I) and pCLD04541 (*Sst* I, although *Sst* I is a 6-bp cutter, it was found to be rare in certain genomes), is needed to construct a LBC library for a particular region of a genome. Therefore, here we only provide the procedure of

preparing clonable DNA fragments spanning the genomic region of interest (Fu and Dooner 2000). This includes two steps: physical mapping of the genomic region of interest and preparation of clonable DNA fragments spanning the region.

Physical Mapping of a Genomic Region of Interest

1. Cut 1–4 100-μl DNA LMP agarose plugs into slices of approximately equal size using a microscope glass slide cover, with nine slices per 100-μl plug, and transfer them into a 50-ml Falcon tube.
2. Add the incubation buffer made as above (Determination of Optimal Partial Digestion Condition) to the 50-ml Falcon tube at 970 μl per reaction and incubate on ice for 30 minutes.
3. After the first 30-minute incubation, remove the incubation buffer from the Falcon tube, add the same amount of fresh incubation buffer as Step 2 to the tube, and incubate on ice for another 30 minutes.
4. Transfer the plug slices into 1.5-ml microtubes, three slices per tube, add 170 μl of the digestion buffer made as above (Determination of Optimal Partial Digestion Condition) per tube, and incubate on ice for 10 minutes.
5. Add a rare-cutting restriction enzyme, such as *Not* I or *Sst* I (depending on the availability of cloning vectors), to each 1.5-ml microtube, in a volume of 2–10 μl per tube.
6. Mix the reaction and incubate the tubes on ice for 100 minutes.
7. Transfer the tubes into a 37°C water bath and incubate at 37°C for 3 hours to allow the DNA to be digested completely.
8. Stop the reactions by transferring the tubes onto ice and adding 1/10 volume of 0.5 M EDTA, pH 8.0 to each tube. Keep the reactions on ice before use.
9. Prepare a 1% agarose gel in 0.5 × TBE and allow the gel to completely solidify. Reserve 3–5 ml of the molten agarose at 60°C for later gel loading.
10. Prechill the gel running buffer (0.5 × TBE) in the electrophoresis chamber.
11. Load the completely digested DNA slices, along with lambda ladder PFG markers, into the gel, seal with the 60°C-reserved 1% agarose, and run the gel under the following conditions:
    - 12.5°C (cooler settings)
    - 80 (pump settings)
    - 120° included angle
    - 6 V/cm
    - initial pulse time of 0.1 seconds
    - final pulse time of 40 seconds for 18–24 hours
12. Stain the gel with ethidium bromide and photograph.
13. Nick the large DNA fragments in the gel either with 60 mJoules of UV light (254 nm) using the GS Gene Linker (BioRad, USA), or by treating the gel with 0.2 N HCl for 10 minutes, followed by washing in 2 × SSC (20 ×: 3 M NaCl, 3.3 M Na citrate, pH 7.0) for 5 minutes.
14. Southern blot the DNA from the gel onto Hybond-N+ (Amersham, USA) membrane with 1 liter of 1.5 M NaCl, 0.4 N NaOH for 48 hours.
15. Wash the membrane in 2 × SSC for 10 minutes with gentle shaking. The membrane can be immediately used for Southern blot hybridization, or wrapped with plastic wrap and stored at 4°C for later use.

16. Probe the pulsed-field gel Southern blot with DNA marker(s) specific for the genomic region of interest just as with a conventional Southern blot hybridization. To show the lambda ladder markers on the blot autoradiograph, a small amount of lambda DNA (0.5–2.0 ng) could be added to the probe DNA before labeling. The size of the DNA fragment spanning the genomic region of interest could be determined by the Southern analysis.

Preparation of Clonable DNA Fragments for Construction of a Genomic Region-specific LBC Library

1. Completely digest the megabase-size DNA of the targeted species on a larger scale (approximately five to ten 100-μl plugs) as described above (Physical Mapping of a Genomic Region of Interest) with the selected enzyme that generates DNA fragments having a size (<300 kb) suitable for LBC cloning and spanning the genomic region of interest.
2. Analyze the digested DNA on a 1% agarose gel by pulsed-field gel electrophoresis as described above (Physical Mapping of a Genomic Region of Interest).
3. Excise the gel zone containing the lambda ladder PFG markers and the outer 2–5 mm edges of the DNA sample trough, stain, and mark the zone containing the DNA fragments of interest with a razor blade. The remaining unstained portion of the gel is maintained at 4°C.
4. Locate the gel zone of desired range on the unstained portion of the gel using the stained gel mark as a reference and excise the gel zone of interest with a razor blade.
5. Elute the DNA fragments from the agarose zone by pulsed-field gel electrophoresis, dialyze in the same tube against ice-cold 0.5 × TE, and collect the DNA in the dialysis tube with a cut-off tip into a 1.5-ml microtube as described above (Whole-genome LBC Library Construction).
   Note: As in the procedure of the Whole-genome LBC Library Construction, the DNA selected can be ligated into a cloning vector containing the cloning site of the enzyme (for instance, in the case of *Not* I, use pNOBAC 1) and transformed into the *E. coli* DH10B cells. The transformed cells are plated on the selective medium, and the resulting recombinant clones are analyzed by pulsed-field gel electrophoresis.

Currently Existing LBC Libraries

Since the LBC system was first reported (Shizuya et al. 1992), at least 500 arrayed LBC libraries have been constructed from a variety of species and genotypes of different organisms, including human, animals, plants, fishes, insects, and microbes, for various research requirements. Most of the libraries are available to the public at http://hbz.tamu.edu or http://hbz7.tamu.edu, http://bacpac.chori.org/, http://www.genome.arizona.edu/, and http://www.genome.clemson.edu/groups/bac/. Of these LBC libraries, a couple of dozen were constructed from fishes, including channel catfish (*Icalurus punctatus*), Atlantic salmon, chinook salmon (*Oncorhynchus tshawytscha*), fresh water stickleback, three-spined stickleback (*Gasterosteus aculeatus*), stickleback fish

(*Gasterosteus aculeatus*), zebrafish, yellowbelly rockcod (*Notothenia coriiceps*), swordtail fish (*Xiphophorus helleri*), Antarctic icefish (*Chaenocephaus aceratus*), platyfish (*Xiphophorus maculatus*), lamprey (*Petromyzon marinus*), Pacific oyster (*Crassostrea gigas*), and eastern oyster and shark (*Ginglymostoma cirratum*). These libraries have provided useful resources for different areas of genomics, molecular genetics, and molecular biology research.

Applications of LBC Libraries in Genomics and Molecular Research

LBC libraries have been widely used in many aspects of genomics, genetics, and biological research. Several examples are listed below to demonstrate the utility of LBC libraries in genomics and molecular research.

Positional or Map-based Cloning of Genes and QTLs

Positional or map-based cloning is a unique approach to cloning genes or QTLs of interest known only by phenotype. Targeted genes or QTLs are first genetically mapped and fine mapped with DNA markers, and then approached by the process of chromosome walking from the flanking markers using a LBC library. The LBC library is screened with the flanking DNA markers, positive clones are identified, an overlapping clone contig is constructed from the positive clones, and the contig end that is closer to the targeted gene or QTL is isolated and used as a probe to continue screening the library until a LBC spanning the gene or QTL is isolated. This process is called chromosome walking. It is apparent that the insert size of the LBC library is of significance to success of the chromosome walk; when a larger-insert DNA library is used, fewer steps, with a higher probability, are needed to complete the chromosome walk (Zhang et al. 1996). Using this approach, a number of genes and QTLs have been cloned. Readers are referred to the following works for good examples: Patocchi and others (1999), Deng and others (2001), El-Din El-Assal and others (2001), Takahashi and others (2001), and Yan and others (2004).

Physical Mapping

Genome physical mapping with LBCs allows the reconstruction of chromosomes or genomes from LBC libraries in the physical order of the clones derived from the chromosomes or genomes. To this end, the LBC fingerprinting and contig assembly has emerged as the method of choice for construction of physical maps from LBCs (Wu et al. 2005). LBC DNA is isolated, digested and/or end-labeled, and fractionated on solid matrices, including agarose gels (Marra et al. 1997), polyacrylamide gels for manual (Zhang and Wing 1997) or automated (Gregory et al. 1997; Ding et al. 1999, 2001) sequencers, and capillary arrays for capillary sequencers (Xu et al. 2004). The physical map of the genome is assembled from the LBC fingerprints using the

computer software FingerPrint Contig (FPC) (Sulston et al. 1988; Soderlund et al. 1997, 2000) according to the clone fingerprint similarities. See Chapter 14. Whole-genome BAC and/or BIBAC-based physical maps have been constructed from a number of species, including *Arabidopsis* (Marra et al. 1999, Chang et al. 2001), *Drosophila* (Hoskins et al. 2000), human (International Human Genome Mapping Consortium 2001), rice (Tao et al. 2001, Chen et al. 2002), chicken (Ren et al. 2003, Wallis et al. 2004), soybean (Wu et al. 2004a), *Penicillium* (Xu et al. 2005), and *Phytophthora* (Zhang et al. 2006). These BAC and/or BIBAC-based physical maps have provided a central platform for many aspects of genome research. A number of BAC-based physical maps have been constructed with fish including that of Atlantic salmon (Ng et al. 2005), tilapia (Katagiri et al. 2005), and channel catfish (Xu et al. 2007).

Genome Sequencing

LBCs have provided a powerful tool for large-scale genome sequencing. For review, see Zhang and Wu (2001). Two strategies have been used for genome sequencing (Zhang and Wu 2001). See Chapter 26, for a review of the clone-by-clone shotgun and the whole-genome shotgun. In the clone-by-clone shotgun approach, LBCs are first used to construct a physical map of the targeted genome. The LBCs in the minimal tiling path of the physical map are selected and used as templates for small-insert shotgun subclone library construction and sequencing. In the whole-genome shotgun approach, genomic DNA libraries with different insert sizes, including 1–2 kb, 10–20 kb, and 100–200 kb, are constructed and used as templates for sequencing from both ends of each clone. Of these libraries, the 1–2 kb library is for sequence production and sequence contig assembly, the 10–20 kb library is used to link the sequence contigs into sequence scaffolds, and the 100–200 kb LBC library is used to link the sequence contigs into sequence scaffolds and the scaffolds into superscaffolds that are an essential step for assembly of chromosome-wide sequences. To facilitate the genome sequence assembly, the LBC library is also used to construct a physical map of the genome as described above. For instance, the genome sequences of *Arabidopsis* (Arabidopsis Genome Initiative 2000), *Drosophila* (Adams et al. 2000), human (International Human Genome Sequencing Consortium 2001), chicken (International Chicken Genome Sequencing Consortium 2004), rice (International Rice Genome Sequencing Project 2005), and *Phytophthora* (Tyler et al. 2006) are all generated and/or assembled based on LBC-based physical maps.

Long-range Genome Comparative Analysis

LBCs have made it possible to comparatively analyze long-range genomic regions of interest from different species or genotypes, including centromeric regions. LBCs derived from the genomic regions of different species are isolated from LBC libraries, sequenced, and comparatively analyzed. For instances, see Chen and others (1997), Nagaki and others (2004), Wu and others (2004c), and Zhang and others (2004).

Chromosome Imbalance Analysis Using LBC-based Microarrays

LBCs have been used as elements to fabricate genome-wide microarrays for analysis of chromosome imbalance, including DNA copy number changes, single-copy sequence deletion/insertion, genomic amplification, and chromosome instability. LBCs in the minimal tilling path of a physical map are selected, LBC DNA is prepared and printed on glass slides, and the resultant microarrays are hybridized with genomic DNA isolated from different genotypes by a technique named the Array Comparative Genome Hybridization (CGH). For examples, see Chung and others (2004) and Ishkanian and others (2004).

Targeting the Genomic Region of Interest Using User-friendly DNA Markers

LBCs have been used to develop DNA markers, such as SSRs (simple sequence repeats or microsatellites) and SNPs (single nucleotide polymorphism), for fine mapping of a particular genomic region or for genome-wide integrative genetic and physical mapping. LBC libraries are screened with SSR synthetic oligos to identify the LBCs containing one or more SSR loci, the positive clones are subcloned and rescreened with the SSR oligos, and the SSR-containing subclones are sequenced and used to design SSR primers. For instances, see Cregan and others (1999) and Lichtenzveig and others (2005).

Functional Analysis of Long-range Genomic Sequences

Of the LBCs, BIBACs, binary PBCs, and TACs can be directly transformed into plants for functional analysis of the cloned DNA fragments in addition to their large fragment cloning capacity as BACs and PACs. Techniques have been developed to transform LBCs in several plant species, including tobacco (Hamilton et al. 1996), tomato (Hamilton et al. 1999), *Arabidopsis* (Liu et al. 1999), and rice (Liu et al. 2002, He et al. 2003). The transformation-competent BIBAC, binary PBC, and TAC libraries not only streamline the positional cloning of genes and QTLs, but also facilitate the genetic engineering of individual genes and QTLs, multiple genes and/or QTLs, and gene clusters for genetic improvement and functional analysis of long-range genomic sequences.

The LBC or megabase recombinant DNA technology represents one of the major technologies resulting from genomics and DNA research in the past decade. Although a large number of arrayed LBC libraries have been constructed from a variety of species or genotypes of different organisms to facilitate different research purposes, new LBC libraries will be needed for different species or genotypes, especially for those of aquaculture and marine species, and economically important microbes. As arrayed LBC libraries have been resources essential for different aspects of genomics and DNA research, it is predicted that they will continue to play a significant role in advanced genomics research.

References

Adams MD, SE Celniker, RA Holt, CA Evans, JD Gocayne, et al. 2000. The genome sequence of *Drosophila melanogaster*. Science, 287, pp. 2185–2195.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature, 408, pp. 796–815.

Burke DT, GF Carle, and MV Olson. 1987. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. Science, 236, pp. 806–812.

Canilleri C, J Lafleuriel, C Macadre, F Varoquaux, Y Parmentier, et al. 1998. A YAC contig map of *Arabidopsis thaliana* chromosome 3. Plant J, 14, pp. 633–642.

Chang Y-L, Q Tao, C Scheuring, K Meksem, and H-B Zhang. 2001. An integrated map of *Arabidopsis thaliana* for functional analysis of its genome sequence. Genetics, 159, pp. 1231–1242.

Chen M, G Presting, WB Barbazuk, JL Goicoechea, B Blackmon, et al. 2002. An integrated physical and genetic map of the rice genome. Plant Cell, 14, pp. 537–545.

Chen M, P San Miguel, AC de Oliveira, S-S Woo, H-B Zhang, et al. 1997. Microcolinearity in *sh2*-homologous regions of the maize, rice and sorghum genomes. Proc. Natl Acad Sci USA, 94, pp. 3431–3435.

Chumakov IM, P Rigault, IL Gall, C Bellanne-Chantelot, A Billault, et al. 1995. A YAC contig map of the human genome. Nature, 377, pp. 175–297.

Chung Y-J, J Jonkers, H Kitson, H Fiegler, S Humphray, et al. 2004. A whole-genome mouse BAC microarray with 1-Mb resolution for analysis of DNA copy number changes by array comparative genomic hybridization. Genome Res, 14, pp. 188–196.

Clarke L and J Carbon. 1976. A colony bank containing synthetic ColE1 plasmid reprehensive of the entire *E. coli* genome. Cell, 9, pp. 91–100.

Cregan PB, J Mudge, EW Fickus, LF Marek, D Danesh, et al. 1999. Targeted isolation of simple sequence repeat markers through the use of bacterial artificial chromosomes. Theor Appl Genet, 98, pp. 919–928.

Deng Z, S Huang, P Ling, C Yu, Q Tao, et al. 2001. Fine genetic mapping and BAC contig development for the citrus tristeza virus resistance gene locus in *Poncirus trifoliata* (Raf.). Mol Genet Genomics, 265, pp. 739–774.

Ding Y, MD Johnson, WQ Chen, D Wong, Y-J Chen, et al. 2001. Five-color-based high-information-content fingerprinting of bacterial artificial chromosome clones using Type IIS restriction endonucleases. Genomics, 74, pp. 142–154.

Ding Y, MD Johnson, R Colayco, YJ Chen, J Melnyk, et al. 1999. Contig assembly of bacterial artificial chromosome clones through multiplexed fluorescence-labeled fingerprinting. Genomics, 56, pp. 237–246.

El-Din El-Assal S, C Alonso-Blanco, AJM Peeters, V Raz, and M Koornneef. 2001. A QTL for flowering time in *Arabidopsis* reveals a novel allele of *CRY2*. Nature Genet, 29, pp. 435–440.

Frengen E, D Weichenhan, B Zhao, K Osoegawa, M van Geel, and PJ de Jong. 1999. A modular, positive selection bacterial artificial chromosome vector with multiple cloning sites. Genomics, 58, pp. 250–253.

Frijters ACJ, Z Zhang, M van Damme, G-L Wang, PC Ronald, and RW Michelmore. 1997. Construction of a bacterial artificial chromosome library containing large *Eco* RI and *Hin*d III genomic fragments of lettuce. Theor Appl Genet, 94, pp. 390–399.

Fu H and HK Dooner. 2000. A Gene-enriched BAC Library for Cloning Large Allele-specific Fragments from Maize: Isolation of a 240-kb Contig of the *bronze* Region. Genome Res, 10, pp. 866–873.

Gregory SG, GR Howell, and DR Bentley. 1997. Genome mapping by fluorescent fingerprinting. Genome Res, 7, pp. 1162–1168.

Hamilton CM, A Frary, C Lewis, and SD Tanksley. 1996. Stable transfer of intact high molecular weight DNA into plant chromosomes. Proc Natl Acad Sci USA, 93, pp. 9975–9979.

Hamilton CM, A Frary, Y Xu, SD Tanksley, and H-B Zhang. 1999. Construction of tomato genomic DNA libraries in a binary-BAC (BIBAC) vector. Plant J, 18, pp. 223–229.

He R-F, Y Wang, Z Shi, X Ren, L Zhu, Q Weng, G-C He. 2003. Construction of a genomic library of wild rice and *Agrobacterium*-mediated transformation of large-insert DNA linked to BPH resistance locus. Gene, 321, pp. 113–121.

Hess G. 2005. Toward positional cloning of everblooming gene (*evb*) in plants: a BAC library of *Rosa chinensis* cv. Old Blush. Thesis (M.S.), Texas A&M University, College Station.

Hoskins RA, CR Nelson, BP Berman, TR Laverty, RA George, et al. 2000. A BAC-based physical map of the major autosomes of *Drosophila melanogaster*. Science, 287, pp. 2271–2274.

International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature, 432, pp. 695–716.

International Human Genome Mapping Consortium. 2001. A physical map of the human genome. Nature, 409, pp. 934–941.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. Nature, 409, pp. 860–921.

International Rice Genome Sequencing Project. 2005. The map-based sequence of the rice genome. Nature, 436, pp. 793–800.

Ioannou I, CT Amemiya, J Garnes, PM Kroisel, H Shizuya, et al. 1994. A new bacteriophage P1-derived vector for propagation of large human DNA fragments. Nature Genet, 6, pp. 84–89.

Ishkanian AS, CA Malloff, SK Watson, RJ deLeeuw, B Chi, et al. 2004. A tiling resolution DNA microarray with complete coverage of the human genome. Nature Genet, 36, pp. 299–303.

Jones JDG, L Shlumukov, F Carland, J English, SR Scofield, et al. 1992. Effective vectors for transformation, expression of heterologous genes, and assaying transposon excision in transgenic plants. Transgenic Res, 1, pp. 285–297.

Katagiri T, C Kidd, E Tomasino, JT Davis , C Wishon, et al. 2005. A BAC-based physical map of the Nile tilapia genome. BMC Genomics, 6, p. 89.

Kim U-J, BW Birren, T Slepak, V Mancino, C Boysen, et al. 1996. Construction and characterization of a human bacterial artificial chromosome library. Genomics, 34, pp. 213–218.

Kurata N, Y Umehara, H Tanoue, and T Sasaki. 1997. Physical mapping of the rice genome with YAC clones. Plant Mol Biol, 35, pp. 101–113.

Lichtenzveig J, C Scheuring, J Dodge, S Abbo, and H-B Zhang. 2005. Construction of BAC and BIBAC libraries and their applications for generation of SSR markers for genome analysis of chickpea, *Cicer arietinum* L. Theor Appl Genet, 110, pp. 492–510.

Liu Y-G, H Liu, L Chen, W Qiu, Q Zhang, et al. 2002. Development of new transformation-competent artificial chromosome vectors and rice genomic libraries for efficient gene cloning. Gene, 282, pp. 247–255.

Liu Y-G, Y Shirano, H Fukaki, Y Yanai, M Tasaka, et al. 1999. Complementation of plant mutants with large genomic DNA fragments by a transformation-competent artificial chromosome vector accelerates positional cloning. Proc Natl Acad Sci USA, 96, pp. 6535–6540.

Marra M, T Kucaba, NL Dietrich, ED Green, B Brownstein, et al. 1997. High throughput fingerprint analysis of large-insert clones. Genome Res, 7, pp. 1072–1084.

Marra M, T Kucaba, M Sekhon, L Hiller, R Martienssen, et al. 1999. A map for sequence analysis of the *Arabidopsis thaliana* genome. Nature Genet, 22, pp. 265–270.

Nagaki K, Z Cheng, S Ouyang, PB Talbert, M Kim, et al. 2004. Sequencing of a rice centromere uncovers active genes. Nature Genet, 36, pp. 138–145.

Ng SHS, CG Artieri, IE Bosdet, R Chiu, RG Danzmann, et al. 2005. A physical map of the genome of Atlantic salmon, *Salmo salar*. Genomics, 86, pp. 396–404

Nusbaum C, DK Slonim, KL Harris, BW Birren, RG Steen, et al. 1999. A YAC-based physical map of the mouse genome. Nature Genet, 22, pp. 388–393.

Patocchi A, BA Vinatzer, S Gianfranceschi, H-B Zhang, S Sansavini, and C Gessler. 1999. Construction of a 550-kb BAC contig spanning the genomic region containing the apple scab resistance gene *Vf*. Mol Gen Genet, 262, pp. 884–891.

Ren C, M-K Lee, B Yan, K Ding, B Cox, et al. 2003. A BAC-based physical map of the chicken genome. Genome Res, 13, pp. 2754–2758.

Ren C, ZY Xu, S Sun, M-K Lee, C Wu, et al. 2005. Genomic DNA Libraries and Physical Mapping. In: Meksem K, Kahl G, Eds. The Handbook of Plant Genome Mapping: Genetic and Physical Mapping. Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 173–213.

Saji S, Y Umehara, BA Antonio, H Yamane, H Tanoue, et al. 2001. A physical map with yeast artificial chromosome (YAC) clones covering 63% of the 12 rice chromosomes. Genome, 44, pp. 32–37.

Schmidt R, J West, K Love, Z Lenehan, C Lister, et al. 1995. Physical map and organization of *Arabidopsis thaliana* chromosome 4. Science, 27, pp. 480–483.

Shizuya H, B Birren, UJ Kim, V Mancino, T Slepak, et al. 1992. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. Proc Natl Acad Sci USA, 89, pp. 8794–8797.

Soderlund C, S Humphray, A Dunham, and L French. 2000. Contigs built with fingerprints, markers, and FPC V4.7. Genome Res, 10, pp. 1772–1787.

Soderlund C, I Longden, and R Mott. 1997. FPC: a system for building contigs from restriction fingerprinted clones. CABIOS, 13, pp. 523–535.

Sulston J, F Mallett, R Staden, R Durbin, T Horsnell, and A Coulson. 1988. Software for genome mapping by fingerprinting techniques. CABIOS, 4, pp. 125–132.

Takahashi Y, A Shomura, T Sasaki, and M Yano. 2001. *Hd6*, a rice quantitative trait locus involved in photoperiod sensitivity, encodes the a subunit of protein kinase CK2. Proc Natl Acad Sci USA, 98, pp. 7922–7927.

Tao Q, Y-L Chang, J Wang, H Chen, C Scheuring, et al. 2001. Bacterial artificial chromosome-based physical map of the rice genome constructed by restriction fingerprint analysis. Genetics, 158, pp. 1711–1724.

Tao Q and H-B Zhang. 1998. Cloning and stable maintenance of DNA fragments over 300 kb in *Escherichia coli* with conventional plasmid-based vectors. Nucleic Acids Res, 26, pp. 4901–4909.

Tyler BM, S Tripathy, X Zhang, P Dehal, RHY Jiang, et al. 2006. *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. Science, 313, pp. 1261–1266.

Wallis JW, J Aerts, MAM Groenen, RPMA Crooijmans, D Layman, et al. 2004. A physical map of the chicken genome. Nature, 432, pp. 761 – 764.

Wu C, S Sun, M-K Lee, ZY Xu, C Ren, and H-B Zhang. 2005. Whole genome physical mapping: An overview on methods for DNA fingerprinting. In: Meksem K, Kahl G, Eds. The Handbook of Plant Genome Mapping: Genetic and Physical Mapping. Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 257–284.

Wu C, S Sun, N Padmavathi, FA Santos, R Springman, et al. 2004a. A BAC and BIBAC-based physical map of the soybean genome. Genome Res, 14, pp. 319–326.

Wu C, Z Xu, and H-B Zhang. 2004b. DNA Libraries. In: Meyers RA, Ed. Encyclopedia of Molecular Cell Biology and Molecular Medicine. Vol. 3 (2nd Edition). Wiley-VCH Verlag GmbH, Weinheim, Germany, pp. 385–425.

Wu J, H Yamagata, M Hayashi-Tsugane, S Hijishita, M Fujisawa, et al. 2004c. Compositon and structure of the centromeric region of rice chromosome 8. The Plant Cell, 16, pp. 967–976.

Xu P, S Wang, L Liu, J Thorsen, H Kucuktas, et al. 2007. A BAC-based physical map of the channel catfish genome. Genomics, in press.

Xu Z, S Sun, L Covaleda, K Ding, A Zhang, et al. 2004. Genome physical mapping with large-insert bacterial clones by fingerprint analysis: methodologies, source clone genome coverage and contig map quality. Genomics, 84, pp. 941–951.

Xu Z, M van den Berg, C Scheuring, L Colaveda, H Lu, et al. 2005. Genome-wide physical mapping from large-insert clones by fingerprint analysis with capillary electrophoresis: A robust physical map of *Penicillium chrysogenum*. Nucleic Acids Res, 33, p. e50.

Yan L, A Loukoianov, A Blechl, G Tranquilli, W Ramakrishna, et al. 2004. The wheat *VRN2* gene is a flowering repressor down-regulated by vernalization. Science, 303, pp. 1640–1644.

Zachgo EA, ML Wang, J Dewney, D Bouchez, C Camilleri, et al. 1996. A physical map of chromosome 2 of Arabidopsis thaliana. Genome Res, 6, pp. 19–25.

Zhang H-B. 2000. Construction and Manipulation of Large-insert Bacterial Clone Libraries—Manual. Texas A&M University College Station,, Texas, USA.

Zhang H-B and RA Wing. 1997. Physical Mapping of the rice genome with BACs. Plant Mol Biol, 35, pp. 115–127.

Zhang H-B, S-S Woo, and RA Wing. 1996. BAC, YAC and Cosmid Library Construction. In: Foster G, Twell D, Eds. Plant Gene Isolation: Principles and Practice. John Wiley & Sons Ltd., England, pp. 75–99.

Zhang H-B and C Wu. 2001. BACs as tools for genome sequencing. Plant Physiol Biochem, 39, pp. 195–209.

Zhang H-B, X-P Zhao, X-L Ding, AH Paterson, and RA Wing. 1995. Preparation of megabase-size DNA from plant nuclei. Plant J, 7, pp. 175–184.

Zhang X, C Scheuring, S Tripathy, Z Xu, C Wu, et al. 2006. An integrated BAC and genome sequence physical map of *Phytophthora sojae*. Molecular Plant-Microbe Interactions, 19, pp. 1302–1310.

Zhang Y, Y Huang, L Zhang, Y Li, T Lu, et al. 2004. Structural features of the rice chromosome 4 centromere. Nucleic Acids Res, 32, pp. 2023–2030.

Zhao X-P, H-B Zhang, RA Wing, and AH Paterson. 1994. A simple method for cotton megabase DNA isolation. Plant Mol Biol Rep, 12, pp. 126–131.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 14
Bacterial Artificial Chromosome Libraries and BAC-based Physical Mapping of Aquaculture Genomes

William S. Davidson

Introduction

The availability of large-insert genomic libraries is considered an essential resource for genomic analysis, especially of complex genomes. The construction of libraries using yeast artificial chromosomes (YAC) allowed the cloning of fragments >500 kilobase pairs (kb) (Burke et al. 1987, Larin et al. 1991), but it was found that a significant proportion of these fragments were the result of co-cloning noncontiguous DNA fragments (i.e., chimeric clones). Moreover, YAC clones are often unstable, and it is difficult to obtain sufficient quantities of pure DNA for analysis by restriction digestion or for the production of small fragment shotgun libraries that are required for DNA sequencing. Bacterial artificial chromosome (BAC) (Shizuya et al. 1992, Osoegawa et al. 1998) and P1-derived artificial chromosome (PAC) (Sternberg et al. 1990, Ioannou et al. 1994) cloning vectors provide a better alternative for vertebrate genome mapping and sequencing. Also see Chapter 13. Although these bacterial clones carry smaller inserts compared to YACs, they are more stable and the recombinant DNA can be isolated using simple plasmid extraction methods. BAC libraries, and the physical maps that are derived from them, have become the mainstay for vertebrate genome sequencing projects. But the BAC clones offer many other opportunities for the genomics community. For example, they can be used for the isolation of specific genetic markers and probes; they allow regions of the genome that contain quantitative trait loci (QTL) to be surveyed for candidate genes; and they facilitate comparative and evolutionary genomic studies. When a series of BACs that cover the entire genome has been identified, this enables the production of arrays that can be used for comparative genomic hybridization (CGH) analysis to identify rearrangements in the genomes of individuals and closely related species. In this chapter, I will focus on how BAC-based physical maps are constructed and how the integration of physical and linkage maps is achieved. The details for the construction of BAC libraries were described in Chapter 13, so here I will only provide a brief discussion as it pertains to the issues related to physical mapping and genome analysis.

Preparation of BAC Libraries

BAC Cloning Vectors

DNA cloned into plasmid vectors based on medium to high copy number replicons is often deleted or becomes rearranged, particularly if the DNA comes from eukaryotic

organisms whose genomes contain families of repetitive sequences. Therefore, to facilitate the construction of large-insert libraries from complex eukaryotic genomes BAC vectors were developed that would be present as a single copy per cell (Shizuya et al. 1992). This is achieved by incorporating the F factor that encodes four genes (*oriS*, *repE*, *parA*, and *parB*) that regulate its replication and control its copy number. The *oriS* and *repE* genes mediate the unidirectional replication of the F factor while *parA* and *parB* maintain copy number at a level of one to two per *E. coli* genome. The cloning site is usually flanked by T7 and Sp6 promoters that can be used to generate BAC-end sequences, and rare-cutting restriction enzyme sites (e.g., *Not*I) that allow the insert DNA to be excised intact or in a few very large fragments. A gene conferring drug resistance (e.g., to chloramphenicol or ampicillin) is present in the BAC vector to permit selection for transformants. In order to be able to use positive selection for the presence of recombinant clones, a SacBII gene that has a "pUClink" stuffer fragment between its promoter and the coding region has been introduced into vectors such as pBACe3.6 and the pTARBAC series (Frengen et al. 1999, Osoegawa et al. 2004). This stuffer fragment is removed prior to ligating in the desired genomic DNA. A cloning vector that self-ligates produces an active *Sac*BII, and the gene product, levansucrase, converts saccharose to levan, which is toxic to the host *E. coli*, whereas in hybrid recombinant clones the insert will keep *Sac*BII in an inactive state. Other BAC cloning vectors (e.g., pKS145) make use of blue/white colony selection through inactivation of the beta-galactosidase gene in the recombinants (Fujiyama et al. 2002). The size of BAC vectors ranges from 7 to 14 kb, so they are small compared to the DNA inserts of 100 to 220 kb.

Selection of Genomic DNA for Cloning

The quality of a BAC library depends on being able to prepare high molecular weight DNA. The tissue of choice should yield intact cell suspensions that can be embedded in agarose plugs. The DNA is prepared in situ by treating the plugs with a detergent and Proteinase K. The agarose plugs with the embedded DNA is then washed extensively, and partial digestion with a restriction enzyme is carried out to generate fragments that are in the range of 200 to 250 kb. The fragments are separated by pulse field gel electrophoresis, recovered, and then ligated into the BAC vector. See Osoegawa and others (1998) for experimental details.

Although obtaining high molecular weight DNA is essential for a good quality BAC library, it is equally important to consider the source of the DNA. The cloned DNA should represent a "typical genome." DNA that is prepared from vertebrate white blood cells or other specialized cells, may contain rearrangements (e.g., in immunoglobulin genes), and could cause problems in future studies. Similarly, the amount of genetic variation in the donor of the DNA should be at a minimum, otherwise it can prove difficult to assemble contigs from BAC fingerprints. It is relatively easy to obtain DNA from inbred strains of model organisms or domesticated animals, and making an effort to do this will pay dividends. For fish species, the production of double haploids is particularly useful (Parsons and Thorgaard 1985, Corley-Smith et al. 1996), especially if the clonal line can be maintained and used to produce other genomic resources such as EST libraries and genomic DNA for whole shotgun library sequencing. Finally, one

should consider the desirability to have DNA from the heterogametic sex represented in the library. In the XY sex determination system, for instance, the use of a male as the source of DNA should provide not only the autosomes, but both the X and Y chromosomes though the X and Y are represented at half the amount.

Construction and Arraying of BAC Library

The procedures for constructing BAC libraries have been described in great detail in Chapter 13 (Osoegawa et al. 1998), and they will not be repeated here. BAC libraries are maintained in 384 well plates, usually with a master plate from which additional copies may be produced. In addition, the clones are printed in duplicate on nylon membranes that can be screened by hybridization with oligonucleotide probes to identify BACs containing sequences of interest. Pools of DNA are sometimes prepared from BAC libraries so that there are plate pools (representing 384 clones) and row and column pools from each plate (representing 24 and 16 clones, respectively). More complicated pooling strategies have also been used (Klein et al. 2000). Although time consuming and rather laborious to prepare, these pooled DNA samples allow researchers to screen the BAC library by PCR for genetic markers, and this is an efficient way to join a physical map with a linkage map.

BAC libraries are available for a variety of aquaculture finfish including Atlantic salmon (Thorsen et al. 2005), carp (Katagiri et al. 2001), catfish (Quiniou et al. 2003), Japanese flounder (Katagiri et al. 2000), rainbow trout (Katagiri et al. 2001, Palti et al. 2004), red sea bream (Katagiri et al. 2002), and tilapia (Katagiri et al. 2001).

Characterization of BAC Libraries

Fold of Genome Coverage

It is important to characterize the BAC library with respect to insert size and overall coverage of the genome. The size distribution of clone insert sizes is determined by digesting a subset of the BAC clones with a rare-cutting restriction enzyme with sites on either side of the cloning site (e.g., *Not*I) and analyzing the digested DNA by pulse field gel electrophoresis. From the knowledge of the average size of the inserts, the number of clones in the library, and the size of the genome, it is possible to estimate the fold coverage of the genome that the library represents. It is obvious that the larger the average insert size, the fewer the clones that are required to give a particular fold genome coverage. For example, it requires approximately 17,000 BACs with an average insert size of 180 kb to cover the Atlantic salmon's genome once (genome size of 3 Mb).

Representation of the Genome

To gain an estimate of how well a BAC library covers the entire genome, it is normal to screen the library for known genes or genetic markers and to determine how often

clones containing them occur. Marker content assessment is conducted by screening the nylon membranes containing the arrayed BACs by hybridization with oligonucleotide probes (Ross et al. 1999). Data production can be accelerated if the probes are hybridized in sets according to a two-dimensional pooling scheme (Khorasani et al. 2004). Alternatively, one can use marker-specific primers and polymerase chain reaction (PCR) analysis of DNA templates from pooled BAC DNA prepared according to pooling schemes that reduce the amplification effort (Crooijmans et al. 2000). Segments of a genome are sometimes not represented in a BAC library. This may be a function of the enzyme used to create the library or simply unstable sequences that are recalcitrant to cloning. For this reason it is desirable to prepare multiple BAC libraries for a species, preferably using the same source of genomic DNA and different restriction enzymes.

Construction of Physical Maps

A physical map is an ordered set of DNA fragments that aims to cover the entire genome. BACs are the preferred building blocks of physical maps, and in today's context a physical map comprises a set of ordered, overlapping BAC clones. The goal is to identify the smallest number of BACs required to represent the genome (i.e., the minimal tiling path). Sequencing these BACs should provide the basis for whole genome sequencing. The most frequent approach to constructing a BAC-based physical map involves DNA fingerprinting. Although several methods have been developed, all are variations on the theme first used to construct a cosmid-based physical map of the nematode, *C. elegans* (Coulson et al. 1986).

Fingerprinting Cosmid Clones

The fingerprinting of *C. elegans* cosmid clones was achieved by extracting the DNA using a simple alkaline SDS procedure, digesting it with a six-base pair (bp) recognition restriction enzyme (*Hin*dIII), attaching a radioactive label to the *Hin*dIII ends, digesting the DNA again but this time with a four-bp recognition enzyme (*Sau*3A1), and then separating the relatively small fragments by electrophoresis through denaturing polyacrylamide gels (Coulson et al. 1986). The banding patterns were visualized by autoradiography, and the sizes of the bands were determined against size standards. It should be noted that not all fragments produced by the double digest are observed in this procedure; rather, it is restricted to those with a radioactive tag at a *Hin*dIII site. The digestion of a nematode cosmid clone using a combination of *Hin*dIII and *Sau*3A1 yielded on average 23 labeled bands.

Fingerprinting Using One Enzyme and Agarose Gels

The procedure devised by Coulson and others (1986) for fingerprinting cosmid clones can be described as using two restriction enzymes and separation of the fragments by

polyacrylamide gel electrophoresis. Marra and others (1997) developed a procedure for the high throughput analysis of BAC clones that uses a single restriction enzyme and separation by agarose gel electrophoresis. This method has several advantages over the procedure described above for cosmid clones. First, only one restriction enzyme (*Hin*dIII) is used; second, the procedure does not involve radioactivity; third, separation of the fragments is carried out using agarose gels rather than polyacrylamide gels; and fourth, the detection of the DNA fragments is by SYBR Green staining, which allows essentially all fragments to be identified. The predictable signal intensity of the bands as a function of fragment size lends itself to fully automated band calling (Fuhrmann et al. 2003), and this provides an estimate of the insert size of the BAC clone based on summing the fragments. The number of detectable bands produced by the method of Marra and others (1997) for BAC clones is comparable to that of Coulson and others (1986) for cosmid clones.

Although the single enzyme digestion and agarose gel separation system continues to be used extensively to construct physical maps for a wide variety of organisms, several other methods have been developed that involve more than one restriction enzyme, differential end-labeling, or partial sequencing of the ends of fragments and separation of DNA fragments on automated DNA sequencers.

Fingerprinting Using Five Enzymes and Differential End-labeling

Fingerprinting BAC clones with the one enzyme system (Marra et al. 1997) requires extensive overlap between adjacent clones to assemble contigs with high confidence. It was reasoned that if more information could be obtained about the fingerprinted fragments, then it would be possible to identify BACs that have smaller overlaps. The combination of *Hin*dIII and *Sau*3A1 used by Coulson and others (1986) generates many small fragments, and should in theory be able to provide the information necessary to detect smaller overlaps in BACs. However, the sheer number of fragments produced when a large BAC clone is digested by this means makes it difficult to differentiate among the fragments. This problem is overcome using the SNaPshot labeling kit and sizing the fragments by capillary electrophoresis (Luo et al. 2003).

In this procedure the BAC DNA is digested with four six-bp recognition enzymes that each leaves a different 3′ recessed end. This allows each of these restriction cut sites to be differentially labeled using ddNTPs with different fluorescent dye labels and Ampli*Taq* FS polymerase (*Eco*RI with ddATP; *Bam*HI with ddGTP; *Xba*I with ddCTP; and *Xho*I with ddTTP). *Hae*III, which has a four-bp recognition site and yields blunt ends, is included in the initial digest to give fragments in the range of 50 to 500 bps that are amenable to separation on an automated sequencer. See Figure 2 in Luo and others (2003).

Fingerprinting Using Two Enzymes and Partial Sequencing

It was reasoned that if the sequences of the ends of small fingerprint fragments were known, then this would provide even more information on which to base overlaps

between pairs of BACs (Brenner and Livak 1989). This is the rationale for the five-color-based high-information-content fingerprinting approach (Ding et al. 2001).

In this system each fragment is characterized by both size and the end terminal sequence of one to five nucleotides. The BAC DNA is digested with *Rsa*I, a four-bp recognition enzyme that generates blunt ends, and *Hga*I, which cuts DNA five bps away from its recognition site leaving a 5′ overhang consisting of five unknown bases. The fragments that are generated by this double digest are then subjected to a DNA sequencing reaction in which the ddNTPs each have a different fluorescent label. This fills in the 5′ overhang at *Hga*I ends. The fragments are then separated using a DNA sequencer. The patterns appear as groups of restriction fragments with each group being the result of the sequencing reaction. See Figure 2 in Ding and others (2001).

Comparison of Fingerprinting Methods

The purpose of fingerprinting is to identify clones that overlap so that they can be grouped into contigs. The effectiveness of this process will depend on the number of bands shared by a pair of clones, and this in turn will be a function of the total number of bands generated by the fingerprinting procedure. This latter point is at the heart of comparing different methods that have been developed for analyzing BAC libraries. However, it should also be realized that other factors such as cost and throughput must be taken into account when choosing which fingerprinting approach to use.

The pros and cons of the different fingerprinting procedures have been reviewed recently, with simulations being made to assist in determining which is the most effective (Meyers et al. 2004, Xu et al. 2004). The agarose fingerprinting method is widely used because of its relative simplicity, its low cost, and its proven effectiveness. It is the only method that detects the vast majority of fragments from a clone. This provides important information for the incorporation of BAC-end sequence data into genome sequences, and for verifying the accuracy of sequence assembly. However, this procedure is probably not the best for producing a high-resolution physical map.

The fluorescence-based fingerprinting methods have technical advantages that enable them to produce more detectable bands, and it is generally considered that an increase in band number corresponds to an increase in information content. In general, methods that produce a greater number of bands can detect overlaps more efficiently than methods that produce fewer bands (Meyers et al. 2004). Labeling the fragments with different colors increases the information content, as does having the partial sequence of a fragment's end. A simulation carried out by Meyers and others (2004) revealed that the procedure of Ding and others (2001) gave the largest number of informative fragments followed by the SNaPshot labeling method (Luo et al. 2003) and finally the methods of Coulson and others (1986) and Marra and others (1997).

It should be noted that analyzing a BAC library with very large BAC insert sizes is far more beneficial to the single enzyme, agarose system (Marra et al. 1997) than it is to the other procedures. A general rule of thumb is that for this procedure to be effective, the average insert size should be >150,000 bps. Two–thirds of the BACs that were not incorporated into the Atlantic salmon physical map gave fewer than 16 *Hin*dIII fragments, whereas the proportion of BACs that assembled into contigs rose

dramatically as the number of *Hin*dIII fragments increased above 25. See Figure 3 in Ng and others (2005).

Another variable that should be considered before starting any BAC fingerprinting is the base composition of the genome to be analyzed. This can influence the choice of method and enzyme(s) to be used. For example, the procedure of Ding and others (2001) uses *Hga*I, which has GACGC as its recognition site sequence. The high G+C composition of the rice genome and its low CpG suppression compared to the low frequency of CpG in human DNA, enabled this procedure to produce approximately fivefold more bands from rice BACs than were obtained from human BAC clones with similar sized inserts (Meyers et al. 2004). Similarly, it is important to avoid restriction enzymes that have two or more cut sites in highly repetitive elements in the genome to be mapped.

Assembly of Physical Maps

The first stage in constructing a physical map from fingerprint data is to transform the fragment sizes and related information into a data set that can be recognized by the automated physical map assembly software, FingerPrint Contig (FPC) (Soderlund et al. 1997). For fragments separated on agarose gels, this involves using the Image software (www.sanger.ac.uk/Software/Image, Sulston et al. 1988) and BandLeader (Fuhrmann et al. 2003), whereas GenoProfiler (wheat.pw.usda.gov/PhysicalMapping) is used to process the SNaPshot data from ABI genetic analyzers (Luo et al. 2003). There does not seem to be a convenient method at present to capture the GeneScan data produced by the method of Ding and others (2001) in a form that can be imported directly into FPC.

FPC considers fragments to be shared by two BAC clones if they have the same size within a given tolerance. The probability that two fragments are shared by chance between clones is calculated as a Sulston cutoff score (Sulston et al. 1988). During the assembly process in FPC, clones are binned together if they satisfy a user-defined cutoff value for fingerprint similarity based on the Sulston score. Automated assemblies are usually performed at high stringency to avoid false inclusions into contigs. These contigs are then manually edited to identify questionable (Q) clones that likely result from cross-contaminated wells in the library (i.e., those with insert sizes and/or numbers of fragments far in excess of the average for the library), and these are cancelled in the FPC database.

The final stage in building the physical map requires manual curation. The stringency of the assembly is relaxed as clones from the ends of contigs are tested to determine if they might overlap one another and so join contigs. In addition, singleton clones (i.e., those not incorporated into contigs) may be incorporated into the map at lower Sulston scores if there is additional information to support this. For example, the supporting data may come from genetic markers that have been mapped to these BACs.

Physical gaps result from segments of the genome that are not present in the BAC libraries. This can be mitigated by increasing the number of genome equivalents (usually to at least tenfold), and by using libraries constructed using different restriction enzymes. The quality of a physical map can be assessed by determining the number of BACs that are positive for known, single-locus probes, and also the proportion of these probes that give at least one positive hit. Ultimately, the test of a physical map

is if it can be integrated with a high-density linkage map, and colinearity is observed between the order of markers placed in the contigs of the physical map and on the linkage map.

Considerations for the Selection of Proper Tolerance and Sulston Cutoff Score

Tolerance deals with the flexibilities of size calling of restriction fragments. Due to technical fluctuations from gel run to run, the same fragment may be determined to have a different called size leading to the exclusion of the true overlapping fragments. Therefore, certain tolerance must be given to allow such technical fluctuations. The extent to which such a tolerance can be given depends on the types of fingerprinting. A tolerance of several bps up to 10 bp can be given to fingerprints from agarose gels. However, most fingerprinting techniques based on sequencing gels use a tolerance of 0.4–0.6 bp. Generally speaking, fingerprinting systems that naturally give more accurate calling of band size require the use of a low tolerance. However, the choice of tolerance is a dilemma where low tolerance may lead to true neighboring BAC clones to be excluded from being binned together in the same contig, while a high tolerance could lead to false contigs involving DNA segments from noncontinuous genomic regions.

Another factor used for contig construction is the Sulston score, which corresponds to the probability that two fingerprints share similar fragment patterns by chance. The lower the Sulston score, the more confident you can be that the overlapping BAC clones are really overlapping. The use of the stringency of the Sulston score is also a dilemma. Too low a Sulston score would lead to splitting of true contigs into multiple contigs or singletons, whereas too high a Sulston score would lead to false contigs. While a Sulston score of 3×10^{-12} was used for the human genome (3×10^9 bp) for automated assembly using FPC (McPherson et al. 2001), a larger score of 1×10^{-9} was used for the smaller *Arabidopsis* genome (Marra et al. 1999). The rice genome was assembled using a Sulston score of 1×10^{-12} (Krzywinski et al. 2004). The larger the genome, the lower the Sulston score should be used. In multicolor fluorescence-labeled fingerprinting, it is possible to use a low Sulston score because the number of fragments would be quadrupled (for four-enzyme, four-color labeling systems), and so should the shared fragments.

Integration of Physical and Linkage Maps

Time is well spent producing and verifying a robust physical map because it provides an invaluable tool for molecular studies on a genome. Although BAC fingerprinting places BACs into contigs, it neither gives the order of these contigs relative to one another nor their relationships to the genome as a whole. The utility of the physical map is greatly enhanced if it can be integrated with other genomic resources such as a high-density linkage map. Identifying contigs that correspond to genetic map assignments provides access to candidate gene regions for QTL and hence the raw material for gene-assisted selection protocols.

Mapping Anchor Loci from Genetic Maps

Microsatellite markers form the basis for most linkage maps because this type of marker has a high information content. Moreover, microsatellites are routinely used for QTL mapping, and thus they define regions of the genome that are of interest to breeders. The oligonucleotide primers used to amplify microsatellite loci often give amplification products in closely related species. Therefore, these markers can act as anchor loci for comparisons of genetic maps among species (Danzmann et al. 2005, Woram et al. 2003). It is desirable to make use of these anchor loci to orient the BAC contigs with respect to the linkage map. If DNA from the BAC library has been isolated and placed in pools (Crooijmans et al. 2000), PCR can be used with the microsatellite primers to identify which BAC(s) contain this segment of the genome. An alternative strategy must be adopted if pooled BAC DNA templates are not available. One approach is to design oligonucleotides from the flanking sequences of microsatellite loci and to use them as hybridization probes with the nylon filters that have DNA from the BAC colonies on them. Although this methodology should work well for genomes that are relatively devoid of repetitive elements, it has been our experience that the flanking regions of microsatellites from Atlantic salmon often contain repetitive sequences that complicate the interpretation of the hybridization data (Davidson et al. unpublished observations).

Genes and Expressed Sequence Tags (EST) are also excellent anchor loci. The coding regions are usually highly conserved, and thus they enable genomic comparisons over longer evolutionary distances than do microsatellite loci. PCR primers designed from coding regions can be used to screen pooled BAC DNA templates as described above. In addition, these amplicons can be used as hybridization probes, either individually or in combination. For example, Ren and others (2003) were able to place 361 contigs on to the chicken genetic map by screening BAC filters with probes in a $6 \times 6 \times 6$ matrix (i.e., 18 hybridizations with 36 probes at a time). This proved to be a very effective and efficient method for integrating the physical and linkage maps. A similar approach was used to join the medaka physical map and linkage maps to produce a resource that will facilitate positional cloning and sequencing the genome using a clone-by-clone strategy (Khorasani et al. 2004).

Genetic Markers from BACs

An alternative approach to placing markers from a genetic map on to the physical map is to identify genetic markers in BACs and then to place them on to the genetic map. Rodriguez and others (2006) have developed a PCR-based screening technique to select microsatellite-containing subclones from a BAC shotgun library. Although this method will be valuable for mapping specific BACs and the contigs to which they belong, it does not lend itself to high throughput, which is a hallmark of genomics. BAC-end sequences provide a wealth of information that can be used for physical mapping, comparative genome analysis, and map integration (Xu et al. 2006). Database mining of BAC-end sequences from catfish has revealed that 17.5% of them contain microsatellites (Xu et al. 2006). A similar proportion is found in BAC-end sequences from Atlantic salmon (Ng and Davidson, unpublished observations). The BAC-end sequences therefore, provide

an excellent source of markers that can be placed on the genetic map, and in so doing they provide a link to the physical map. See Chapter 15. Even in the absence of a microsatellite in a BAC-end sequence, there is the opportunity to identify a genetic marker. PCR primers can be designed from the BAC-end sequence and used to screen the parents of mapping families for single nucleotide polymorphisms (SNP), either by direct sequencing or by an indirect means such as challenging the amplicon with a suite of restriction enzymes. It appears likely that BAC-end sequences will become a rich source of genetic markers, especially for novel aquaculture species that do not have a tradition of breeding and pedigree analysis.

Integration of Genetic Maps and Karyotypes

There should be a direct correlation between the number of linkage groups in a genetic map and the number of chromosomes in an organism. However, even when a sufficiently dense linkage map is achieved, the relationship between the linkage groups and karyotype is not immediately obvious. BACs can provide a link between the two. Using fluorescent in situ hybridization (FISH), also discussed in Chapter 17, with BAC probes containing genes mapped to each linkage group, Phillips and others (2006a) were able to assign rainbow trout linkage groups to specific chromosomes. Zebrafish genetic linkage groups have similarly been assigned to chromosomes (Phillips et al. 2006b). These studies showed that there is a rough correlation between chromosome size and size of the linkage group based on total recombination frequency. Having sets of BACs that correspond to chromosome arms for a particular species will prove beneficial for investigating chromosomal evolution in related species. There are several diverse sex-determining mechanisms in fish. But to date, only one teleost master sex-determining gene (*SEX*) has been identified, namely, *DMY* in medaka (Matsuda et al. 2002). The search for the chromosomes on which *SEX* lies in different fish has been aided by the development of BAC libraries. For example, the integration of linkage group 1 in Atlantic salmon with the physical map and subsequent FISH analysis using BACs from selected contigs revealed that *SEX* is located on the long arm of a large metacentric chromosome (Artieri et al. 2006). BAC physical maps have been constructed for the regions surrounding the sex-determining genes of the platyfish (Froschauer et al. 2002) and the three-spined stickleback (Peichel et al. 2004). It is expected that BAC libraries will be essential resources for the identification of the actual genes in these and other species, as well as studying the evolution of sex-determination (Kondo et al. 2006).

Genome Sequencing and Comparative Genomics

Based on experience gained from assembling the sequences of large complex vertebrate genomes, it appears that a combination of whole genome shotgun sequencing and map-assisted sequencing is the most efficient strategy, as described in Chapter 26 (Mouse Genome Sequencing Consortium 2002, Rat Genome Sequencing Consortium 2004, International Chicken Genome Sequencing Consortium 2004). Low coverage sequencing from ordered BACs (the clone-by-clone approach) undoubtedly improves genome sequence quality. Physical maps and linkage maps provide the framework for organizing the sequence. In addition, they provide validation of the

sequence assembly order and supply long-range linking information for closing gaps. In turn, when the genome sequence has been produced, the integrated physical and genetic maps become a major resource for the broader research community, providing it with access to clones with known genomic locations. See for example Wallis and others (2004). For animal breeders, this gives tools to identify candidate genes for QTL by positional cloning (Demars et al. 2006).

BAC-end sequences have a dual role in that they provide raw sequence data as well as anchors to the physical map. These sequences provide a snapshot of the genome, allowing estimates of the number and types of repetitive elements and their partial characterization. Data mining of BAC-end sequences for genes allows the anchoring of BACs, and hence contigs if the BAC library has been fingerprinted, to genomes that have been sequenced. For example, Xu and others (2006) were able to anchor 1,074 and 773 BACs from 25,195 catfish BAC-end sequences to the genomes of zebrafish and *Tetraodon*, respectively. This approach provides a rich source for comparative genomics. Equally importantly, it enables the testing of the prediction that the locations of genes that have been identified close to the anchors in the sequenced fish genomes have retained their syntenic relationships in the genomes of other fish.

Summary and Future Prospects

In this chapter I have introduced the reader to BAC libraries and physical maps, and in particular, their construction, characterization, importance for whole genome sequencing and comparative genomics, and their potential as resources for identifying candidate genes for QTL and ultimately markers for gene-assisted breeding programs. It will be a long time before the genome of every species of interest to aquaculture has been sequenced. In the meantime, it will be possible to glean information from the sequenced genomes of zebrafish and the two pufferfish (*Tetraodon* and *Takifugu*) and others such as medaka and the three-spined stickleback as they become available, through comparative genomics using BAC libraries. As pointed out above, BAC-end sequences enable predictions to be made concerning the locations of genes based on the conservation of synteny. This will likely pay huge dividends over relatively short regions of the genome. Similarly, hybridization of EST probes from a species whose genome has been sequenced can be used to identify and anchor BACs from both closely and quite distantly related species. For example, Romanov and Dodgson (2006) used this "one sequence, multiple genomes" strategy effectively to compare the genomes of turkey and zebrafinch with that of the chicken.

If a novel species is being targeted for aquaculture and there are few genetic or genomic resources available for it, then a first step should surely be the construction of a BAC library. I would suggest that the two next most important steps follow: (1) fingerprinting the BAC library to construct a physical map, and (2) BAC-end sequencing to obtain genetic markers in the form of microsatellites by data mining and SNPs by BAC-end resequencing. Taken together, these resources can be used not only to build a genetic map but also to integrate it with the physical map. Comparative genomics, using a sequenced genome from the species most closely related to the novel species under investigation, will quickly pave the way for advances at the molecular level including pedigree analysis, searches for QTL, and marker-assisted selection breeding programs. Furthermore, these resources will form a solid foundation for whole genome sequencing.

Acknowledgments

I thank everyone associated with the Genomics Research on Atlantic Salmon Project (GRASP) and the Consortium for Genomics Research on All Salmon Project (cGRASP) for their input into this review. I would particularly like to thank Siemon Ng and Krzysztof Lubieniecki for many stimulating discussions and Dawne Dadswell for editorial assistance. Funding from Genome Canada and Genome BC greatly facilitated the writing of this chapter.

References

Artieri CG, LA Mitchell, SHS Ng, SE Parisott, RG Danzmann, B Hoyheim, RB Phillips, M Morasch, BF Koop, and WS Davidson. 2006. Identification of the sex-determining locus of Atlantic salmon (*Salmo salar*) on chromosome 2. Cytogenetics and Genome Research, 112, pp. 152–159.

Brenner S and KJ Livak. 1989. DNA fingerprinting by sampled sequencing. Proc Natl Acad Sci USA, 86, pp. 8902–8906.

Burke DT, GF Carle, and MV Olson. 1987. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. Science, 236, pp. 806–812.

Corley-Smith GE, CJ Lim, and BP Brandhorst. 1996. Production of androgenetic zebrafish (*Danio rerio*). Genetics, 142, pp. 1265–1276.

Coulson A, J Sulston, S Brenner, and J Karn. 1986. Toward a physical map of the genome of the nematode *Caenorhabditis elegans*. Proc Natl Acad Sci USA, 83, pp. 7821–7825.

Crooijmans RPMA, J Vrebalov, RJM Dijkhof, JJ van der Poel, and MAM Groenen. 2000. Two-dimensional screening of the Wageningen chicken BAC library. Mammalian Genome, 11, pp. 360–363.

Danzmann RG, M Cairney, WS Davidson, MM Ferguson, K Gharbi, R Guyomard, LE Holm, E Leder, N Okamoto, A Ozaki, CE Rexroad 3rd, T Sakamoto, JB Taggart, and RA Woram. 2005. A comparative analysis of the rainbow trout genome with 2 other species of fish (Arctic charr and Atlantic salmon) within the tetraploid derivative Salmonidae family (subfamily: Salmoninae). Genome, 48, pp. 1037–1051.

Demars J, J Riquet, K Feve, M Gautier, M Morisson, O Demeure, C Renard, P Chardon, and D Milan. 2006. High resolution physical map of porcine chromosome 7 QTL region and comparative mapping of this regional among vertebrate genomes. BMC Genomics, 7, 13.

Ding Y, MD Johnson, WQ Chen, D Wong, Y-D Chen, SC Benson, JY Lam, Y-M Kim, and H Shizuya. 2001. Five-Color-Based High-Information-Content Fingerprinting of Bacterial Artificial Chromosome Clones Using Type II Restriction Endonucleases. Genomics, 74, pp. 142–154.

Frengen E, D Weichenhan, B Zhao, K Osoegawa, M van Geel, and PJ de Jong. 1999. A Modular, Positive Selection Bacterial Artificial Chromosome Vector with Multiple Cloning Sites. Genomics, 58, pp. 250–253.

Froschauer A, C Korting, T Katagiri, T Aoki, S Asakawa, N Shimizu, M Schartl, and JN Volff. 2002. Construction and initial analysis of bacterial artificial chromosome (BAC) contigs from the sex-determining region of the platyfish Xiphophorus maculates. Genetics, 295, pp. 247–54.

Fuhrmann DR, MI Krzywinski, R Chiu, P Saeedi, JE Schein, IA Bosdet, A Chinwalla, LW Hillier, RH Waterston, JD McPherson, SJM Jones, and MA Marra. 2003. Software for Automated Analysis of DNA Fingerprinting Gels. Genome Research, 13, pp. 940–953.

Fujiyama A, H Watanabe, TD Toyoda, T Itoh, SF Tsai, HS Park, ML Yaspo, H Lehrach, Z Chen, G Fu, N Saitou, K Osoegawa, PJ de Jong, Y Suto, M Hattori, and Y Sakaki. 2002. Construction and analysis of a human-chimpanzee comparative clone map. Science, 295, pp. 131–134.

Ioannou PA, CT Amemiya, J Garnes, PM Kroisel, H Shizuya, C Chen, MA Batzer, and JP de Jong. 1994. A new bacteriophage P1-derived vector for the propagation of large human DNA fragments. Nature Genetics, 6, pp. 84–89.

Katagiri T, S Asakawa, I Hirono, T Aoki, and N Shimizu. 2000. Genomic bacterial artificial chromosome library of the Japanese flounder *Paralichthys olivaceus*. Marine Biotechnology, 2, pp. 571–6.

Katagiri T, S Asakawa, S Minagawa, N Shimizu, I Hirono, and T Aoki. 2001. Construction and characterization of BAC libraries for three fish species; rainbow trout, carp and tilapia. Animal Genetics, 32, pp. 200–204.

Katagiri T, S Minagawa, I Hirono, K Kato, M Miyata, S Asakawa, N Shimizu, and T Aoki. 2002. Construction of a BAC library for the red sea bream *Pagrus major*. Fisheries Science, 68, p. 942.

Khorasani MZ, S Hennig, G Imre, S Asakawa, S Palczewski, A Berger, H Hori, K Naruse, H Mitani, A Shima, H Lehrach, J Wittbrodt, H Kondoh, N Shimizu, and H Himmelbauer. 2004. A first generation physical map of the medaka genome in BACs essential for positional cloning and clone-by-clone based genomic sequencing. Mechanisms of Development, 121, pp. 903–913.

Klein PE, RR Klein, SW Cartinhour, PE Ulanch, J Dong, JA Obert, DT Morishige, SD Schlueter, KL Childs, M Ale, and JE Mullet. 2000. A High-throughput AFLP-based Method for Constructing Integrated Genetic and Physical Maps: Progress Toward a Sorghum Genome Map. Genome Research, 10, pp. 789–807.

Kondo M, U Hornung, I Nanda, S Imai, T Sasaki, A Shimizu, S Asakawa, H Hori, M Schmid, N Shimizu, and M Schartl. 2006. Genomic organization of the sex-determining and adjacent regions of the sex chromosomes of medaka. Genome Research, 16, pp. 815–826.

Krzywinski M, I Bosdet, D Smailus, R Chiu, C Mathewson, N Wye, S Barber, M Brown-John, S Chan, S Chand, A Cloutier, N Girn, D Lee, A Masson, M Mayo, T Olson, P Pandoh, AL Prabhu, E Schoenmakers, M Tsai, D Albertson, W Lam, CO Choy, K Osoegawa, S Zhao, PJ de Jong, J Schein, S Jones, and MA Marra. 2004. A set of BAC clones spanning the human genome. Nucleic Acids Res, 32, pp. 3651–3660.

Larin Z, AP Monaco, and H Lehrach. 1991. Yeast artificial chromosome libraries containing large inserts from mouse and human DNA. Proc Natl Acad Sci, 88, pp. 4123–4127.

Luo M-C, C Thomas, FM You, J Hsiao, S Ouyang, CR Buell, M Malandro, PE McGuire, OD Anderson, and J Dvorak. 2003. High-throughput fingerprinting of bacterial artificial chromosomes using the SnaPshot labeling kit and sizing of restriction fragments by capillary electrophoresis. Genomics, 82, pp. 378–389.

Marra M, T Kucaba, M Sekhon, L Hillier, R Martienssen, A Chinwalla, JM Crockett, J Fedele, H Grover, C Gund, WR McCombie, K McDonald, J McPherson, N Mudd, L Parnell, J Schein, R Seim, P Shelby, R Waterston, and R Wilson. 1999. A map for sequence analysis of the Arabidopsis thaliana genome. Nature Genetics, 22, pp. 265–270.

Marra MA, TA Kucaba, NL Dietrich, ED Green, B Brownstein, RK Wilson, KM McDonald, LW Hillier, JD McPherson, and RH Waterston. 1997. High Throughput Fingerprint Analysis of Large-Insert Clones. Genome Research, 7, pp. 1072–1084.

Matsuda M, Y Nagahama, A Shinomiya, T Sato, C Matsuda, T Kobayashi, CE Morrey, N Shibata, S Asakawa, N Shimizu, H Hori, S Hamaguchi, and M Sakaizumi. 2002. DMY is a Y-specific DM-domain gene required for male development in the medaka fish. Nature, 417, pp. 559–563.

McPherson JD, M Marra, L Hillier, RH Waterston, A Chinwalla, J Wallis, M Sekhon, K Wylie, ER Mardis, RK Wilson, R Fulton, TA Kucaba, C Wagner-McPherson, WB Barbazuk, SG Gregory, SJ Humphray, L French, RS Evans, G Bethel, A Whittaker, JL Holden, OT McCann, A Dunham, C Soderlund, CE Scott, DR Bentley, G Schuler, HC Chen, W Jang, ED Green, JR Idol, VV Maduro, KT Montgomery, E Lee, A Miller, S Emerling, RS Kucherlapati, R Gibbs, S Scherer, JH Gorrell, E Sodergren, K Clerc-Blankenburg, P Tabor, S Naylor, D Garcia, PJ de Jong, JJ Catanese, N Nowak, K Osoegawa, S Qin, L Rowen, A Madan, M Dors, L Hood,

B Trask, C Friedman, H Massa, VG Cheung, IR Kirsch, T Reid, R Yonescu, J Weissenbach, T Bruls, R Heilig, E Branscomb, A Olsen, N Doggett, JF Cheng, T Hawkins, RM Myers, J Shang, L Ramirez, J Schmutz, O Velasquez, K Dixon, NR Stone, DR Cox, D Haussler, WJ Kent, T Furey, S Rogic, S Kennedy, S Jones, A Rosenthal, G Wen, M Schilhabel, G Gloeckner, G Nyakatura, R Siebert, B Schlegelberger, J Korenberg, XN Chen, A Fujiyama, M Hattori, A Toyoda, T Yada, HS Park, Y Sakaki, N Shimizu, S Asakawa, K Kawasaki, T Sasaki, A Shintani, A Shimizu, K Shibuya, J Kudoh, S Minoshima, J Ramser, P Seranski, C Hoff, A Poustka, R Reinhardt, and H Lehrach. 2001. International Human Genome Mapping Consortium. A physical map of the human genome. Nature, 409, pp. 934–941.

Myers BC, S Scalabrin, and M Morgante. 2004. Mapping and Sequencing Complex Genomes: Let's Get Physical! Nature Reviews Genetics, 5, pp. 578–588.

Ness SR, W Terpstra, M Krzywinski, MA Marra, and SJM Jones. 2002. Assembly of fingerprint contigs: parallelized FPC. Bioinformatics, 18, pp. 484–485.

Ng SHS, CG Artieri, IE Bosdet, R Chiu, RG Danzmann, WS Davidson, MM Ferguson, CD Fjell, B Hoyheim, SJM Jones, PJ de Jong, BF Koop, MI Krzywinski, K Lubieniecki, MA Marra, LA Mitchell, C Mathewson, K Osoegawa, SE Parisotto, RB Phillips, ML Rise, KR von Schalburg, JE Schein, H Shin, A Siddiqui, J Thorsen, N Wye, G Yang, and B Zhu. 2005. A physical map of the genome of Atlantic salmon, Salmo salar. Genomics, 86, pp. 396–404.

Osoegawa K, PY Woon, B Zhao, E Frengen, M Tateno, JJ Catanese, and PJ de Jong. 1998. An Improved Approach for Construction of Bacterial Artificial Chromosome Libraries. Genomics, 52, pp. 1–8.

Osoegawa K, B Zhu, CL Shu, T Ren, Q Cao, GM Vessere, MM Lutz, MI Jensen-Seaman, S Zhao, and PJ de Jong. 2004. BAC Resources for the Rat Genome Project. Genome Research, 14, pp. 780–785.

Palti Y, SA Gahr, JD Hansen, and CE Rexroad III. 2004. Characterization of a new BAC library for rainbow trout: evidence for multi-locus duplication. Animal Genetics, 35, pp. 130–133.

Parsons JE and GH Thorgaard. 1985. Production of androgenetic diploid rainbow trout. Journal of Heredity, 76, pp. 177–81.

Peichel CL, JA Ross, CK Matson, M Dickson, J Grimwood, J Schmutz, RM Myers, S Mori, D Schluter, and DM Kingsley. 2004. The master sex-determination locus in threespine sticklebacks is on a nascent Y chromosome. Current Biology, 14, pp. 1416–1424.

Phillips RB, A Amores, MR Morasch, C Wilson, and JH Postlethwait. 2006b. Assignment of zebrafish genetic linkage groups to chromosomes. Cytogenetic and Genome Research, 114, pp. 155–62.

Phillips RB, KM Nichols, JJ Dekoning, MR Morasch, KA Keatley, C Rexroad III, SA Gahr, RG Danzmann, RE Drew, and GH Thorgaard. 2006a. Assignment of Rainbow Trout Linkage Groups to Specific Chromosomes. Genetics, 174, pp. 1661–1670.

Quiniou SMA, T Katagiri, NW Miller, M Wilson, WR Wolters, and GC Waldbieser. 2003. Construction and characterization of a BAC library from a gynogenetic channel catfish *Ictalurus punctatus.* Genetics and Selection Evolution, 35, pp. 673–683.

Rodriquez MF, SA Gahr, CE Rexroad 3rd, and Y Palti. 2006. A polymerase chain reaction screening method for rapid detection of microsatellites in bacterial artificial chromosomes. Marine Biotechnology, 8, pp. 346–350.

Romanov MN and JB Dodgson. 2006. Cross-species overgo hybridization and comparative physical mapping with avian genomes. Animal Genetics, 37, pp. 397–399.

Ross MT, SM Labire, J McPherson, and JV Stanton. 1999. In: Current Protocols in Human Genetics, Boyle A, Ed. 5.6.1–5.6.52, Wiley, New York.

Shizuya H, B Birren, U-J Kim, V Mancino, T Slepak, Y Tachiiri, and M Simon. 1992. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. Proc Natl Acad Sci USA, 89, pp. 8794–8797.

Soderlund C, I Longden, and R Mott. 1997. FPC: a system for building contigs from restriction fingerprinted clones. Computational and Applied Biosciences, 13, pp. 523–535.

Soderlund C, S Humphray, A Dunham, and L French. 2000. Contigs Built with Fingerprints, Markers, and FPC V4.7. Genome Research, 10, pp. 1772–1787.

Sternberg N, J Ruether, and K de Riel. 1990. Generation of a 50,000-member human DNA library with an average DNA insert of 75–100 kbp in a bacteriophage P1 cloning vector. New Biologist, 2, pp. 151–162.

Sulston J, F Mallett, R Staden, R Durbin, T Horsnell, and A Coulson. 1988. Software for genome mapping by fingerprinting techniques. Computational and Applied Biosciences, 4, pp. 125–32.

Thorsen J, B Zhu, E Frengen, K Osoegawa, PJ de Jong, BF Koop, WS Davidson, and B Hoyheim. 2005. A highly redundant BAC library of Atlantic salmon *Salmo salar*: an important tool for salmon projects. BMC Genomics, 6, 50.

Wallis JW, J Aerts, MAM Groenen, RPMA Crooijmans, D Layman, T Graves, D Scheer, C Kremitzki, MJ Fedele, NK Mudd, M Gardenas, J Higginbotham, J Carter, R McGrane, T Galge, K Mead, J Walker, D Albract, J Davito, S-P Yan, S Leong, A Chinwalla, M Sekhon, K Wylie, J Dodgson, MN Romanov, H Cheng, PJ deJong, K Osoegawa, M Nefedov, H Zhang, JD McPherson, M Krzywinski, J Schein, LW Hillier, ER Mardis, RK Wilson, and WC Warren. 2004. A physical map of the chicken genome. Nature, 432, pp. 761–764.

Woram RA, K Gharbi, T Sakamoto, B Hoyheim, LE Holm, K Naish, C McGowan, MM Ferguson, RB Phillips, J Stein, R Guymard, M Cairney, JB Taggart, R Powell, WS Davidson, and RG Danzmann. 2003. Comparative genome analysis of the primary sex-determining locus in Salmonid fishes. Genome Research, 13, pp. 272–80.

Xu P, S Wang, L Liu, E Peatman, B Somridhivej, J Thimmapuram, G Gong, and Z Liu. 2006. Channel catfish BAC-end sequences for marker development and assessment of syntenic conservation with other fish species. Animal Genetics, 37, pp. 321–326.

Xu Z, S Sun, L Covaleda, K Ding, A Zhang, C Wu, C Scheuring, and H-B Zhang. 2004. Genome physical mapping with large-insert bacterial clones by fingerprint analysis: methodologies, source clone genome coverage, and contig map quality. Genomics, 84, pp. 941–951.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 15
Physical Characterization of Genomes Through BAC End Sequencing

Peng Xu, Shaolin Wang, and Zhanjiang Liu

BAC (bacterial artificial chromosome) libraries are large-insert genomic libraries suitable for physical mapping, as well as serving as the basis for clone-by-clone whole genome sequencing. They are important genome resources that can be exploited for other purposes. Particularly for aquaculture genomics, where whole genome sequencing is not likely for most of the hundreds of species used in aquaculture, BAC libraries provide opportunity to gain much genomic information that would be otherwise difficult to obtain. The construction of BAC libraries and BAC-based fingerprinting for contig construction are covered in Chapter 13 and Chapter 14, respectively. In this chapter, we will briefly describe physical characterization of the genome through BAC-end sequencing analysis, and the value of BAC-end sequences for genome analysis.

Generation of BAC End Sequences

BAC-end sequences can be generated by direct sequencing of BAC clones using sequencing primers designed based on the BAC vector sequences at the border of the genomic insert. Typically, Sp6 and T7 sequencing primers can be used because these sequencing primer sequences have been incorporated into the BAC vectors. The sequencing reactions are straightforward using the dideoxy chain termination sequencing reactions (Sanger's sequencing method) except that a large number of cycles is required for cycle sequencing, usually 80–100 cycles (Xu et al. 2006) because of low copy numbers of BAC DNA.

Although BAC-end sequencing is straightforward, it is important to emphasize the significance of tracking and quality issues. As long-term genome resources, BAC-end sequences must be properly tracked with great quality assurances. Resequencing of a small fraction of the clones is generally recommended. For instance, eight clones can be resequenced from each 384-plate from positions A1, A2, B1, B2, C1, C2, D1, and D2, or better off, 16 clones can be sequenced from the diagonal positions of A1, B2, C3, D4, E5, F6, etc. Quality assessment can be performed using the raw chromatogram files directly before any trimming. The raw, untrimmed files can be processed by Phred software (Ewing and Green 1998, Ewing et al. 1998). Phred quality score cut off value is usually set at 20 for the acquisition of Q20 values. A Q20 curve can be generated to exhibit the length distribution of sequences that passed the Q20 threshold.

Sequence Processing and Routine Bioinformatic Analysis of BAC-end Sequences

BAC-end sequences need to be processed and submitted to GSS database of the GenBank. An example of the routine analysis before sequence submission is shown in Figure 15.1. The BAC-end sequences need to be trimmed of vector sequences and filtered of bacterial sequences, stored in a local Oracle database after base calling and quality assessment. We have used the Genome Project Management System, a local laboratory information management system, for large-scale DNA sequencing projects (Liu et al. 2000). Quality assessment was performed using Phred software (Ewing and Green 1998, Ewing et al. 1998) using $Q \geq 20$ as a cutoff. Repeats were masked using RepeatMasker software (http://www.repeatmasker.org) before Basic Local Alignment Search Tool (BLAST) analysis.

BLASTX searches of the repeat masked BES were conducted against Non-Redundant Protein database. A cut off value of e^{-5} was used as the significance similarity threshold for the comparison. The BLASTX result was parsed out in a tab-delimited format. In order to anchor the catfish BES to zebrafish and *Tetraodon* genomes, BLASTN searches of the repeat masked catfish BES were conducted against zebrafish

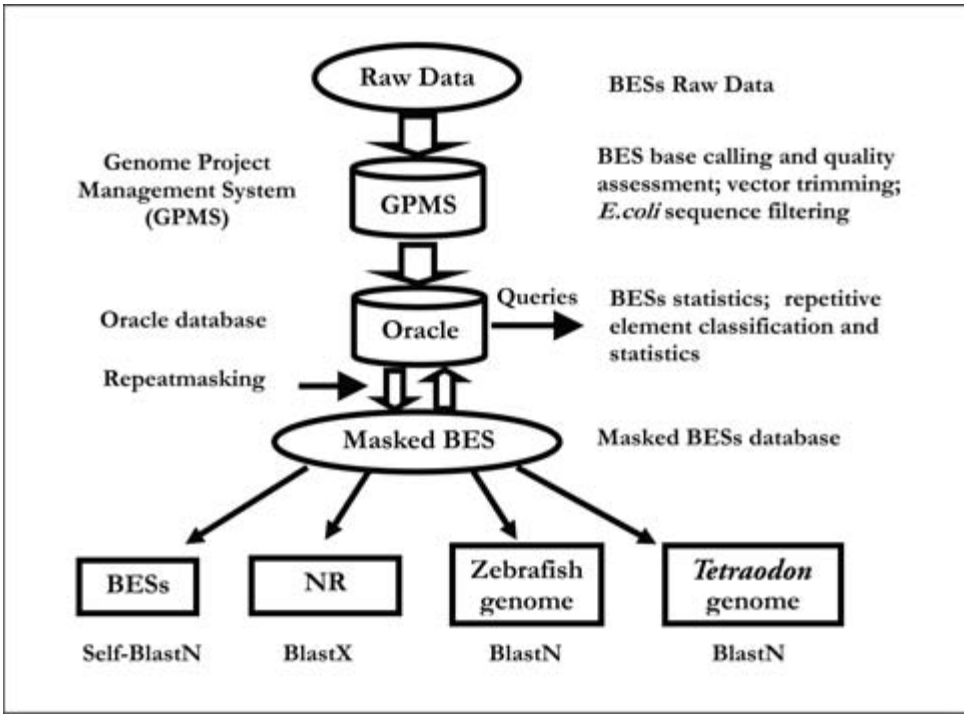


Figure 15.1. An example of routine informatic analysis of BAC-end sequences (BES) before their submission to GSS of the GenBank.

and *Tetraodon* genome sequences. The location and chromosome number of each top hit were collected from the results and parsed out in tab-delimited format.

Microsatellites and other simple sequence repeats were analyzed by using Repeatmasker as well as by using Vector NTI Suite 9.0 (Invitrogen, Carlsbad, CA) as we previously described (Serapion et al. 2004). *Microsatfinder (http://www.eusebius.mysteria.cz/microsatfinder/index.php)* is another user-friendly program for the identification of microsatellites within known sequences.

BAC End Sequences Provide an Unbiased Survey of Genomic Sequences

The average insert size of BAC libraries is usually 100–200 kilobase (kb), and such inserts are usually prepared by partial restriction digest of genomic DNA. Assuming restriction sites are randomly distributed (which would be expected to be the case except within repetitive elements, but often proved to be not), then the BAC ends represent random genomic sequences. For the sake of understanding, we will use the catfish BAC-end sequencing as an example. Channel catfish have a genome size of 1×10^9 base pairs (bp). One of its BAC libraries, CHORI 212 (http://bacpac.chori.org/catfish212.htm) has an average insert size of 160 kb, and thus every 6,250 BAC clones cover one genome. Sequencing 6,250 BAC clones from both ends would effectively generate a sequence tag per 80 kb of the genomic DNA on average. In most cases, BAC libraries have a multiple fold (at least 6–8 folds) of genome coverage to assure a near complete coverage of the entire genome. For instance, the channel catfish CHORI 212 BAC library has a 10.6× genome coverage with 72,067 clones. Sequencing all of the BAC ends from this library would generate one sequence tag per 8 kb genomic DNA, on average. Every BAC-end sequencing reaction can easily produce 500–600 bp. Sequencing all of the BAC ends of the CHORI 212 BAC library should generate 7–8% of 1× genome sequences. With exception of its expense, BAC-end sequencing can rapidly and effectively produce a reasonably unbiased survey of the genome of interest. Such a sequence survey should allow estimation of A/T (G/C) content of the genome, assessment of the repeat structure of the genome, discovery of the microsatellite markers for linkage mapping, production of genomic resources for comparative mapping, and virtually mapping genes to BACs.

Assessment of Repeat Structure of the Genome from BAC End Sequences

The repeat structure of a genome can be assessed from BAC-end sequences. Most often, such an assessment can be accomplished by answering the following two questions:

1. What is the fraction of the genomic sequence survey sharing repeat sequences with species from which the entire genome sequences are available?
2. What types of species-specific novel repeat sequences exist in the species under study?

For the first question, the answer can be derived from analysis of the BAC-end sequences using RepeatMasker (http://www.repeatmasker.org/), a program for the masking of repetitive sequences using various repeat libraries. For instance, analysis using the 11,414,601 bp channel catfish BAC-end sequences resulted in 10.86% masked using *Danio* repeat database, and 7.31% masked using *Takifugu* repeat database. Use of both *Danio* and *Takifugu* repeat databases masked 11.91% (see Chapter 16). This suggested that teleost fishes share a high level of repetitive elements and also that a significant fraction of taxa-specific repeats exists. The analysis indicated that a larger pool of repetitive elements is shared between the genomes of zebrafish and catfish than between the fugu and catfish genomes. It is also obvious that the catfish and the zebrafish genomes harbor a larger percentage of repetitive elements than the fugu genome, an expected result because the fugu genome is much more compact.

Analysis of the catfish BES against the human repeat database suggested the presence of a significant fraction of similar repetitive elements between teleost fish and mammals. In addition to the expected simple repeats (2.7%) and low complexity repeats (1.1%), the RepeatMasker masked 1.08% of the channel catfish BES in the category of DNA repetitive elements, of which the vast majority (1.06%) were the MER2 type of repeats.

The identification of novel repetitive elements in the genome can be approached by bioinformatic analysis of the BAC-end sequences. Several old computer software packages are available (Devereux et al. 1984, Agarwal and States 1994, Rivals et al. 1997). However, most of the earlier programs have a limit on the maximal sequence that can be analyzed. For instance, the Repeat Finder of the GCG package (version 7.0) has a maximal sequence limit of 350,000 bp (Kurtz and Schleiermacher 1999). BLAST (Altschul et al. 1997) and MegaBLAST (Zhang et al. 2000) are quite efficient in the analysis of sequences, but are also limited by sequence size they can process. For the identification of novel repeats in the catfish genome, we have used self BLASTN to detect sequences that share a high level of similarity with other sequences, an indication of repetitive elements. The cut off e-value was set at e^{-2}, and the identify threshold was set to 90% with a minimal alignment length of 100 bp. The BLASTN results were parsed to tab-delimited format to count the redundant queries and other statistics.

Several recently developed software programs can handle large data sets up to the size of the human genome for the identification of repeats. These programs include *REPuter* (Kurtz and Schleiermacher 1999, Kurtz et al. 2000), MUMmer (Delcher et al. 1999), and FORRepeats (Lefebvre et al. 2003). The *REPuter* program (http://www.genomes.de/) is a powerful program for the analysis of repeats in large data sets. The search engine *REPfind* of *REPuter* uses an efficient and compact implementation of suffix trees in order to locate exact repeats in linear space and time. These exact repeats are used as seeds from which significant degenerate repeats are constructed allowing for mismatches, insertions, and deletions. The user has the option for defining the parameters of the search including minimum length and maximum number of errors (sequence divergence). Most often, a minimum length of 20 bp and a 10% error (90% sequence conservation) should be highly sufficient for the detection of repetitive sequences. The output is sorted by significance scores (*E*-values). In addition to finding degenerate direct repeats, *REPfind* is capable of detecting degenerate palindromic repeats (Kurtz et al. 2001). The FORRepeats is more suitable for comparison of genomes between species.

In spite of the fact that BAC-end sequences can be used to effectively estimate repeat structures, there are limitations to using BAC-end sequences for the characterization of the repeat structure of the genome, usually leading to the underestimation of the repeats in the genome. The major problem is its inability to identify the tandem repeats that had not been cloned into the BAC library. For instance, we previously described the presence of a major class of tandem repeats named *Xba* elements (Liu et al. 1998) that accounted for about 5% of the catfish genome. These elements were not detected in the BES, because they lack the *Eco*R I restriction sites necessary for insertion into BAC clones.

Identification of Microsatellites from BAC-end Sequences

BAC-end sequences are a genome resource that can be used to mine microsatellite markers. As a matter of fact, they are rich in microsatellites. For instance, during the analysis of 20,366 BAC-end sequences, it was found that 3,748 BAC-end sequences (18.4%) contain one or more stretches of microsatellite sequences. Of these, 2,365 (63%) had sufficient flanking sequences on both sides, making them potentially useful as markers for genetic mapping; 403 BES harbor microsatellite sequences at the immediate beginning of the BES, making them more difficult to be developed as markers; the remaining 980 clones had microsatellite sequences at the end of BES. It is obvious that the number of microsatellites at the end of BES was much larger than the number of microsatellites at the beginning of BES. That is because many sequencing reactions terminated due to the presence of simple sequence repeats. For this last category, additional sequencing can be conducted using primers close to the end of the BES to generate sufficient flanking sequences on both sides. With additional efforts including generation of sufficient flanking sequences and testing of polymorphism, these microsatellites should be useful for genetic linkage mapping and integration of the catfish linkage maps with the BAC-based physical map. Bioinformatic mining for microsatellites is perhaps the most productive and economic approach if the genome resources such as BES exist.

Several Web-based programs are sufficient for the identification of microsatellites within BAC-end sequences. Microsatfinder *(http://www.eusebius.mysteria.cz/microsatfinder/index.php)*, Tandem Repeat Finder (Benson 1999, http://tandem.bu.edu/trf/trf.html), and RepeatFinder (http://www.genet.sickkids.on.ca/~ali/repeatfinder.html) are all quite user-friendly. MICAS (http://210.212.212.7/MIC/index.html) is another highly user-friendly, interactive Web-based server to find nonredundant microsatellites in a given nucleotide sequence/genome sequence (Sreenu et al. 2003). The Informax Vector NTI software packages are also very efficient for the identification of microsatellites within sequences (Serapion et al. 2004), allowing establishment of a microsatellite database.

BAC-end sequences also allow an overall glance at the types and relative abundance of microsatellites in an organism. For instance, in catfish, the most abundant microsatellite type is CA/GT. Overall, AT-rich microsatellites are more abundant than GC-rich microsatellites (Figure 15.2).

The BAC-anchored microsatellites are a valuable resource for integration of genetic linkage and physical maps (Figure 15.3). Because they are identified through

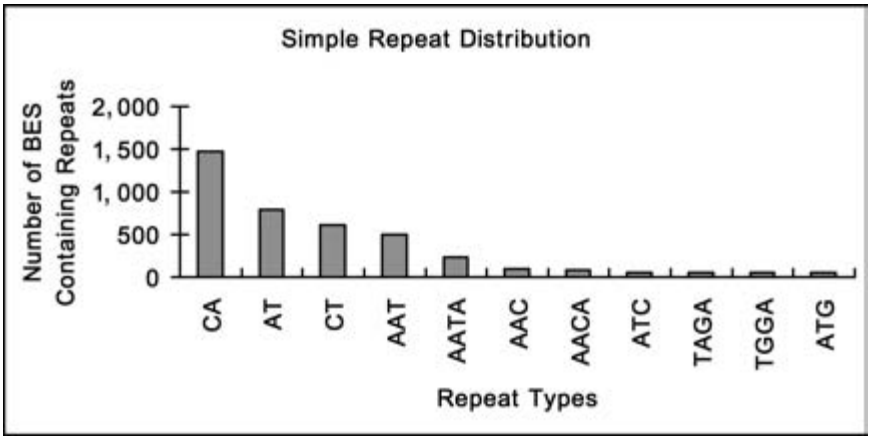


Figure 15.2. Microsatellite types and their relative abundance in the catfish genome as revealed by informatic analysis of BAC-end sequences.

BAC-end sequencing, their location on the physical map can be determined by BAC contig construction. After they are mapped to genetic linkage maps by genotyping them in a resource family, they allow alignment of the linkage and the physical maps (Figure 15.3).

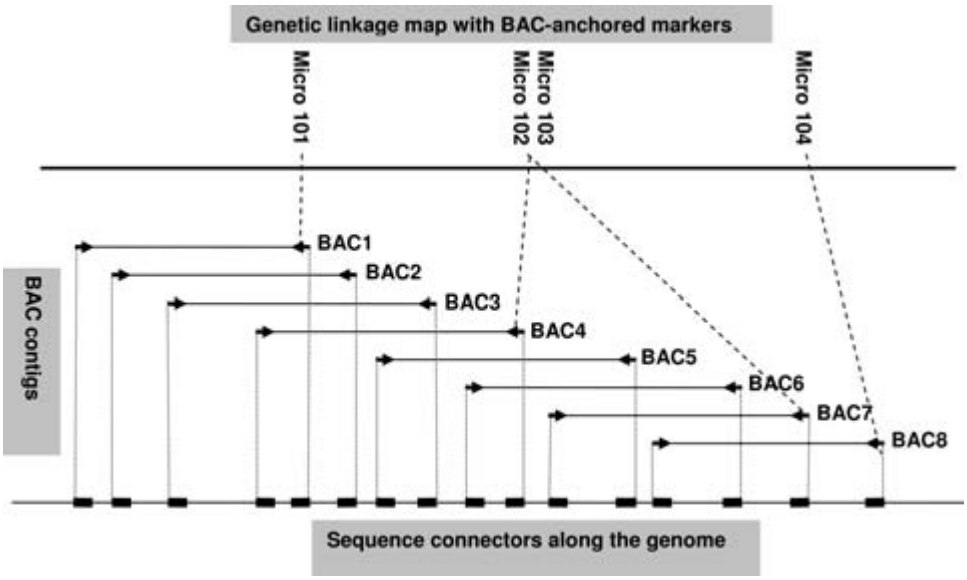


Figure 15.3. BAC-end sequences are produced from both ends (arrows in opposite directions on the BAC clones within the contig). After the BAC contigs are constructed, the BAC-end sequences become sequence tag connectors (shown as black rectangles along the line, bottom) along the genome as dictated by the distribution of BAC contigs. The microsatellite markers (micro 101–104) identified from BAC-end sequences allow integration of linkage map (top) and the BAC-based physical maps (middle).

BAC-End Sequencing Allows Virtual Mapping of Genes to Physical Maps

The entire business of gene mapping deals with the positions of genes on chromosomes and in the genome. Genes can be mapped in various ways including genetic linkage mapping, and mapping of genes to physical maps through hybridization, however, BAC-end sequencing is probably the most efficient way to map genes to BACs. After the BAC-based contigs are constructed, the genes located on the BACs are placed on the physical map.

Genes can be placed on BACs by simple BLASTX searches of BAC-end sequences. For instance, BLASTX searches of the 20,366 channel catfish BES resulted in 2,351 BES with significant hits to 1,877 unique genes (E value $<e^{-5}$), demonstrating the efficiency. However, gene identification by BLASTX searches is complicated by the presence of erroneously annotated sequences in the GenBank, especially in cases when repetitive elements are involved. In addition, it is a challenging task to set a solid benchmark for significant hits. Because different genes may have different evolving rates, researchers have to be flexible enough to make a judgment as to what level of E-values is sufficient to putatively identify the sequence. At times, it is safe to provide the significant hits with their associated E-values, as well as the alignment length. A short alignment length may indicate only part of an exon is involved in the BAC end sequences, which even provide more concrete evidence than alignments with long sequence but scattered identities. For instance, of the 1,877 gene hits in the catfish example, 1,130 had an e-value smaller than e^{-10} in BLASTX searches with average alignment length of 73 amino acids and a range of 31–100% identity (Table 15.1). Although it is difficult to conclude what level of e-values would provide a stringent confidence on the putative gene identities of the BES, it is obvious that the lower the E-values, the more likely the BES are to be related to the putatively identified genes.

Table 15.1. Mapping of genes to BACs through BAC-end sequencing. Listed are a number of BLASTX hits of genes by BAC-end sequences, excluding redundant hits. E-values, alignment length range, average alignment length, and percentage of identities are provided as an indication for the level of similarities.

E-values	Number of hits	Alignment length (amino acid)	Average alignment length (amino acids)	(%) Identity
$<10^{-50}$	58	101–228	167	48–99
10^{-40}–10^{-50}	54	81–207	134	43–97
10^{-30}–10^{-40}	77	66–217	103	40–100
10^{-20}–10^{-30}	253	45–175	75	34–100
10^{-15}–10^{-20}	275	37–199	62	30–100
10^{-10}–10^{-15}	413	30–186	54	31–100
Subtotal	1130	30–228	73	31–100
10^{-5}–10^{-10}	747	19–193	47	23–100
Total	1,877	19–228	63	23–100

Anchoring BES to Existing Genome Sequences for Comparative Genomics

The fundamental basis for comparative genomics is the genomes are conserved in their gene sequences, as well as in the arrangement of genes and other important functional domains. The BAC-end sequences, though not continuous, provide a string of sequences that can be used to compare with genome sequences for the presence of related sequences arranged in a similar fashion. Such comparisons can be conducted by direct BLAST searches. In the case of fish, the entire genome sequences exist for zebrafish, fugu, and *Tetraodon* species. A direct comparison of BAC-end sequences and the genome sequences of the closely related species should identify chromosome-anchored conserved sequences within the species of interest. For instance, searches of the 20,366 catfish BES against *D. rerio* and *T. nigroviridis* genome sequences resulted in 3,251 (16%) and 1,670 (8.2%) significant hits (E value $<e^{-5}$), respectively. However, many BES had many hits in different genomic regions of the zebrafish genome sequence, suggesting that they are repetitive in nature. In order to obtain unique significant hits that are meaningful as resources for comparative genome analysis, the significant hits were tabulated in Excel with hit ID, chromosome information, and beginning and ends of the region of similarity. Repeated hits were then removed, resulting in 1,074 unique significant hits to the zebrafish genome sequence. Similar BLAST searches against *T. nigroviridis* genome resulted in 773 unique significant hits, suggesting that the number of genomic regions containing evolutionarily conserved sequences was greater between catfish and zebrafish than between catfish and *T. nigroviridis*, consistent with their phylogenetic relationships.

In order to understand the nature of the conserved genomic sequence blocks between catfish and zebrafish or *T. nigroviridis*, the BES with unique BLASTN significant hits were searched against the NR database using BLASTX to assess the number of genes among the conserved genomic sequences. Of the 1,074 unique significant hits to the zebrafish genome, 417 (38.8%) had significant BLASTX hits. Similarly, with *Tetraodon*, of the 773 unique significant hits, 406 had significant BLASTX hits. Clearly, of the unique BLASTN hits, the number of significant BLASTX hits was similar with the zebrafish and the *Tetraodon* genomes, suggesting that the vast majority of genes were conserved among catfish, zebrafish, and *Tetraodon*. The greater number of unique BLASTN hits to the zebrafish genome was accounted for, in the most part, by the nongene genomic sequences, reflecting the close relatedness of the catfish and zebrafish genomes.

To anchor the catfish BES to the chromosomes of the zebrafish and *Tetraodon* genomes, BLASTN and BLASTX search results were tabulated according to chromosome locations (Table 15.2). The catfish BES had significant BLASTN hits to every one of the 25 zebrafish chromosomes with a range of 16–66 hits per chromosome. Of the unique BLASTN hits, all 25 chromosomes had significant BLASTX hits to the catfish BES, with 4–30 hits per chromosome. Similar but fewer unique hits were found with the *Tetraodon* genome. The notable low number of hits was found with *Tetraodon* chromosome 20, with only a single significant hit with BLASTN or BLASTX. This was mainly because of the low sequence coverage of the *Tetraodon* chromosome 20 to date (http://www.genoscope.cns.fr/externe/tetranew/), and perhaps also because of the small size of chromosome 20.

Table 15.2. Distribution of unique BES significant BLASTN hits in the *Danio rerio* and *Tetraodon nigroviridis* genomes. The cut-off value was set at 1 × E-05.

	Zebrafish		*Tetraodon*	
Chromosome	Unique BLASTN hits	BLASTX hits of unique BLASTN hits	Unique BLASTN hits	BLASTX hits of unique BLASTN hits
1	58	22	44	32
2	39	13	55	27
3	58	28	36	19
4	36	14	13	10
5	66	27	23	13
6	40	21	15	4
7	56	22	20	11
8	36	15	20	14
9	59	24	23	12
10	42	16	32	21
11	38	15	21	11
12	37	16	24	16
13	45	16	74	17
14	54	16	14	9
15	31	9	30	25
16	53	18	23	15
17	47	15	16	9
18	28	4	21	14
19	44	20	10	8
20	52	30	1	1
21	34	10	19	14
22	28	11		
23	54	20		
24	23	7		
25	16	8		
Undesignated			239	104
Total	1,074	417	773	406

Identification of Conserved Syntenies Using BAC-end Sequences

Large-scale BAC-end sequencing is currently the most efficient strategy for building whole-genome comparatively anchored physical maps in map-poor species (Larkin et al. 2003). Conservation of genes and their locations can also be approached in a local genomic environ, allowing identification of conserved syntenies. Assuming genes and their genomic locations are conserved through evolution, it is reasonable to assume that any two given genes that are close to each other may have the same arrangement in the same genomic environ as well as in a closely related species. Based on this notion, BAC-end sequences are analyzed for the presence of genes on both ends of the same BAC clone. Given that the average insert size of BAC libraries is 100–200 kb, these genes on both ends of the same BAC clone are physically linked with a distance of 100–200 kb apart. This information can be used to determine if the same genes are arranged in the

same genomic environ in species with entire genome sequences by BLAST searches. If the answer is yes, the process identifies a conserved synteny between the two species. This approach was demonstrated to be very efficient for the identification of conserved syntenies. For instance, of the 20,366 BES, 17,478 BES were mate pair sequences from 8,739 BAC clones. BLASTX searches indicated that 141 sequenced BACs harbor genes on both ends. These paired BAC-ends with genes allowed us to compare whether the same set of genes were located on the similar genomic environs in the zebrafish and *Tetraodon* genomes. Of the 141 paired BAC-ends with genes, 43 (30.5%) were located on the same chromosomes of zebrafish or *Tetraodon*, of which 23 (16.3%) appeared to exhibit a high level of conserved synteny. The number of conserved syntenies was greater between the catfish genome and the zebrafish genome than between the catfish genome and the *Tetraodon* genome. Of the 23 conserved syntenies, 21 were present between the catfish and zebrafish genomes. Additional experiments using comparative analysis and direct BAC sequencing in catfish revealed that many of the syntenies could be extended. These encouraging results suggest that comparative mapping, especially with zebrafish, will be a sound approach for future catfish genomics research. Furthermore, information gained in mapping and gene discovery projects in catfish may help to explain aspects of zebrafish and teleost genome evolution.

BAC-end Sequences and the Minimal Tiling Path for Entire Genome Sequencing

The successful sequencing of the human and mouse genomes stirred up a wave of excitement in genome biology. Currently, whole genome sequencing has been completed or is being completed for a number of vertebrate animals including important agricultural animals such as cattle, porcine, and chicken. Sequencing in fish species, however, has been limited to several model species such as the zebrafish (*Danio rerio*), *Takifugu rubripes*, *Tetraodon nigroviridis*, and medaka. As the whole genome sequences have been crucial for the study of genome expression and function, scientists are now seriously considering a genome-sequencing project for important aquaculture species. Large-scale BAC-end sequencing is not only a necessary step as a survey of the genome in the assessment of genome composition and architecture, but also a required element for the identification of minimal tiling pass for clone-by-clone based strategy of entire genome sequencing. See Chapter 26.

Minimal tiling path (MTP) is a set of minimally overlapping BAC clones in the physical map picked for use in clone-by-clone sequencing of the entire genome. For most efficient whole genome sequencing, the ideal situation is to have minimally repeated sequencing and also cover all gaps so that the entire genome sequences can be assembled (Engler et al. 2003). Three different strategies were used to generate draft sequences for the human, mouse, and rat genomes, namely, the "clone-by-clone" for human, the whole genome shotgun (WGS) for mouse, and a hybrid strategy for rat (Lander et al. 2001, Waterston et al. 2002, Rat Genome Sequencing Project Consortium 2004). With the clone-by-clone strategy, individual BAC clones are shotgun-sequenced, the sequence of each BAC clone is generated by assembling the corresponding sequencing reads, and the sequence of the whole genome is obtained by merging overlapping BAC clone sequences. To minimize sequencing the

same genomic region multiple times, a set of minimally overlapping clones covering the whole genome is determined beforehand, and such a set of BAC clones are called MTP. BES serving as sequence connectors are essential for the identification of a minimum tiling path of BAC clones for whole genome sequencing (Siegel et al. 1999, Engler et al. 2003, Chen et al. 2004).

Two approaches are available traditionally for the selection of the MTP. The first is a map-based approach as used by the *Caenorhabditis elegans* project (Coulson et al. 1986) and human chromosomes 1, 6, 20, 22, and X (Bentley et al. 2001). Fingerprints of clone pairs that appear to have a minimum overlap are analyzed in the FPC Gel Image display. Viewing the gel images of neighboring clones helps identify false-positive and false-negative bands. With this method, a complete MTP can be picked before any sequencing is started, so that all clones can be sequenced in parallel. However, the amount of overlap may be large (e.g., 47.5 kb overlap for the MTP picked by the International Human Genome Consortium), because many BAC contigs were constructed with fingerprinting using restriction enzyme with 6 bp recognition sequences. Clones need to share multiple bands to have enough evidence of overlap, and on average, one band is produced every 4,096 bp. In addition, manual selection of one minimally overlapping pair is highly labor intensive.

The second approach is based on BAC-end sequences (Venter et al. 1996). In this approach, a seed clone is picked and completely sequenced. The BAC-end sequences are queried for hits to the finished sequence. The BAC clones with minimal overlapping is picked for extending sequencing. The new set of MTP clones is sequenced, and new clones are picked off the ends of this set based on minimal overlapping with BAC-end sequences. This process is repeated until the entire region is sequenced (Figure 15.4).

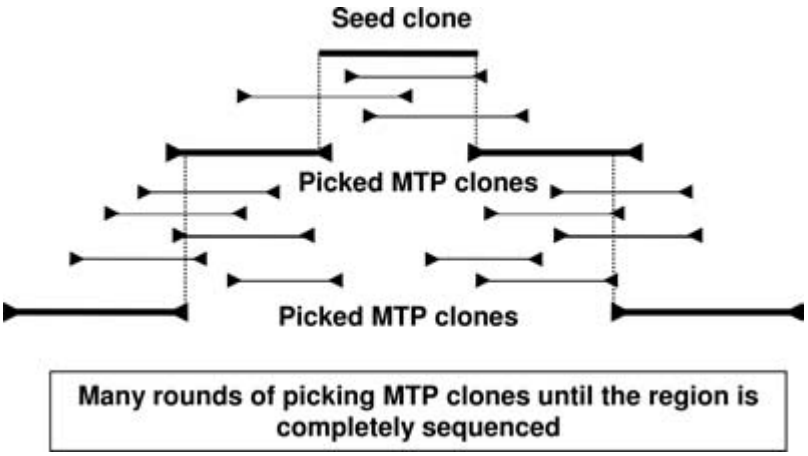


Figure 15.4. Selection of minimal tiling path (MTP) clones using BAC-end sequences (arrows). The MTP clones are selected by picking BAC clones with minimal sequence overlapping between their BAC-end sequences (triangle at the end of BAC clones) and the already sequenced seed clone. After picking the two MTP clones, they are completely sequenced and aligned to the BAC-end sequences to identify the next set of MTP clones with minimal overlap with the first round of MTP clones. This process is continued until complete sequencing of the region. Note that the BAC-end sequences can be coupled to the restriction fingerprinting information to minimize false positives and to reduce the overlapping regions.

By use of sequence information rather than the restriction fingerprints, the amount of overlap required for MTP pairs is reduced drastically. However, the risk of false positives is high, especially when considering repeats and errors in the low-quality BES. This approach, when coupled with the BAC-based physical map (i.e., the hybrid approach) has been the most effective for the selection of MTP. The MTP with minimal overlapping and maximal accuracy guarantees efficient sequencing of the genome and proper assembly of the entire genome sequences once they are sequenced.

Recently, a software pipeline CLONEPICKER was developed that integrates sequence data with BAC clone fingerprints to dynamically select a minimal overlapping clone set covering the whole genome (Chen et al. 2004). CLONEPICKER uses restriction enzyme fingerprint data, BAC-end sequence data, and sequences generated from individual BAC clones as well as whole genome shotgun sequencing reads. Incorporation of all available genomic information in this software allows accurate selection of tiling path with truly minimal overlaps.

Conclusion

BAC-end sequencing is simple enough not to refer to it as a genome technology. However, BAC-end sequences are extremely rich in genome information. The analysis of BAC-end sequences allows an unbiased sampling of the genome as to its composition and architecture. Bioinformatic mining of BAC-end sequences allows the uncovering of the repeat structure of the genome and the identification of polymorphic microsatellites. Many markers discovered from BAC-end sequences can be used to integrate the genetic linkage maps and the physical map as they can be placed on both maps. BAC-end sequencing is one of the most efficient approaches to locate genes to physical maps. BAC-end sequences can be exploited for comparative genome analysis including characterization of evolutionarily conserved syntenies. For many aquaculture species, the assessment of their genomes by BAC-end sequencing is probably as good as they can get as their genomes may never be sequenced.

Acknowledgments

Research in my laboratory is supported by grants from the USDA NRI Animal Genome and Genetic Mechanisms Program, the USDA NRI Basic Genome Reagents and Tools Program, the Mississippi-Alabama Sea Grant Consortium, the Alabama Department of Conservation, the USAID, and the BARD.

References

Agarwal P and DJ States. 1994. The Repeat Pattern Toolkit (RPT): analyzing the structure and evolution of the *C. elegans* genome. Proc Int Conf Intell Syst Mol Biol, 2, pp. 1–9.

Altschul SF, TL Madden, AA Schaffer, J Zhang, Z Zhang, W Miller, and DJ Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 25, pp. 3389–3402.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res, 27, pp. 573–580.

Bentley DR, P Deloukas, A Dunham, L French, SG Gregory, SJ Humphray, AJ Mungall, MT Ross, NP Carter, and I Dunham. 2001. The physical maps for sequencing human chromosomes 1, 6, 9, 10, 13, 20 and X. Nature, 409, pp. 942–943.

Chen R, E Sodergren, GM Weinstock, and RA Gibbs. 2004. Dynamic building of a BAC clone tiling path for the Rat Genome Sequencing Project. Genome Res, 14, pp. 679–684.

Coulson A, J Sulston, S Brenner, and J Karn. 1986. Towards a physical map of the genome of the nematode *C. elegans*. Proc Natl Acad Sci, 83, pp. 7821–7825.

Delcher AL, S Kasif, RD Fleischmann, J Peterson, O White, and SL Salzberg. 1999. Alignment of whole genomes. Nucleic Acids Res, 27, pp. 2369–2376.

Devereux J, P Haeberli, and O Smithies. 1984. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res, 12, pp. 387–395.

Engler FW, J Hatfield, W Nelson, and CA Soderlund. 2003. Locating sequence on FPC maps and selecting a minimal tiling path. Genome Res, 13, pp. 2152–2163.

Ewing B and P Green. 1998. Base-calling of automated sequencer traces using Phred. II. Error probabilities. Genome Res, 8, pp. 186–194.

Ewing B, L Hillier, MC Wendl, and P Green. 1998. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. Genome Res, 8, pp. 175–185.

Kurtz S, JV Choudhuri, E Ohlebusch, C Schleiermacher, J Stoye, and R Giegerich. 2001. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res, 29, pp. 4633–4642.

Kurtz S, E Ohlebusch, C Schleiermacher, J Stoye, and R Giegerich. 2000. Computation and visualization of degenerate repeats in complete genomes. Proc Int Conf Intell Syst Mol Biol, 8, pp. 228–238.

Lander ES, LM Linton, B Birren, C Nusbaum, MC Zody, J Baldwin, K Devon, K Dewar, M Doyle, W FitzHugh, R Funke, D Gage, K Harris, A Heaford, J Howland, L Kann, J Lehoczky, R LeVine, P McEwan, K McKernan, et al. 2001. Initial sequencing and analysis of the human genome. Nature, 409, pp. 860–921.

Larkin DM, A Everts-van der Wind, M Rebeiz, PA Schweitzer, S Bachman, C Green, CL Wright, EJ Campos, LD Benson, J Edwards, L Liu, K Osoegawa, JE Womack, PJ de Jong, and HA Lewin. 2003. A cattle-human comparative map built with cattle BAC-ends and human genome sequence. Genome Res, 13, pp. 1966–1972.

Lefebvre A, T Lecroq, H Dauchel, and J Alexandre. 2003. FORRepeats: detects repeats on entire chromosomes and between genomes. Bioinformatics, 19, pp. 319–326.

Liu L, L Roinishvili, X Pan, Z Liu, and C Kumar. 2000. GPMS: A web based genome project management system. The Proceeding of 4th World Multiconference on Systematics, Cybernectics, and Informatics SCI 2000, pp. 62–67.

Liu ZJ, P Li, and R Dunham. 1998. Characterization of an A/T-rich family of sequences from the channel catfish (*Ictalurus punctatus*). Mol Mar Biol Biotechnol, 7, pp. 232–239.

Rivals E, O Delgrange, JP Delahaye, M Dauchet, MO Delorme, A Henaut, and E Ollivier. 1997. Detection of significant patterns by compression algorithms: the case of approximate tandem repeats in DNA sequences. Comput Appl Biosci, 13, pp. 131–136.

Serapion J, H Kucuktas, J Feng, and Z Liu. 2004. Bioinformatic mining of Type I microsatellites from expressed sequence tags of channel catfish (*Ictalurus punctatus*). Mar Biotechnol, 6, pp. 364–377.

Siegel AF, B Trask, JC Roach, GG Mahairas, L Hood, and G van den Engh. 1999. Analysis of sequence-tagged-connector strategies for DNA sequencing. Genome Res, 9, pp. 297–307.

Sreenu VB, G Ranjitkumar, S Swaminathan, S Priya, B Bose, MN Pavan, G Thanu, J Nagaraju, and HA Nagarajaram. 2003. MICAS: A fully automated web server for microsatellite extraction and analysis from prokaryote and viral genomic sequences. Applied Bioinformatics, 2, pp. 165–168.

The International Human Genome Mapping Consortium. 2001. A physical map of the human genome. Nature, 409, pp. 934–941.

Venter JC, HO Smith, and L Hood. 1996. A new strategy for genome sequencing. Nature, 381, pp. 364–366.

Waterston RH, K Lindblad-Toh, E Birney, J Rogers, JF Abril, P Agarwal, R Agarwala, R Ainscough, M Alexandersson, P An, SE Antonarakis, J Attwood, R Baertsch, J Bailey, K Barlow, S Beck, E Berry, B Birren, T Bloom, P Bork, et al. 2002. Initial sequencing and comparative analysis of the mouse genome. Nature, 420, pp. 520–562.

Xu P, S Wang, L Liu, E Peatman, B Somridhivej, J Thimmapuram, G Gong, and Z Liu. 2006. Generation of channel catfish BAC end sequences for marker development and assessment of syntenic conservation with other fish species. Anim Genet, 37, pp. 321–326.

Zhang Z, S Schwartz, L Wagner, and W Miller. 2000. A greedy algorithm for aligning DNA sequences. J Comput Biol, 7, pp. 203–214.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 16
Genomescape: Characterizing the Repeat Structure of the Genome

Zhanjiang Liu

Understanding genome landscape, the genomescape, is important for genome biology because repetitive elements form a major fraction of eukaryotic genomes. Specifically, characterization of repeat structures of a genome can significantly reduce the complexities involved in genome studies, facilitating linkage mapping, physical mapping, comparative mapping, and laying the groundwork for whole genome sequencing. Repetitive elements, once dismissed as mere junk DNA, are now recognized as "drivers of genome evolution" (Kazazian 2004) whose evolutionary role can be "symbiotic (rather than parasitic)" (Holmes 2002). Examples of potentially beneficial evolutionary events in which repetitive elements have been implicated include genome rearrangements (Kazazian 2004), gene-rich segmental duplications (Bailey et al. 2003), random drift to new biological function (Kidwell and Lisch 2001, Brosius 2003), and increased rates of evolution during times of stress (Capy et al. 2000, Shapiro 1999). For these reasons, the study of repeat elements and their evolution is emerging as a key area in genome biology (Zhi et al. 2006).

Based on their known functions, repetitive elements can be divided into functional repetitive gene clusters and nonfunctional repetitive elements; based on their arrangements and distribution in the genome, they can be divided into tandem repetitive elements and dispersed elements; and based on their abundance, they can be divided into high, intermediate, and low abundance repetitive elements. In this chapter, I will briefly introduce several major classes of repetitive elements, and present several selected methodologies for the characterization of genomescape.

Classical Approaches for the Characterization of Repeat Structures of Genome

Roy Britten and his colleague were the first to study the genome using genomic approaches (Britten and Kohne 1968). They studied reassociation kinetics of genomic DNA in solution using a technique termed "Cot analysis." When a solution of denatured genomic DNA is placed in an environment conducive to renaturation, the rate at which a particular sequence reassociates is correlated to the copy number of the sequence present in the genome. This principle forms the basis of Cot analysis (Britten and Kohne 1968 and for a recent review of Cot principles, see Peterson et al. 2002). Cot value is equal to the product of nucleotide concentration in moles per liter (C_0 or Co) and reassociation time in seconds (t), and, if applicable, a factor based upon the cationic concentration of the buffer. Samples of sheared genomic DNA are heat-denatured and

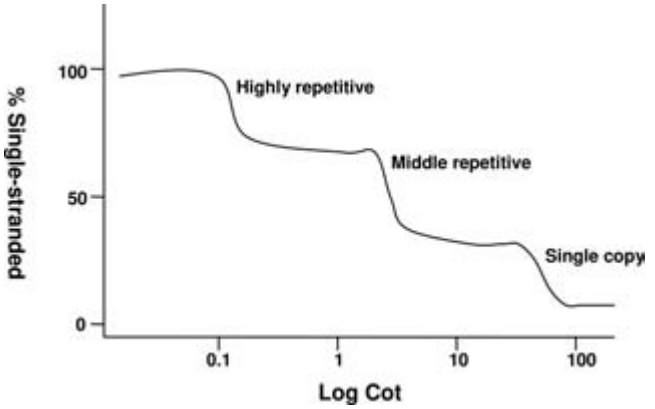


Figure 16.1. A schematic presentation of the Cot reassociation curve.

allowed to reassociate to different Cot values. For each sample, renatured DNA is separated from single-stranded DNA using hydroxyapatite (HAP) chromatography, and the percentage of the sample that has not reassociated (percentage of ssDNA) is determined. The logarithm of a sample's Cot value is plotted against its corresponding percentage of ssDNA to yield a Cot point, and a graph of Cot points ranging from the start to completion of the reassociation is called a Cot curve (Peterson et al. 1998) (see Figure 16.1). Mathematical analysis of a Cot curve permits estimation of genome size, the proportion of the genome contained in the single-copy and repetitive DNA components, and the kinetic complexity of each component Peterson et al. 2002). Cot curves (Figure 16.1) reflect the major three fractions of genomic DNA: the highly repetitive fraction that renatures rapidly; the intermediate repetitive fraction that renatures relatively slower; and the low copy number repetitive elements and the single copy genes that reassociate very slowly.

Cot analysis also allows many of the basic characteristics of genomes to be compared between organisms. Based on the reassociation kenetics, interspecific comparison of Cot data has provided considerable insight into the structure and evolution of eukaryotic genomes (e.g., Britten and Kohne 1968). Even though the Cot analysis was developed over three decades ago even before the emergence of the molecular biology era, its principles are still highly useful for genome characterizations in the genomics era.

Characterization of Tandem Repeats with High Genomic Copy Numbers

Tandem repetitive noncoding DNA sequences make up a large fraction of the genomes of eukaryotes. The sequence complexities of these repetitive sequences can vary from a single base pair (bp) to over two kilobase pairs (kb). Copies of the tandem repeats can vary greatly, ranging from a total size of 100 bp to more than 100 mega base pairs (Mb) (Milklos and Gill 1982). Based on the repeat lengths and array sizes, they have been divided into three classes: microsatellite, minisatellite, and satellite

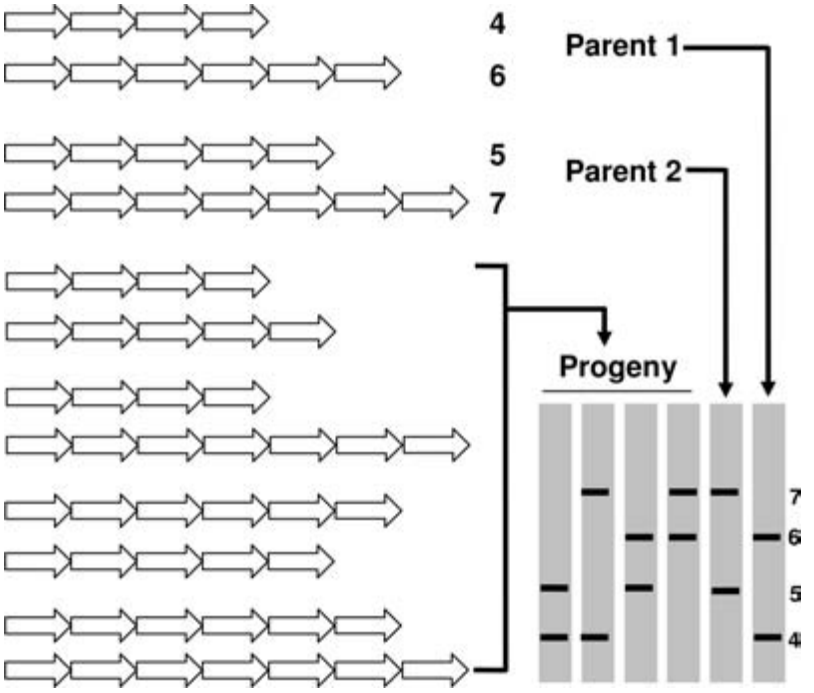


Figure 16.2. Minisatellites can be used as polymorphic markers. They are inherited as co-dominant markers.

DNAs (Levinson and Gutman 1987). As their names indicate, microsatellites include repetitive sequences with very simple sequence complexity and short arrays of the repeat (see Chapter 5). Satellite DNA exhibits more sequence complexity (generally 100 bp or longer) with long arrays of total repeat lengths. Minisatellites show intermediate features between satellites and microsatellites. The three classes of the repetitive sequences also differ in where they reside in the eukaryotic genomes. Satellite DNA is often concentrated in regions of very low recombination (Charlesworth et al. 1986) such as heterochromatic regions near centromeres and telomeres where meiotic recombination is suppressed. Minisatellite sequences are usually located in the euchromatic portions of chromosomes. Recombination at minisatellite loci appears to be higher than at satellite DNA loci (Stephan 1989, Stephan and Cho 1994). Variable number of tandem repeats (VNTR) can be regarded as minisatellites with 5–50 repeats. VNTRs are hot spots for meiotic recombination. They exhibit high levels of polymorphism, and therefore, are useful as polymorphic markers (Figure 16.2). It is believed that the high mutation rate leading to polymorphism is caused by uneven meiotic recombination. Microsatellites are tandem repeats of very short sequences (1–6 bp, as shown in Chapter 5) interspersed in various regions in chromosomes including within or near expressed genes (e.g., Gong et al. 1997, Serapion et al. 2004).

Technically, the identification of tandemly duplicated repeats is the most straightforward. These elements can be readily detected by hybridizations. For highly abundant repetitive elements, they can be observed by digesting genomic DNA with restriction

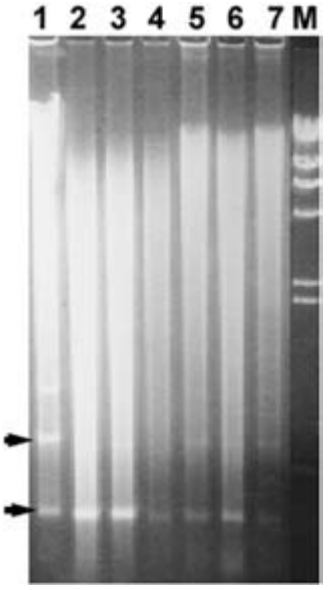


Figure 16.3. Detection of repetitive elements by direct restriction enzyme digestion. The example shown here was adopted from Liu and others (1998) from a catfish study. Genomic DNA from various strains of catfish (lanes 1–7) was digested with restriction endonuclease *Xba* I, separated on an agarose gel, and visualized by ethidium bromide staining, with molecular weight (M) on the last lane. Note the generation of discrete bands (arrows) above the background of a smear.

enzymes; direct restriction enzyme digestion of genomic DNA should produce bands above the background of a smear (Figure 16.3). Using this approach, two types of repetitive elements were found in the zebrafish genome, accounting for 5% and 0.5% of the zebrafish genome (He et al. 1992). Using a similar approach, Liu and others (1998) identified a family of A/T-rich *Xba* elements that are arranged in tandem and account for 5–6% of the catfish genome.

Differentiation of tandem repeat from interspersed repeats (see below) can be accomplished through several approaches. As discussed above, tandem repeats tend to generate discrete DNA bands (if restriction sites exist within the repeat unit). However, many interspersed repeats also generate discrete DNA bands if two or more restriction sites for the same enzyme exist within the repeat unit. One way to differentiate these two types is through the use of partial restriction digest. For tandem repeats, with limited amounts of restriction enzyme, partial digestion products will generate a band pattern exhibiting monomers, dimers, trimers, tetramers, pentamers, hexamers, etc., of the repeat unit. With incremental amounts of restriction enzyme, the higher molecular weight bands will disappear resulting in the final product, the monomer of the repeat unit. This is not the case for interspersed repeats that generate smears regardless of the amount of restriction enzyme used (Figure 16.4).

Another approach to differentiate tandem repeats from interspersed repeats is the use of fluorescent in situ hybridization. See Chapter 17. Tandem repeat types generate focused hybridization patterns with very high fluorescent signals, whereas interspersed repeats produce hybridization patterns that are scattered throughout the genome.

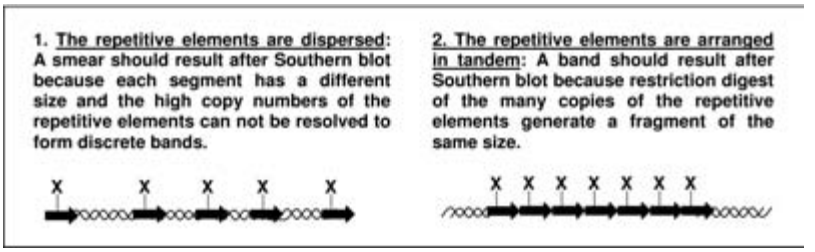


Figure 16.4. Two different types of arrangements and organizations of repetitive elements can be differentiated by Southern blot analysis followed by partial restriction enzyme digest.

Characterization of Intersperse Repetitive Elements

Dispersed repetitive elements included mostly long interspersed elements (LINEs), short interspersed elements (SINEs), and DNA transposons. LINEs and SINEs are retrotransposable DNA elements that duplicate in the genome via a "copy and paste" mechanism. Functional LINE elements encode a reverse transcriptase, allowing the transcribed LINE RNA to be converted into DNA, and an endonuclease for cleaving the genomic DNA at the new insertion site. SINE elements are shorter repetitive elements that require the "copy and paste" machinery of the LINEs. SINEs resemble small nuclear RNAs such as tRNA, and are transcribed by RNA polymerase III. SINEs then rely on the reverse transcriptase and endonuclease from the LINE elements to reinsert and multiply in the genome.

In addition to LINEs and SINEs, another major class of interspersed elements is DNA transposons. DNA-mediated transposons move through DNA intermediates and depend on transposases. While retrotransposons are more common and abundant in vertebrates, DNA transposons are more common in bacteria, plants, and invertebrates such as fruit flies and nematodes. DNA transposons harbor inverted repeats required *in cis* for transposition. Autonomous DNA transposons also harbor a functional transposase gene to encode the transposase enzyme required for transposition *in trans.* Despite a wide range of distribution (Radice et al. 1994), DNA transposons were not discovered in vertebrates until the 1990s (Heierhorst et al. 1992, Henikoff 1992). Since the discovery of the first vertebrate DNA transposons in channel catfish by database analysis (Henikoff 1992), search for an active copy of DNA transposons has been a keen research area. Recently, DNA transposons in vertebrates have been isolated from a number of vertebrate species, but they have been best characterized in teleosts (Heierhorst et al. 1992; Henikoff 1992; Goodier and Davidson 1994; Radice et al. 1994; Izsvak et al. 1995; Ivics et al. 1996; Lam et al. 1996; Koga and Hori 2000, 2001). Members of the *Tc1*/mariner superfamily represent most of the DNA transposons discovered in vertebrates. In addition to teleosts, *Tc1*/*mariner*, *tiggers*, and other transposon-like elements have been discovered in amphibians (Lam et al. 1996) and human genomes (Oosumi et al. 1995, Smit and Riggs 1996).

Until recently *Tc1*-like elements were all identified as nonautonomous from vertebrate genomes, resulting from extensive insertions/deletions, and in-frame termination codons in their transposase genes (Lam et al. 1996; Ivics et al. 1996, 1997). Kawakami

and his colleagues (2000) identified the first endogenous autonomous transposon, an *Ac*-like *Tol2* element from the Japanese *medaka* fish that can transpose both transiently in fertilized zebrafish eggs and in the zebrafish germ lineage (Koga and Hori 2000, 2001). Several additional transposases with intact open reading frames (ORF) were recently identified from the amphibian *Xenopus tropicalis* and may still be involved in active transposition (Sinzelle et al. 2005).

Characterization of dispersed repetitive elements is a great challenge. First, these elements are highly interspersed throughout the genome; second, they are present in high copy numbers, making hybridization-based approaches ineffective; third, their sequences are highly related, but differ between copies. Genomic approaches using data mining appears to be much more effective for the characterization of dispersed repetitive elements. Their abundance, composition, distribution, and nature of the sequences can all be analyzed using data mining software. REPuter (Kurtz and Schleiermacher 1999) (http://www.genomes.de/), Basic Local Alignment Search Tool (BLAST), and RepeatMasker, among many others, are frequently used to assess the repeat structures of the genome. However, the effectiveness of the software depends on the existence of repeat databases. REPuter also allows identification of new repeat types in a given organism. Often searches can be conducted by searching for the presence of repeated sequences with parameters set at, for example, a minimum length of 20 bp and a sequence variation of 10% (90% conserved). Self BLAST searches are also effective for the identification of repeated sequences in an organism upon generation of genomic sequence surveys.

RepeatMasker is highly effective for the identification of common types of repeats. For instance, the entire genome sequences of zebrafish, fugu, and *Tetraodon* are already available. Repeat masking using repeat libraries of these species allowed a good estimation of the shared repeats of catfish with these species (Xu et al. 2006), as shown in Table 16.1. However, RepeatMasker lacks the ability to detect species-specific and novel repeat types.

At the genomic level, dot plot alignments are also often used for the detection of repetitive elements (Sonnhammer and Durbin 1995). Within a genome, a dot plot illustrates duplicated sequences: repeated sequences and gene families. Between genomes, dot plots reveal levels of conservation of sequence and of gene orders.

Characterization of Gene Clusters

Many genes with related functions are arranged in clusters and are duplicated into many copies. As a rule of thumb, genes whose products are most highly demanded by cellular functions likely have multigene families, and often they are arranged together in tandem repeats. Some examples of tandemly repeated gene families include globin genes, immunoglobulin genes, rRNA genes, tRNA genes, and histone genes. Such an arrangement is beneficial in order to provide coordinated expression. For instance, the ribosomal RNA is required for the structure of ribosomes that demand a constant 1:1 ratio. The genes for rRNA are organized not only together, but also transcribed under control of the same promoter. A single transcription unit (cistron) contains the 18S, 5.8S, and the 28S rRNA genes. In addition to functional genes, pseudogenes are often

Table 16.1. Repeat composition of the channel catfish genome as assessed by RepeatMasker of the BAC-end sequences using the combined repeat database of zebrafish and *Takifugu rubripes.*

Repetitive Elements	Number of Elements	Length Occupied	% of Sequence
Retroelements	1,972	349,917 bp	3.13
SINEs:	1,083	139,373 bp	1.24
Penelope	1	269 bp	0
LINEs:	643	135,520 bp	1.21
L2/CR1/Rex	525	107,885 bp	0.96
R1/LOA/Jockey	14	2,773 bp	0.02
R2/R4/NeSL	1	50 bp	0
RTE/Bov-B	37	6,137 bp	0.05
L1/CIN4	65	18,406 bp	0.16
LTR elements:	246	75,024 bp	0.67
BEL/Pao	12	3,604 bp	0.03
Gypsy/DIRS1	179	60,047 bp	0.54
Retroviral	21	3,726 bp	0.03
DNA transposons:	2,591	563,923 bp	4.57
hobo-Activator	220	23,205 bp	0.16
Tc1-IS630-Pogo	2,077	507,735 bp	4.12
En-Spm	3	160 bp	0
PiggyBac	57	6,399 bp	0.06
Tourist/Harbinger	77	7,982 bp	0.07
Unclassified:	9	751 bp	0.01
Total interspersed repeats:		914,591 bp	8.17
Satellites:	25	2,602 bp	0.02
Simple repeats:	5,798	289,211 bp	2.58
Low complexity:	3,202	127,064 bp	1.13

found in tandem clusters. Pseudogenes are defined by their possession of sequences that are related to those of the functional genes, but that cannot be translated into a functional protein. Pseudogenes are often denoted by ψ. They are formed when transcription signals such as CAAT box or TATA are abolished, or when the splicing junction is mutated so that the transcripts cannot be properly processed, or when in-frame premature termination codons are evolved. Most gene families have members that are pseudogenes constituting a minority of the family.

Histone Gene Clusters

In most organisms, five major types of histone proteins (H1, H2A, HZB, H3, and H4) combine to form a functional histone unit. Two structures are created by the dimerization of H2A/H2B and H3/H4. These structures then self-dimerize forming two tetramers (H2A-H2B-H2A-H2B and H3-H4-H3-H4) and the two tetramers join to create the cylindrical histone octamer. About 145 bp of DNA coils around this functional

octamer to create a nucleosome, the basic unit of chromatin structure. Due to this highly conserved function, histone genes are among the most evolutionarily conserved genes.

Histones make up the chromatin structure of the genome and they are highly expressed to meet the high demand. High levels of histone expression are accomplished by gene duplications that form large gene families existing in tandem copies in the genome. The number of histone genes varies among species. In the yeast *Saccharomyces cerevisiae*, there are two identical copies of each of H2A, H2B, H3, and H4 (Mardian and Isenberg 1978). The H1 is present as a single copy, making a total of nine histone genes in the 12 Mb of the yeast. In higher eukaryotic organisms, two types of histone gene organization are commonly seen: (1) the histone genes are clustered and tandemly repeated, or (2) the histone genes are clustered, but no tandem duplication is seen. Sea urchins, worms, frogs, fish, and *Drosophila* all have histone genes organized in a tandemly duplicated manner. In these species, the core histone genes are in repeating units ranging from 750 to 2,000 times in the genome. Species such as chicken, human, and mouse have histone genes that are located in various clusters in the genome without a clear repeating pattern. For these species, the number of histone genes is between 50 and 200, and, while the histone genes are located in a cluster, they show little resemblance to a tandemly repeating unit.

The highly repetitive nature of histone genes makes characterization difficult. However, the highly conserved sequences offer advantages for the adoption of polymerase chain reaction (PCR). PCR primers can be designed based on gene or EST sequences of histone genes from the species of interest, or even from a closely related species. The gene sequences can then be obtained without any problem. The gene arrangement of the histone genes within the repeat unit can be determined by using PCR primers in the intergenic regions, allowing amplification across different histone genes. In rainbow trout and Atlantic salmon, the histone genes are arranged in an order of H4-H2B-H1-H2A-H3, with an external gene region of approximately 13 kb. Along with the gene region of approximately 5.1 kb, the histone cluster repeat unit in the salmonids appears to be approximately 19–20 kb (Pendas et al. 1994; Ng and Davidson, personal communications).

Determination of the number of units of a given repeat can be difficult. Most often, this can be assessed by hybridization of histone probes to bacterial artificial chromosome (BAC) filters. For instance, if 600 BACs are positive to the histone gene probes on a 10 $\times$ BAC filter, 60 BACs per genome contain histone genes. The total size of the 60 BACs can be estimated from the average size of the BAC library, minus the average overlapping regions of BACs. For instance, the channel catfish CHORI BAC library has an average insert size of 161 kb. The maximum size that can be covered by the 60 BACs is 60×161 kb $= 9{,}660$ kb. If the repeating unit is 20-kb long, then the maximal number of repeat units is $9{,}660/20 = 483$. However, this estimate is on the high side because the 600 BACs have large overlapping regions. The number of repeat units can be adjusted by subtracting the average overlapping region out of the average length of the BAC. For example, if on average, 30 kb overlapping regions exist in the BAC-based contigs, then the total length covered by the 60 BAC/genome is 131 kb $\times$ 60 $= 7{,}860$ kb. Considering that each repeat unit is 20 kb long, then the number of repeat units is $7{,}860/20 = 393$. Another approach that can provide a more accurate estimation for the number of repeat units is perhaps through the use of quantitative real time PCR.

The rDNA Cistron

Ribosomal RNAs are involved in structures of ribosomes for the translation machinery. There are four rRNAs: 18S, 5.8S, 28S, and 5S rRNA. The 18S, 5.8S, and 28S rRNAs are encoded as a cistronic unit in the genome. This rRNA precursor undergoes posttranscriptional cleavage to process the cistronic RNA into three functional rRNAs. The 18S rRNA becomes a part of the small ribosomal subunit whereas the 28S and 5.8S rRNA are components of the large ribosomal subunit.

Over 80% of the cellular RNA mass is composed of rRNAs. The large demand for rRNA is met in the cells by gene duplication of these genes to form the large tandem repeat. In addition, their gene products are required in a stoichiometric one to one ratio, and thus their expression must be coordinated. The evolutionary processes have produced some of the most efficient ways for this coordination by using cistrons (i.e., these genes are present together and their expression is under the control of the same promoter). Cistrons are common in prokaryotic organisms, but rare in eukaryotes. The rDNA cistrons are one of the rare examples for the use of cistrons in eukaryotic organisms.

The rDNA organization varies according to species. In humans, the 250 or so copies of the rDNA repeats are clustered on acrocentric chromosome arms of 13p, 14p, 15p, 21p, and 22p. The location of rDNA in fish and their organization is not yet well understood. In zebrafish, it is believed that 6–8 chromosomes harbor rDNA repeat clusters, but 6–12 chromosomes may harbor rDNA clusters in sturgeons (Gornung et al. 1997, Phillips and Reed 2000, Fontana et al. 2003). PCR primers can be designed for 18S, 28S, and 5.8S rRNAs based on gene sequences of these genes from the species of interest. In order to estimate the cistron size, external primers facing toward the external transcribed spacer can be used. As discussed above with the histone gene clusters, the estimation for the number of repeat units can be difficult, but BAC-based hybridization can provide a fairly accurate estimate. The Atlantic salmon genome is estimated to contain 1,000–1,500 copies of the rDNA repeats (Moran et al. 1997). Very similar approaches can be used for the characterization of tRNA genes that also exist in tandem repeats.

It is logical to think that highly duplicated genes should be subjected to more mutations due to reduced evolutionary pressure. How then can many highly duplicated genes such as rRNA or histone genes maintain their integrity? Several theories exist including coincidental evolution (concerted evolution or coevolution to explain the situation). Coincidental evolution describes the ability of two related genes to evolve together as though constituting a single locus. When members of a repetitive family are compared, greater sequence similarity is found within a species than between species, suggesting that members within a repetitive family do not evolve independently of each other. The detailed mechanisms may differ: a mutation can be removed; or the mutated copy takes over leading to both copies with a mutation; or homogenization by enzymes through strand change and editing (Liao 1999). Another interesting hypothesis is sudden correction: Every so often the entire gene cluster is replaced by a new set of copies derived from one or a few of the copies. While these models may explain the high sequence conservation of rRNA and histone genes, the truth may be that the evolutionary pressure for these genes has never been reduced as assumed by many. The demand for the gene products of highly duplicated genes may

be so high that each duplicated copy is still absolutely required so that any mutation of even a single copy may lead to a detrimental impact on the organism.

Determination of Copy Numbers of Repetitive Elements

It is difficult to determine the exact copy numbers of repetitive elements. However, their copy numbers in the genome can be assessed by hybridization, or in some cases by real time PCR (e.g., Chen et al. 2006). The choice of approaches depends on the nature of the repetitive elements. If the repetitive elements or genes have exactly the same or highly conserved sequences, then real time quantitative PCR can be used to determine the copy numbers. However, in most cases, the repetitive elements are highly related in sequence, but carry many base substitutions, or exist as remnants with only part of the entire sequences. In this later case, hybridization is an effective strategy to assess the copy numbers. The copy number of the repetitive elements of interest can be assessed by comparing the hybridization signals of the genomic sample with that of a serially diluted sample with known copy number of the same target molecules. The major problem of a hybridization-based approach for the assessment of copy numbers is the saturation of signals (the after-black-is-black nature of signals). Therefore, serial dilutions should be made with genomic DNA, as well as the control DNA such as a plasmid containing the target sequence. Each copy of plasmid contains one copy of the target sequence. For instance, the hybridization signal of 2 μg genomic DNA is equivalent to that of 8 nanogram (ng) plasmid containing one copy of the target sequence. There are 1.4×10^9 molecules (copies) of plasmids in the 8 ng plasmid DNA (moles of plasmid = 8×10^{-9} gram/molecular weight [5,000 bp × 660/bp], number of molecules = moles × 6.023×10^{23} per mole). There are 2,000,000 genomes in 2 μg of genomic DNA (the catfish genome size is 1 pg DNA per cell). Thus, the copy number of the repetitive element = 1.4×10^9 copies/2,000,000 = 700 copies. When determining copy numbers of repetitive elements, one should carefully select the hybridization probes so that they cover the typical region of the repetitive elements under consideration.

Conclusion

Repetitive elements comprise a major fraction of eukaryotic genomes. In humans, over 50% of their 3 billion bp genome is composed of repetitive elements. The proportion of repetitive elements in the genomes of aquaculture species is yet to be discovered. Once regarded as junk DNA, repetitive elements are once again gaining their popularity not only because of their abundance, but also because of their potentially unrealized biological functions. Understanding of repetitive elements and their organizations is by no means an easy task. Many of the current approaches involve bioinformatic analysis, though the role of solid experimental approaches cannot be neglected. Recent studies also indicate that regulation of biological functions can also be achieved through gene copy numbers, thus accurate determination of gene families and their copy numbers may prove to be important. Understanding of the repeat structure in aquaculture species will facilitate studies in linkage mapping, physical

mapping, and lay the foundation for entire genome sequencing. Even after the entire genome is sequenced, correct genome assembly will rely on the understanding and comprehension of repeat structures of the genome.

Acknowledgments

Research in my laboratory is supported by grants from the USDA NRI Animal Genome and Genetic Mechanisms Program, the USDA NRI Basic Genome Reagents and Tools Program, the Mississippi-Alabama Sea Grant Consortium, the Alabama Department of Conservation, the USAID, and the BARD.

References

Bailey JA, G Liu, and EE Eichler. 2003. An *Alu* transposition model for the origin and expansion of human segmental duplications. Am J Hum Genet, 73, pp. 823–834.

Britten RJ and DE Kohne. 1968. Repeated sequences in DNA. Science, 161, pp. 529–540.

Brosius J. 2003. How significant is 98.5% "junk" in mammalian genomes. Bioinformatics, 19(Suppl 2), II35.

Capy P, G Gasperi, C Biemont, and C Bazin. 2000. Stress and transposable elements: co-evolution or useful parasites? Heredity, 85, pp. 101–106.

Charlesworth B, C Langley, and W Stephan. 1986. The evolution of restricted recombination and the accumulation of repeated DNA sequences. Genetics, 112, pp. 947–962.

Chen Q, M Book, X Fang, A Hoeft, and F Stuber. 2006. Screening of copy number polymorphisms in human beta-defensin genes using modified real-time quantitative PCR. J Immunol Methods, 308, pp. 231–240.

Fontana F, M Lanfredi, L Congiu, M Leis, M Chicca, and R Rossi. 2003. Chromosomal mapping of 18S–28S and 5S rRNA genes by two-colour fluorescent *in situ* hybridization in six sturgeon species. Genome, 46, pp. 473–477.

Gong Z, T Yan, J Liao, SE Lee, J He, and CL Hew. 1997. Rapid identification and isolation of zebrafish cDNA clones. Gene, 201, pp. 87–98.

Goodier JL and WS Davidson. 1994. *Tc1* transposon-like sequences are widely distributed in salmonids. J Mol Biol, 241, pp. 26–34.

Gornung E, I Gabrielli, S Cataudella, and L Sola. 1997. CMA3-banding pattern and fluorescence *in situ* hybridization with 18S rRNA genes in zebrafish chromosomes. Chromosome Res, 5, pp. 40–46.

He L, Z Zhu, AJ Faras, KS Guise, PB Hackett, and AR Kapuscinski. 1992. Characterization of *Alu* I repeats of zebrafish (*Brachydanio rerio*). Mol Marine Biol Biotechnol, 1, pp. 125–135.

Heierhorst J, K Lederis, and D Richter. 1992. Presence of a member of the *Tc1*-like transposon family from nematodes and *Drosophila* within the vasotocin gene of a primitive vertebrate, the Pacific hagfish *Eptatretus stouti.* Proc Natl Acad Sci USA, 89, pp. 6798–6802.

Henikoff S. 1992. Detection of Caenorhabditis transposon homologs in diverse organisms. New Biol, 4, pp. 382–388.

Holmes I. 2002. Transcendent elements: whole-genome transposon screens and open evolutionary questions. Genome Res, 12, pp. 1152–1155.

Hong SH, JS Kim, SY Lee, YH In, SS Choi, J-K Rih, CH Kim, H Jeong, CG Hur, and JJ Kim. 2004. The genome sequence of the capnophilic rumen bacterium *Mannheimia succiniciproducens.* Nat Biotechnol, 22, pp. 1275–1281.

Ivics Z, PB Hackett, RH Plasterk, and Z Izsvak. 1997. Molecular reconstruction of *Sleeping Beauty,* a *Tc1*-like transposon from fish, and its transposition in human cells. Cell, 91, pp. 501–510.

Ivics Z, Z Izsvak, A Minter, and PB Hackett. 1996. Identification of functional domains and evolution of *Tc1*-like transposable elements. Proc Natl Acad Sci USA, 93, pp. 5008–5013.

Izsvak Z, Z Ivics, and PB Hackett. 1995. Characterization of a *Tc1*-like transposable element in zebrafish (*Danio rerio*). Mol Gen Genet, 247, pp. 312–322.

Kawakami K, A Shima, and N Kawakami. 2000. Identification of a functional transposase of the Tol2 element, an Ac-like element from the Japanese medaka fish, and its transposition in the zebrafish germ lineage. Proc Natl Acad Sci USA, 97, pp. 11403–11408.

Kazazian HH. 2004. Mobile elements: drivers of genome evolution. *Science,* 303, pp. 1626–1632.

Kidwell MG and DR Lisch. 2001. Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution Int J Org Evolution, 55, pp. 1–24.

Koga A and H Hori. 2000. Detection of *de novo* insertion of the medaka fish transposable element Tol2. Genetics, 156, pp. 1243–1247.

Koga A and H Hori. 2001. The Tol2 transposable element of the medaka fish: an active DNA-based element naturally occurring in a vertebrate genome. Genes Genet Syst, 76, pp. 1–8.

Kurtz S and C Schleiermacher. 1999. REPuter: fast computation of maximal repeats in complete genomes. Bioinformatics, 15, pp. 426–427.

Lam WL, P Seo, K Robison, S Virk, and W Gilbert. 1996. Discovery of amphibian *Tc1*-like transposon families. J Mol Biol, 257, pp. 359–366.

Levinson G and GA Gutman. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol, 4, pp. 203–221.

Liao D. 1999. Concerted evolution: molecular mechanism and biological implications. Am J Hum Genet, 64, pp. 24–30.

Liu Z, P Li, and RA Dunham. 1998. Characterization of an A/T-rich family of sequences from channel catfish (*Ictalurus punctatus*). Mol Mar Biol Biotechnol, 7, pp. 232–239.

Mardian JK and I Isenberg. 1978. Yeast inner histones and the evolutionary conservation of histone-histone interactions. Biochemistry, 17, pp. 3825–3833.

Milklos G and A Gill. 1982. Nucleotide sequences of highly repeated DNAs: compilation and comments. Genet Res, 39, pp. 1–30.

Moran P, KM Reed, J Perez, TH Oakley, RB Phillips, E Garcia-Vazquez, and AM Pendas. 1997. Physical localization and characterization of the *Bgl* I element in the genomes of Atlantic salmon (*Salmo salar* L.) and brown trout (*S. trutta* L.). Gene, 194, pp. 9–18.

Oosumi T, WR Belknap, and B Garlick. 1995. Mariner transposons in humans. Nature, 378, p. 672.

Pendas AM, P Moran, and E Garcia-Vazquez. 1994. Organization and chromosomal location of the major histone cluster in brown trout, Atlantic salmon and rainbow trout. Chromosoma, 103, pp. 147–52.

Peterson DG, WR Pearson, and SM Stack. 1998. Characterization of the tomato (*Lycopersicon esculentum*) genome using *in vitro* and *in situ* DNA reassociation. Genome, 41, pp. 346–356.

Peterson DG, SR Schulze, EB Sciara, SA Lee, JE Bowers, A Nagel, N Jiang, DC Tibbitts, SR Wessler, and AH Paterson. 2002. Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. Genome Res, 12, pp. 795–807.

Phillips RB and KM Reed. 2000. Localization of repetitive DNAs to zebrafish (*Danio rerio*) chromosomes by fluorescence *in situ* hybridization (FISH). Chromosome Res, 8, pp. 27–35.

Radice AD, B Bugaj, DH Fitch, and SW Emmons. 1994. Widespread occurrence of the *Tc1* transposon family: *Tc1*-like transposons from teleost fish. Mol Gen Genet, 244, pp. 606–612.

Serapion J, H Kucuktas, J Feng, and Z Liu. 2004. Bioinformatic mining of Type I microsatellites from expressed sequence tags of channel catfish (*Ictalurus punctatus*). Mar Biotechnol, 6, pp. 364–377.

Shapiro JA. 1999. Transposable elements as the key to a 21st century view of evolution. Genetica, 107, pp. 171–179.

Sinzelle L, N Pollet, Y Bigot, and A Mazabraud. 2005. Characterization of multiple lineages of *Tc1*-like elements within the genome of the amphibian *Xenopus tropicalis.* Gene, 349, pp. 187–196.

Smit AF and AD Riggs. 1996. *Tiggers* and DNA transposon fossils in the human genome. Proc Natl Acad Sci USA, 93, pp. 1443–1448.

Sonnhammer EL and R Durbin. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene, 167, pp. GC1–10.

Stephan W. 1989. Tandem-repetitive noncoding DNA: forms and forces. Mol Biol Evol, 6, pp. 198–212.

Stephan W and S Cho. 1994. Possible role of natural selection in the formation of tandem-repetitive noncoding DNA. Genetics, 136, pp. 333–341.

Xu P, S Wang, L Liu, E Peatman, B Somridhivej, J Thimmapuram, G Gong, and Z Liu. 2006. Generation of channel catfish BAC end sequences for marker development and assessment of syntenic conservation with fish model species. Anim Genet, 37, pp. 321–326.

Zhi D, BJ Raphael, AL Price, H Tang, and PA Pevzner. 2006. Identifying repeat domains in large genomes. Genome Biol, 7, p. R7.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 17
Genomic Analyses Using Fluorescence *In Situ* Hybridization

Ximing Guo, Yongping Wang, and Zhe Xu

Introduction

Fluorescence *in situ* hybridization (FISH) is the detection of specific DNA sequences through hybridization and fluorescence microscopy. Based on the base-pairing property of DNA molecules and direct visualization, FISH offers fast and specific mapping of DNA sequences to their *in situ* targets.

FISH has broad applications in biomedical research. It has been widely used in human genetics and genomics. It is used to label and identify chromosomes and subchromosomal regions using chromosome-specific probes (Lichter and Cremer 1992, Macville et al. 1997). Chromosomal identification is essential for studies on aneuploidy and chromosomal rearrangements. Changes in chromosomal number and structure often have serious consequences in gene expression and development. Trisomy 21 is a case where the gaining of an extra copy of chromosome 21 alters growth and mental development in humans leading to Down syndrome. Reciprocal translocation between chromosome 9 and 22 generates an abnormal chromosome 22, the *Philadelphia chromosome*, which is associated with chronic myelogenous leukemia (De Klein et al. 1982).

FISH is a powerful tool in genome mapping. It is widely used for chromosomal mapping of newly identified genes and physical mapping. Compared with linkage analysis, mapping genes by FISH is direct and fast. Chromosomal location is a basic property of every gene and an essential part to our understanding of genome structure and organization. Genetic and physical maps are different measures of chromosomes. By mapping DNA sequences directly to chromosomes, FISH can anchor genetic maps to their physical entities and therefore provides the ultimate verification. It can greatly empower contig assembly with the near-kilobase (kb) resolution of DNA-fiber FISH (Weier 2001).

FISH or immunocytochemistry-based *in situ* hybridization is also useful in determining tissue or cellular location of gene expression or pathogens. Spatial and temporal changes in gene expression are critical in studies of gene function. FISH provides excellent resolution of the spatial distribution of DNA or messenger RNA (mRNA) sequences.

As a powerful research tool, FISH has made significant contributions to human genetics. FISH has the potential for similar applications in aquaculture genomics. In recent years, interest in FISH has grown, and considerable progress has been made in applying FISH to aquaculture species. This chapter provides a general introduction of FISH, a review of studies in aquaculture species, and discussions on challenges and future directions.

Principles of FISH

Techniques for *in situ* hybridization were first developed in the late 1960s using isotope-based detection (John et al. 1969, Pardue and Gall 1969). The use of radioactive isotopes limited its wide acceptance. The development of enzymatic methods for introducing biotinylated nucleotides into DNA made nonisotopic labeling possible and consequently revolutionized *in situ* hybridization (Langer et al. 1981, Singer and Ward 1982, Leary et al. 1983). Fluorescence-based detection flourished in the 1980s, and the past two decades has brought tremendous development in FISH technology (Levsky and Singer 2003). Now standard reagents and kits are readily available from commercial suppliers. Protocols for diverse applications have been developed and optimized. FISH can be used in both fixed samples and live cells, and can detect from repetitive sequences such as satellite DNA to single-copy fragments of just a few hundred nucleotides (Schriml et al. 1999). Technologies for multicolor detection and quantitative analysis are now available, and so is high-resolution DNA fiber FISH.

Regardless of types of probes and targets, FISH relies on the base-pairing property of DNA sequences and the ability to detect fluorescence signals through microscopy. Successful FISH is dependent on the following basic steps:

1. proper preparation of target materials
2. production and labeling of DNA probes with a fluorescence dye or a reporter molecular
3. hybridization of the denatured probe and target DNA under proper conditions, and posthybridization fluorescence detection, with or without amplification, under microscope (Figure 17.1)

While the basic idea is simple, FISH involves many steps and variables. Closely following protocols is critically important. A thorough understanding of the principles behind each step is also helpful.

Target Material

The target material for FISH can be chromosomes, interphase nuclei, DNA fibers, cells, or tissues that carry the targeted DNA or RNA. Target materials have to be properly prepared so that the *in situ* site is accessible to FISH probes.

Chromosome is the highly condensed form of macromolecules consisting of DNA and histones. The quality of metaphase chromosomes is critically important for the success of FISH. Elongated chromosomes that are free of cytoplasmic material are ideal. The degree of chromosome decondensation directly affects signal strength and resolution. FISH using extended DNA fibers allows the mapping of DNA fragments of a few hundred base pairs (bp) with near-kb resolution (Weier 2001). To obtain elongated chromosomes, materials with high mitotic indexes are required. Cell lines are the preferred material for the preparation of elongated chromosomes. Because of the high mitotic index, cultured cells can be easily synchronized and harvested at metaphase. The availability of cell lines is one reason why FISH has been so successful in humans and other mammalian systems. In most aquaculture species, however, cell lines are not available. Sometimes primary cell cultures can be obtained and used

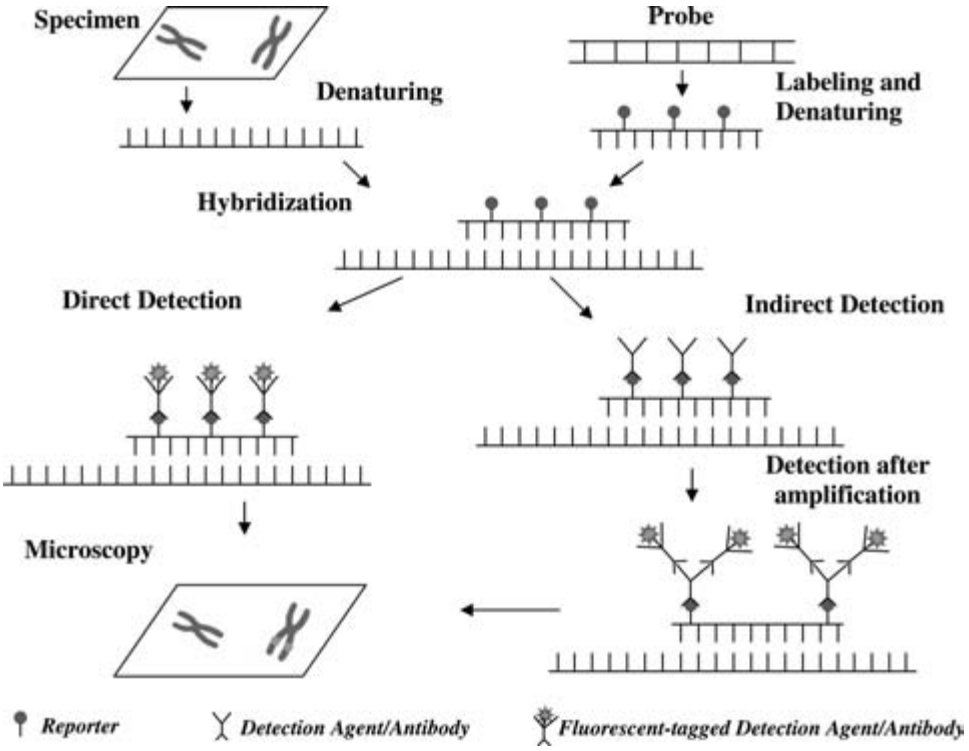


Figure 17.1. Schematic presentation of basic steps in fluorescence *in situ* hybridization. (Also see color plate.)

for FISH (Phillips and Reed 1996, Fujiwara et al. 2001). In most species, however, the only available materials are adult tissues or embryos. Embryos have higher mitotic indices and are better materials than adult tissues.

Colchicine or other microtubule inhibitors are used to arrest cells at metaphase. The duration of colchicine treatment affects chromosome condensation and metaphase quality. Short treatments may not produce enough metaphases, while long treatments will produce highly condensed chromosomes, making FISH and chromosome identification difficult. The optimal treatment duration varies among species or tissue types. It has to be determined through experimentation.

Cells and tissue materials for the detection of gene expression or pathogens are usually fixed in a formaldehyde-based fixative to preserve structure integrity. Fixed cells and tissue are subjected to a pepsin treatment (0.1% in 0.1 N hydrochloric acid [HCl]) to increase their permeability to probes.

Probes

Probe Type and Size

Probes for FISH can be made from any DNA sample or sequences. They can be a gene, a DNA fragment, a repetitive element, a clone, or DNA from a chromosome or

genome. DNA probes can be double- or single-stranded. Single-stranded DNA probes have the advantage of not being able to reanneal, therefore are more sensitive and produce less background; however, double-stranded DNA probes are usually sufficient for most applications. The type and size of DNA probes are important to the success of FISH. Tandemly repeated genes and sequences are excellent targets for FISH. These sequences have a relative short unit that repeats for hundreds or thousands of times. The repeating unit permits efficient labeling, and the entire sequence presents a large and easy target for hybridization and detection. The hybridization of large numbers of labeled probes to the same region makes detection easy with or without amplification. Ribosomal RNA genes are good examples of tandemly repeated genes that are easily detected by FISH. For some tandemly repeated sequences, oligonucleotides as short as 20–50 bp can be used as FISH probes. One example is the vertebrate telomeric sequence, (TTAGGG)n. A 42-mer probe has been successfully used for FISH (Gornung et al. 1998).

For unique sequences, the longer the sequence is, the easier it is to be detected by FISH. Sequences larger than 20 kb can be routinely detected. The detection of target sequences of less than 1 kb, although possible, will require advanced amplification and detection systems (Schriml et al. 1999). Large sequences such as Bacterial Artificial Chromosome (BAC) clones contain both unique and repetitive sequences. Repetitive sequences may include short intersperse nuclear elements (SINE) and long intersperse nuclear elements (LINE). The repetitive sequences, when labeled, produce strong background signals and therefore must be blocked (Lichter and Cremer 1992). Unlabeled total genomic DNA or the Cot-1 fraction is often used as a suppression agent.

The probe size should not be confused with the target DNA size. FISH probes should have a proper size for successful hybridization. Probes that are too short may not be specific or stable enough for hybridization. Probes that are too long may have low penetration and hybridization efficiencies. When long sequences are used as probes, they should be fragmented to the desired length. Probe length should be kept between 100 and 600 bp (Hafen and Levine 1986, Lichter and Cremer 1992, Tautz et al. 1992).

Probe Production and Labeling

FISH probes can be produced and labeled using several methods including nick translation, random priming, polymerase chain reaction (PCR) amplification, end labeling, and direct synthesis. Labeling is achieved by end tagging or incorporating modified nucleotides into probe DNA. The modified nucleotides carry either a fluorescence dye or a reporter molecule that can be later detected. The most commonly used reporters are biotin-11-dUTP and digoxigenin (DIG)-11-dUTP. Nucleotides conjugated with fluorescence dyes are also available for labeling, which makes posthybridization detection unnecessary.

Nick translation is one of the most commonly used methods for probe labeling. Rigby and others (1977) first described it. This method relies on the action of two enzymes: deoxyribonuclease I (DNase I) and *Escherichia coli* DNA polymerase I. At low concentrations, DNase I creates nicks in double-stranded DNA, exposing the 3′ hydroxyl termini and 5′ phosphate termini. The DNA polymerase I functions both as an exonuclease and as a polymerase. It excises nucleotides off one strand of the

DNA and then adds new nucleotides through its polymerase function to fill the gaps. Consequently, labeled nucleotides in the reaction mixture are incorporated into the newly synthesized DNA. Nick translation is most suited for the labeling of large-insert clones such as cosmids (30–40 kb), BACs (120–180 kb), P1s (Phage, 80–250 kb), and yeast artificial chromosomes (YAC, 0.1–2 megabase [Mb]). PCR incorporation is most effective for labeling short DNA fragments (<2 kb). It incorporates modified nucleotides while amplifying the DNA fragment, and generates large quantities of well-labeled probes from small amounts of template. Any DNA fragment that is amplified by PCR can potentially be labeled. The most important factor for PCR-labeling is the normal/labeled nucleotide ratio. Too little labeled nucleotides would result in low labeling efficiency, and too much would reduce PCR yield. The optimal ratio has to be determined empirically. Generally the ratio should be around 2:1 normal to labeled nucleotides. Randomly priming is also an effective method for probe labeling (Feinberg and Vogelstein 1984). It relies on hexanucleotide primers to randomly amplify template DNA and incorporate labeled nucleotides. The advantage of random priming is that it works well with little amount of template DNA (10–200 nanogram [ng]).

Probe size can be controlled during production and labeling. For nick translation, the length is determined by the amount of DNase and reaction time. When random priming is used for probe labeling, probe size can be controlled by adjusting primer concentration; high primer concentrations would generate shorter probes.

Short oligonucleotides can be labeled by end labeling or direct synthesis. End labeling relies on the enzyme terminal transferase to add labeled nucleotides to the 3′ ends of DNA fragments (Schmitz et al. 1990). It is mainly used for labeling short oligonucleotide probes (<100 bp), for which other methods are less effective. More often, short oligonucleotide probes are labeled by direct synthesis. Labeled nucleotides can be placed at desired density and positions during synthesis.

Hybridization and Detection

Target and probe DNA must be denatured before hybridization. Several factors can affect DNA denaturing, including temperature, pH, and formamide concentration. A typical denaturing condition is incubation at 72–75°C in 70% formamide (in 2 × SSC, pH 7.0) for 2–5 minutes. Denatured probes and target sequences must be immediately chilled on ice to prevent reannealing.

Temperature, pH, ionic strength, and duration are important factors for hybridization. Hybridization is usually conducted on slides in a humidified chamber for hours or overnight. After hybridization, nonspecific binding of probes is removed by washing at desired stringencies. Washing stringency is determined by salt concentration with lower concentrations corresponding to higher stringencies. Usually, posthybridization washing is conducted at 2× SSC. If background signals are high, slides are washed at higher stringencies, 0.5× or 1×.

Biotin-labeled probes are detected with fluorescence-labeled avidin. DIG-labeled probes are detected with fluorescence-labeled anti-DIG antibodies. The most commonly used fluorochromes are fluorescein isothiocyanate (FITC), Rhodamine, Texas Red, and CY3. Each fluorochrome has a unique excitation and emission spectrum,

and requires proper filters for visualization (Monk 1992). Fluorochomes with wide differences in excitation/emission spectrum can be used together for dual or triple color FISH.

Depending on the type of DNA sequences and signal strength, amplification of FISH signals may be necessary. Tandem repeats such as ribosomal RNA genes, telomeric repeats, and centromeric repeats produce strong signals without amplification. Large-insert clones also produce strong signals that may not require amplification. Unique sequences shorter than 80 kb usually require amplification. Amplification is achieved using an unlabeled primary antibody (rabbit antisheep) and labeled secondary antibody (antirabbit). Counterstains are used to stain the entire chromosomes for visualization.

FISH signals are documented using digital cameras and preferably cooled 3CCD cameras. Fluorochromes fade rapidly under excitation, and antifade agents should be included as part of the counterstain or mounting medium (Lichter and Cremer 1992). View time should be reduced to a minimum before documentation. Screening for metaphases should be conducted with a filter specific for the counterstain so that the fluorochromes are not unnecessarily exposed.

Methods and Protocols for FISH

FISH protocols are well developed for various applications in human genetics. Working with aquaculture species faces some additional challenges. The main challenge is the lack of resources such as cell lines, probes, and adequate funding for commercial kits. Most of the difficulties can be overcome by using modified protocols and alternative strategies. This section provides some basic protocols that are useful in aquaculture species.

Chromosome Preparation

Many different protocols for chromosome preparation have been reported. They all share the same basic elements: (1) the selection of materials that provide abundant dividing cells; (2) the arrest of metaphases with a colchicine treatment; (3) a hypotonic treatment that conditions chromosomes and the cells for spreading; and (4) thorough fixation with a proper fixative and chromosome spreading using the air-drying method. The following are selected protocols for chromosome preparation that are commonly used in aquaculture species.

Chromosome Preparation from Embryos

There are no cell lines available in marine molluscs and most fish species. Primary cell culture is often a challenge. Embryonic material can be easily obtained in aquaculture species due to external fertilization and provides a good alternative. In molluscs, embryos are small (about 50–100 μm) and contain relatively little yolk materials. Yolk is problematic for FISH analysis. It contains abundant proteins and lipids that affect

detection by creating high background signals. For most molluscs, embryos with 100–2,000 cells give the best results.

Solutions:

1. Carnoy's fixative: 1:3 (v:v) of glacial acetic acid and absolute methanol, freshly made.
2. Phosphate Buffer Saline (PBS): 0.025 M KH_2PO_4, adjust pH to 6.8 with sodium hydroxide (NaOH).
3. Leishman's stain: 1:5 Leishman's stock and PBS. The stock is made by dissolving 150 milligrams (mg) stain in 100 milliliter (ml) methanol.

Procedures:
For oysters, use 6-hour old embryos, cultured at 25°C:

1. Collect embryos in a 15-ml test tube, and treat with colchicine (0.005% in seawater) for 15–20 minutes.
2. Concentrate embryos by centrifugation, remove colchicine, add nine parts of 0.075 moles (M) potassium chloride (KCl) to each part of embryo suspension, and treat for 8 minutes.
3. Concentrate embryos by centrifugation, remove KCl as completely as possible, add 10 ml Carnoy's fixative, change fixative twice within 2 hours, and store at 4°C.
4. Load one–three drops of embryo suspension onto a clean slide, flood with two–three drops of 1:1 acetic acid and methanol, lay the slide at a 45-degree angle, air-dry, and store slide at -20°C.
5. After drying, stain chromosomes with freshly made Leishman's stain for 3–5 minutes, rinse with water, dry it immediately by blowing, and inspect under microscope. For permanent storage, a coverglass can be applied with a drop of Permont mounting medium (Sigma).

Chromosome preparation can also be made from fish embryos. Because fish eggs are usually large and contain too much yolk material, embryonic tissue must be separated from the yolk by dissection. This can be done before or after colchicine treatment. Because fish embryos are large, longer hypotonic treatment, 1–2 hours, is needed.

Chromosome Preparation from Adult Tissues

Embryonic material may not be readily available for some species and sometimes, chromosomes have to be prepared from adult tissues. For some molluscs, growth is seasonal and it is important to use animals that are actively growing. The most commonly used adult tissues are gills, testis, liver, digestive gland, and regenerating fins. The following is a protocol for chromosome preparation using oyster gills or gonad/digestive gland. Solutions are the same as described above.

Procedures:

1. Collect healthy oysters, and condition them for 2–3 days at 23–25°C with intensive feeding and aeration.
2. Transfer oysters to 5 liters (L) seawater with algae, aeration and 0.005% colchicine (0.25 g/5 L, dissolve in 100 ml seawater first), and treat for 5–8 hours.

3. After colchicine treatment, a piece (about 0.25 cm^3) of gill or gonad/digestive gland is removed, treated with 0.075 M KCl for 20–30 minutes, and chopped in two–five drops of Carnoy's fixative. The minced tissue is fixed in 10 ml of fixative. Shake well, change fixative two times, and store at 4°C.
4. Before making slides, suspend the fixed sample and let the large tissue debris settle, take three–five drops of cell suspension from the top and drop onto a clean slide, and overlay with two–three drops of 1:1 acetic acid and methanol. Lean the slide with an angle of about 45 degrees, and let it air dry.
5. After drying, overlay freshly made Leishman's stain on the slide, and stain for 3–7 minutes The staining process can be monitored under 10× objective.
6. After staining, rinse the slide with water, blow dry, and inspect under microscope. Seal with Permount (Sigma) for storage.

Chromosomes can be prepared from fish tissues using similar procedures. When fish are large, injecting colchicine (0.005% of body weight) is more effective. Phytohemagglutinin (PHA) is often used to stimulate mitotic division. Fujiwara and others (2001) described a method for chromosome preparation using lymphocyte culture.

Probe Preparation and Labeling

Among methods for probe labeling, nick translation and PCR incorporation are most popular. Nick translation is suitable for making probes from total genomic DNA or large-insert clones. PCR labeling is mainly used for labeling sequences less than 2 kb. Random priming is used with small amounts of template DNA. Protocols for random priming are not presented here because kits are commercially available, and the procedures are simple. Probes less than 100 bp are best labeled by direct synthesis. Oligonucleotide synthesis service is widely available.

Nick Translation

Several companies such as Roche, Life Technologies, and Sigma supply nick translation kits for DNA labeling. These kits come with all necessary reagents and detailed instructions. A key component of nick translation kits is the optimized enzyme mixture of DNA polymerase I and DNase I that offer highly efficient labeling. When kits are not available, the enzyme ratio and treatments can be easily tested and optimized. The following protocol is based on Lichter and Cremer (1992).

Solutions:

1. 10× Buffer containing 0.5 M Tris-HCl, pH 8.0, 50 mM $MgCl_2$, and 0.5 mg/BSA.
2. DNase I (RNase free) solution: dissolve 3 mg DNase I in 0.5 ml 0.3 M NaCl, add 0.5 glycerol, and store at −20°C. Dilute 1:1000 with double-distilled water just before use.
3. *E. coli* DNA polymerase I: 10 units/μl in double-distilled water.
4. Unlabeled nucleotide set: 2.0 mM each of dATP, dCTP, dGTP, and 0.25 mM dTTP.
5. Labeled nucleotides: 2.5 mM biotin-11-dUTP or DIG-11-dUTP.
6. Stop solution: 50 mM EDTA.

Procedures:

1. Mix the following:

1–2 μg DNA probe	10.0 μl
DNase I (0.003 μg/μl)	1 – 10 μl
DNA polymerase (10 u/μl)	1 – 3 μl
dATP (2 mM)	2.5 μl
dCTP (2 mM)	2.5 μl
dGTP (2 mM)	2.5 μl
dTTP (0.2 mM)	1.6 μl
Biotin-11-dUTP (2 mM)	2.5 μl
Double-distilled water	appropriate
Total	50.0 μl

(17.1)

2. Incubate the mixture in a 15°C water bath for 3 hours.
3. Stop the reaction by adding 50 μl of 50 mM EDTA to a final concentration of 25 mM.
4. Aliquot 7 μl sample for electrophoresis on 2% agarose gel to check fragment size.
5. The probe molecules should be visible as a smear. A smear between 200–600 bp is optimal.
   A. If the DNA is between 200–600 bp, proceed to step 6.
   B. If the probe is too long, add more DNase I or incubate for longer duration until the optimal size is reached.
   C. If the probe is smaller than 100 bp, shorten the incubation time or reduce the amount of DNase I.
6. Clean the probe using protocols below, and store the probe at -20°C till use.

PCR Labeling

PCR incorporation is widely used for labeling probes shorter than 2 kb. Labeling by PCR is simple and robust as long as PCR is optimized. The annealing temperature and Mg^{2+} concentration are important in PCR. The labeled to unlabeled nucleotides ratio is critical. The following protocol is based on Wang and Guo (2004).

Solutions:

1. Primers: dilute the stock primer solution to a 25 μM working solution.
2. Labeled nucleotides: 1 mM stock solution.
3. Unlabeled nucleotides (dNTPs): 10 mM of dATP, dCTP, dGTP, and dTTP each.
4. 10× PCR buffer containing 15 mM $MgCl_2$ that is supplied with the *Taq* polymerase.
5. *Taq* polymerase, 5U/μl.

Procedure:

1. Prepare the following mix on ice, in a 1.5 ml microcentrifuge tube:

1μg DNA probe/ ddH2O	16.8 μl
10X buffer (Mg^{2+})	2.5 μl
BSA (10mg/ml)	1.0 μl

dATP (10 mM)	0.5 μl	
dCTP (10 mM)	0.5 μl	
dGTP (10 mM)	0.5 μl	
dTTP (10 mM)	0.325 μl	
X-dUTP (1mM)	1.75 μl	
Primer (25μM)	1.0 μl	
Taq polymerase (5U/μl)	0.125 μl	
Total	25.0 μl	(17.2)

Note: X-dUTP, X can be DIG, biotin or fluorescein.

2. Mix completely and concentrate by brief centrifugation.
3. Place the tube in a thermocycler and run PCR using the following profile: an initial 5-minute denature at 95°C; 30–35 cycles of 1-minute denature at 95°C, 1 minute annealing at annealing temperature and 1-minute extension at 72°C; and a final extension at 72°C for 5 minutes, and followed by holding at 4°C.
4. Load an aliquot (3 μl) of the reaction onto agarose gel to check the label is successful. A reduction in fragment mobility is an indication of successful labeling.
5. Purify labeled PCR products with G-50 column or ethanol precipitation (see below), and store at −20°C.

Purification of Labeled Probes

Labeled probes should be purified to remove unincorporated labeled nucleotides. The labeled nucleotides, if not removed, would create high background signals. Several kits are available for quick DNA clean up. The G-50 column from Roche (Cat. No. 100609 and 100611) offers fast and effective cleaning of PCR fragments. Free nucleotides can also be removed by ethanol precipitation of DNA using the following protocol.

Solutions:

1. 3 M sodium acetate (or 4 M lithium chloride), filtered and sterilized.
2. 100% ethanol at −20°C.
3. 70% ethanol at −20°C.

Procedure:

1. For every 20 μl of products, add 2 μl of 3 M sodium acetate (or 4 M LiCl) and 60 μl of cold 100% ethanol, mix well.
2. Precipitate the DNA at -20°C overnight, or for 1–2 hours at −80°C.
3. Centrifuge at 12,000 g for 30 minutes at 4°C.
4. Discard the supernatant, wash the pellet by carefully adding 0.5 ml of ice-cold 70% ethanol (v/v), and spin for 5 minutes, 12,000 g at 4°C.
5. Discard the supernatant and leave the pellet dry (be careful so that it does not fall out or blow away).
6. Dissolve the pellet in 10–20 μl of ddH_2O or TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8.0)
7. Store the probe at −20°C until use

FISH Protocol

FISH consists of denaturing, hybridization, posthybridization washing, detection, secondary amplification (optional), and documentation. The following is a protocol for mapping DNA fragments to chromosomes adapted in our laboratory. It is based on several published protocols including Lichter and Cremer (1992) and Wang and others (2005).

Solutions:

1. 20× SSC: dissolve 175.3 g sodium chloride and 88.2 g sodium citrate in 900 ml of water, bring the volume to 1,000 ml, adjust pH to 7.0 with HCl, and store at RT.
2. 50 ml 2× SSC in a Coplin jar: 5 ml 20× SSC and 45 ml distilled water.
3. 40 ml denaturing solution: 4 ml 20× SSC, 8 ml distilled water, 28 ml formamide, adjust pH to 7.0 with HCl, freshly made.
4. 10× PBS: dissolve 80 g sodium chloride, 2 g KCl, 14.4 g Na_2HPO_4, and 2.4 g KH_2PO_4 in 900 ml of water, bring the volume to 1,000 ml, adjust pH to 7.4, filter with 0.2 μm filter, sterilize, and store at RT.
5. Hybridization buffer: 65% formamide in 2× SSC.
6. 1× washing buffer (WB): 1× PBS, 0.4% BSA (w/v), 0.1% Tween-20 (v/v), adjust pH to 7.4 and filter.
7. Ethanol series: prepare fresh 70%, 80%, and 100% ethanol solution in Coplin jars for use at RT, and prepare a second set and place on ice.
8. Counterstain/antifade: mix 100 μl propidium iodide (PI, 500 ng/ml in 4× SSC) with 100 μl antifade (mix 0.2 g 1,4-diazabicyclo[2.2.2]octane in 0.8 ml water, 0.2 ml 1 M Tris-HCl pH 8.0, and 9 ml glycerol, stored in dark at 4°C); or replace PI with 200 ng/ml 4′,6-diamidino-2-phenylindole (DAPI) in 4× SSC.

Procedures:

Slides can be stained with Leishman's stain and screened for suitable metaphases prior to FISH. In that case, slides should be de-stained in 2× SSC before use. Negative controls where FISH probes are replaced with distilled water should be used to test for possible nonspecific hybridization.

1. Prewarm 40 ml of 2× SSC in a Coplin jar to 37°C in a water bath, incubate slides in 2× SSC for 30 minutes.
2. Dehydrate slides in 70%, 80%, and 95% ethanol at RT for 2 minutes each, air dry.
3. Incubate slides in 40 ml of denaturing solution at 72°C ±2°C for 2 minutes.
4. Dehydrate in a cold series of 70%, 80%, and 95% ethanol (Coplin jars on ice), 2 minutes each, and air dry.
5. Denature probe: mix 1 μl of probe with 9–19 μl hybridization buffer in a 0.5 ml microtube, mix well, incubate at 72°C for 5 minutes, and chill on ice.
6. Place 10 μl of probe mix on slide, cover with a coverglass (22 × 22 mm). Use large (15–20 μl) volumes if large areas of the slide (22 × 30 or 22 × 40 mm) are used for hybridization.
7. Seal with rubber cement, incubate at 37°C in a humidified chamber overnight. Care should be taken to avoid air bubbles.
8. Remove cover glass by soaking in 2× SSC at 37°C.
9. Wash slides in 2× SSC at 72°C for 5 minutes, and use a higher wash stringency (0.1–1× SSC) if background signals are too high.

10. Wash in WB twice for 2 minutes each. Do not let slides dry.
11. Let excess WB run off (without letting it dry), apply 10–30 μl (depending on the size of the coverslip) of the appropriate detection reagent (avidin for biotin or antibody for DIG), cover with plastic slip, and incubate in a dark humidified chamber at 37°C for 5–15 minutes.
12. Remove coverslip, wash slides in WB in the dark three times, 2 minutes each.
13. Optional amplification when signals are too weak (usually for short unique sequences):
    A. Apply 60 μl of the appropriate primary antibody (Rabbit antisheep for DIG or antiavidin antibody), cover with plastic slip, incubate at 37°C for 15 minutes, wash slides in WB at RT, in the dark three times, 2 minutes each.
    B. Apply 60 μl of the secondary antibody (labeled antirabbit) or labeled avidin, cover with plastic slip, incubate at 37°C for 15 minutes, and wash slides in WB in the dark three times, 2 minutes each.
14. Apply 15 μl of the appropriate counterstain/antifade (PI for FITC, DAPI for Rhodamine, Texas Red, and dual colors), cover with a glass coverslip, and incubate in the dark for 10 minutes.
15. Blot off excess stain, observe with an epi-fluorescence microscope or store slides at −20°C in the dark.

The FISH protocol works well with a variety of probes made from repetitive and unique sequences. Example metaphases showing FISH signals produced with the protocol are presented in Figure 17.2.

FISH Studies in Aquaculture Species

FISH studies have been reported in many aquaculture species. The studies are concentrated in four areas:

1. chromosomal mapping of tandemly repeated genes such as ribosomal RNA genes
2. the mapping of repetitive elements such as telomeric and centromeric repeats
3. the mapping of large-insert clones
4. the location of gene expression of pathogens in cells and tissues

Studies in fish species have been covered by two excellent reviews (Phillips and Reed 1996, Phillips 2001). This section mainly focuses on studies in molluscan shellfish.

Mapping of Multicopy Genes

The most frequently reported FISH studies in aquaculture species are for chromosomal mapping of tandemly repeated genes. This is because tandemly repeated genes are the best target for FISH detection. The most commonly mapped repetitive gene belongs to the ribosomal RNA family.

Ribosomal RNA genes are one of the best understood gene families in aquaculture species. Ribosomal RNAs encode major (18S-5.8S-28S) and minor (5S) RNAs that together with ribosomal proteins make up the ribosome. Because of their critical

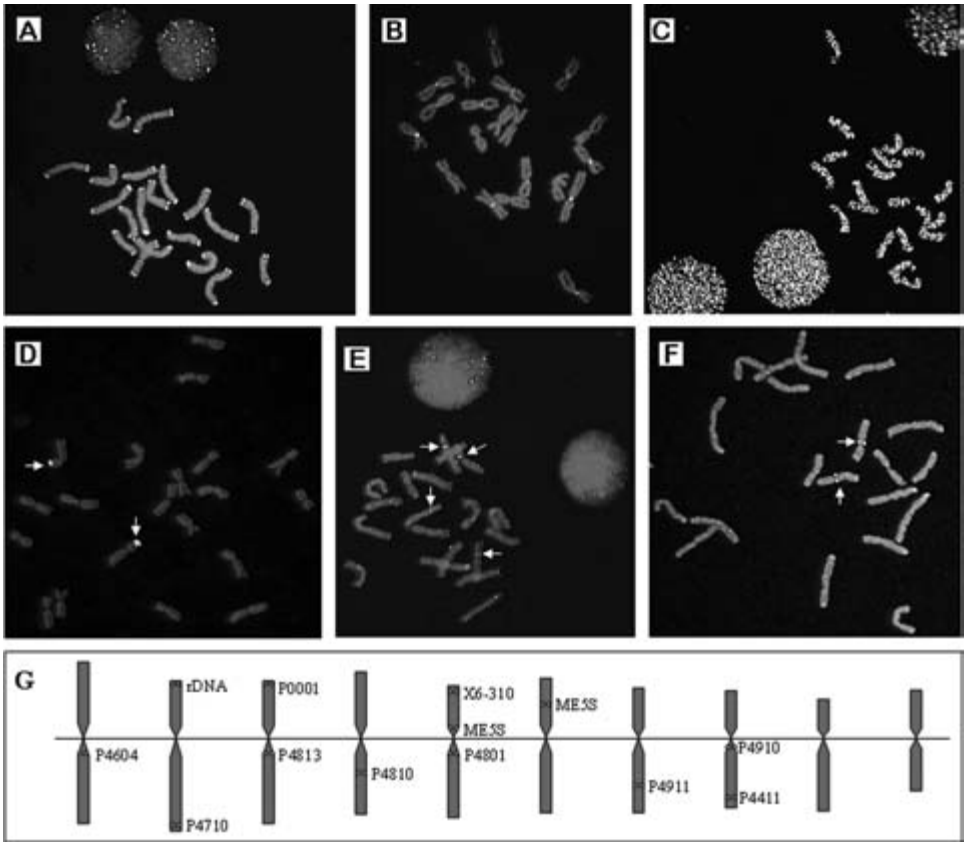


Figure 17.2. Fluorescence *in situ* hybridization with various probes in the eastern oyster. A, the vertebrate telomeric sequence; B, a centromeric element; C, a small interspersed repetitive element; D, the major rRNA genes (18S-5.8S-28S); E, the 5S rRNA gene; F, a P1 clone; and G, a preliminary cytogenetic map. (Also see color plate.)

role in translation, rRNA genes are highly conserved through evolution. They can be easily amplified and sequenced using universal primers. In higher eukaryotes, major and minor rRNA genes are relatively independent of each other and often organized into separate loci on the same or different chromosomes (Martins and Galetti 2001, Liu et al. 2002). Both gene families are organized in large numbers of tandem repeats, making them ideal targets for FISH. Actively transcribed major rRNA genes correspond to NORs and can be visualized by silver staining. Compared with FISH, the detection of rRNA loci by Nucleolar Organizing Region (NOR) staining is not always reliable (Wang et al. 2004).

The rRNA loci have been mapped in many organisms including aquaculture species (Phillips 2001). In molluscs, the rRNA genes have been mapped with FISH in 38 species (Table 17.1). In most studies, rRNA loci were mapped for the purpose of chromosomal identification. The major rRNA genes have one to two loci in most species studies so far. In some studies, comparative mapping of rRNA loci in related species has led to interesting findings about the role of chromosomal changes in evolution. In *Crassostrea* oysters, for example, a difference in the rRNA-bearing chromosome has

Table 17.1. Chromosomal mapping of genes and DNA sequences by FISH in molluscs.

Sequence/Species	FISH Location	Reference
Major rRNA genes, 18S-5.8S-28S		
Pacific oyster *Crassostrea gigas*	10q	Xu et al. 2001, Wang et al. 2004
Suminoe oyster *C. ariakensis*	10q	Wang et al. 2004
Zhe oyster *C. plicatula*	10q	Wang et al. 2004
Kumamoto oyster *C. sikamea*	10q	Wang et al. 2004
Portuguese oyster *C. angulata*	10q	Cross et al. 2003
Eastern oyster *C. virginica*	2p	Zhang et al. 1999, Xu et al. 2001, Wang et al. 2004
Mangrove oyster *C. rhizophorae*	2p	Wang et al. 2004
Queen scallop *Aequipecten opercularis*	7q	Insua et al. 1998
Bay scallop *Argopecten i. irradians*	4p, 8p	Wang and Guo 2004
Zhikong scallop *Chlamys farreri*	5p	Wang and Guo 2004
Scallop *Hinnites distortus*	Centromere of 9 and12	López-Pinón et al. 2005
Scallop *Pecten maximus*	1 pair	Insua et al. 2006
Scallop *Mimachlamys varia*	1 pair, short arm	Insua et al. 2006
Antarctic scallop *Adamussium colbecki*	2p	Odierna et al. 2006
Razor clam *Solen marginatus*	9q, 12p	Fernández-Tajes et al. 2003
Wedgeshell clam *Donax trunculus*	9p	Martínez et al. 2002
Hard clam *Mercenaria mercenaria*	10q, 12p	Wang and Guo 2006
Macoma clam *Macoma nasuta*	15q	González-Tizón et al. 2000
Mahogany clam *Nuttallia nuttallii*	8p	González-Tizón et al. 2000
Clam *Dosinia exoleta*	1 pair, short arm	Hurtado and Pasantes 2005
Cockle *Cerastoderma edule*	1 pair	Insua et al. 1999
Blue mussel *Mytilus edulis*	2 pairs	Insua et al. 2001
Mussel *M. galloprovincialis*	6q, 7q	Martínez-Expósito et al. 1997 Insua and Méndez 1998
Mussel *M. californianus*	3p, 5p, 12p	González-Tizón et al. 2000
Mussel *M. trossulus*	3p, 7p, 7q, 9p	González-Tizón et al. 2000
Mussel *Barchidontes rodriguezi*	2 pairs, short arm	Torrerio et al. 1999
Mussel *Brachidontes pharaonis*	14q	Vitturi et al. 2000b
Cockle *Cerastoderma edule*	1 pair, short arm	Insua et al. 1999
Tulip shell *Fasciolaria lignaria*	4 pairs	Vitturi et al. 2000a
Land snail *Cantareus aspersus*	5 pairs	Vitturi et al. 2005
Land snail *C. mazzullii*	5 pairs	Vitturi et al. 2005
Atlantic dog-whelk *Nucella lapillus*	Variable	Pascoe et al. 1996
Snail *Cerithium vulgatum*	1 pair, short arm	Vitturi et al. 2002
Sea slug *Oxynoe olivacea*	One of pair 12	Vitturi et al. 2000c
Slug *Milax nigricans*	1 pair	Vitturi et al. 2004
Periwinkle *Melarhaphe neritoides*	1 pair	Colomba et al. 2002

Table 17.1. (*Continued*)

Sequence/Species	FISH Location	Reference
Blue abalone *Haliotis fulgens*	4q, 11q	Gallardo-Escarate et al. 2005
Yellow abalone *H. corrugata*	2q, 4q	Gallardo-Escarate et al. 2005
5S rRNA		
Eastern oyster *C. virginica*	5p, 6p	Wang and Guo 2005
Portuguese oyster *C. angulata*	4q, 5q	Cross et al. 2005
Queen scallop *Aequipecten opercularis*	1q, 1q	Insua et al. 1998
Bay scallop *Argopecten i. irradians*	10q	Wang and Guo 2004
Zhikong scallop *Chlamys farreri*	5q	Wang and Guo 2004
Scallop *Hinnites distortus*	13q	López-Pinón et al. 2005
Scallop *Pecten maximus*	1 pair, long arm	Insua et al. 2006
Scallop *Mimachlamys varia*	1 pair, short arm	Insua et al. 2006
Blue mussel *Mytilus edulis*	4 loci on 3 pairs	Insua et al. 2001
Mussel *M. galloprovincialis*	4 loci on 3 pairs	Insua et al. 2001
Cockle *Cerastoderma edule*	9 loci on 5 pairs	Insua et al. 1999
Slug *Milax nigricans*	1 pair	Vitturi et al. 2004
Periwinkle *Melarhaphe neritoides*	1 pair	Colomba et al. 2002
Satellite DNA		
Pacific oyster *C. gigas*	2 pairs	Clabby et al. 1996
	7 pairs, centromere	Wang and Guo 2001
Antarctic scallop *Adamussium colbecki*	Centromere of some	Odierna et al. 2006
Telomeric Repeat, (TTAGGG)n		
Pacific oyster *C. gigas*	telomere	Guo and Allen 1997
Eastern oyster *C. virginica*	telomere	Wang and Guo 2001
Mangrove oyster *C. rhizophorae*	telomere	Wang and Guo 2001
Portuguese oyster *C. angulata*	telomere	Cross et al. 2005
Dwarfsurf clam *Mulinia lateralis*	telomere	Wang and Guo 2001
Hard clam *Mercenaria mercenaria*	telomere	Wang and Guo 2001
Wedgeshell clam *Donax trunculus*	telomere	González-Tizón et al. 1998, Plohl et al. 2002
Clam *Dosinia exoleta*	telomere	Hurtado and Pasartes 2005
Mussel *Mytilus galloprovincialis*	telomere	Plohl et al. 2002
Sea cucumber *Holothuria tubulosa*	telomere	Plohl et al. 2002
Sea slug *Oxynoe olivacea*	No signal	Vitturi et al. 2000c
Slug *Milax nigricans*	telomere	Vitturi et al. 2004
Tulip shell *Fasciolaria lignaria*	telomere	Vitturi et al. 2000a
Land snail *Cantareus aspersus*	telomere	Vitturi et al. 2005
Land snail *C. mazzullii*	telomere	Vitturi et al. 2005
Freshwater snail *Biwamelania habei*	telomere	Nomoto et al. 2001
The "large" *Cerithium vulgatum*	telomere	Vitturi et al. 2002
Blue abalone *Haliotis fulgens*	telomere	Gallardo-Escarate et al. 2005
Yellow abalone *H. corrugata*	telomere	Gallardo-Escarate et al. 2005
Pacific abalone *H. discus*	telomere	Sakai et al. 2005
Periwinkle *Melarhaphe neritoides*	telomere	Colomba et al. 2002
Tetranucleotide Repeat, (GATA)n		
Tulip shell *Fasciolaria lignaria*	Not on Y	Vitturi et al. 2000a

(*Continued*)

Table 17.1. (*Continued*)

Sequence/Species	FISH Location	Reference
Portuguese oyster *C. angulata*	all chromosome	Cross et al. 2005
Land snail *Cantareus aspersus*	all chromosome	Vitturi et al. 2005
Land snail *C. mazzullii*	all chromosome	Vitturi et al. 2005
Slug *Milax nigricans*	all chromosome	Vitturi et al. 2004
Snail *Cerithium vulgatum*	No signal	Vitturi et al. 2002
Blue abalone *Haliotis fulgens*	interstitial and at the ends of some	Gallardo-Escarate et al. 2005
Yellow abalone *H. corrugata*	interstitial and at the ends of some	Gallardo-Escarate et al. 2005
Periwinkle *Melarhaphe neritoides*	No signal	Colomba et al. 2002
Large-insert (P1) Clones		
Eastern oyster *C. virginica*	9 clones to 7 pairs	Wang et al. 2005

revealed a major divide between Pacific and Atlantic species (Wang et al. 2004). The chromosome bearing the major rRNA locus is the smallest in all Pacific species, but the second largest in all Atlantic species. The chromosomal difference coincides with a postzygotic barrier to hybridization, which raises the question if chromosomal changes have played any role during the speciation of oysters. Chromosomal changes have played an important role in the evolution of many taxa (White 1978, King 1993).

FISH analysis of the rRNA locus in scallops (*Pectinidae*) has led to the advance of the genome duplication hypothesis for bivalve evolution (Wang and Guo 2004). The observed number and distribution of the rRNA genes suggest that the karyotype of *Chlamys farreri* with 19 chromosomes is pleisomorphic, and that of *Argopecten irradians* with 16 chromosomes is derived through significant chromosomal loss. The ability of scallops to tolerate significant loss of chromosomes suggests that they may be polyploid in origin. Most scallops and clams have 19 chromosomes, and variants can be explained by chromosomal rearrangement. Chromosome number and DNA content data seem to support the genome duplication hypothesis (Wang and Guo 2004).

The 5S rRNA gene has been mapped in 13 species of molluscs with most species having one to two loci (Table 17.1).

Mapping of Repetitive Elements

Repetitive DNA is a major component of all eukaryotic genomes. Their functions are not well understood at this time. Repetitive sequences may be important for chromosomal structure and crossover. In fish, a large number of repetitive DNA elements have been mapped by FISH (Phillips and Reed 1996, Phillips 2001). In molluscs, only a few repetitive elements have been identified and mapped. They include telomeric and centromeric repeats, and some anonymous repeats.

Telomere is the special protein-DNA structures found at the ends of all eukaryotic chromosomes. It protects chromosomes from degradation and may play important roles in aging and oncogenesis (Blackburn 1990, Harley et al. 1990, Hastie et al. 1990). Telomeric DNA consists of tandem repeats of simple sequences. It is highly conserved through evolution, and all vertebrates share the same telomeric sequence, (TTAGGG)n. This has

been confirmed in many fish species (Phillips and Reed 1996). In molluscs, telomeric sequences have been studied in 21 species (Table 17.1). FISH studies clearly demonstrate the presence of the vertebrate telomeric sequence in all molluscs studied so far, with the exception of the sea slug *oxynoe olivacea* (Vitturi et al. 2000). Molluscs are one of the few invertebrate taxa that share the vertebrate lineage in telomeric sequence (Guo and Allen 1997, Wang and Guo 2001, Plohl et al. 2002). Insects and most protozoa have telomeric sequences that are different from the vertebrate sequence (Zakian 1995). In shrimp, the telomeric sequence is a pentanucleotide repeat, (TAACC)n, that may present at interstitial sites and in less than perfect repeats (Alcivar-Warren et al. 2006). A tetranucleotide repeat, (GATA)n, has been mapped to chromosomal ends and interstitial sites in several molluscs (Vitturi et al. 2005).

In the Pacific oyster, a 166-bp repetitive element has been identified (Clabby et al. 1996). It accounts for 1–4% of the Pacific oyster's genome. The repetitive element is mapped to centromeric regions of seven chromosomes (Wang et al. 2001). A centromeric repeat has also been mapped in the Antarctic scallop *Adamussium colbecki* (Odierna et al. 2006). A highly abundant SINE is found throughout the chromosomes of the eastern oyster (Figure 17.2, C). Several families of repetitive elements have been mapped in a number of fish species (Phillips and Reed 1996, Phillips 2001).

Mapping of Unique Sequences

Chromosomal mapping of genes is one of the most powerful applications of FISH. However, few genes have been successfully mapped in aquaculture species. This is mainly because the mapping of small DNA fragments remains a challenge. Alternatively, genes can be mapped by identifying and mapping the large-insert clones that carry them. DNA fragments larger than 20 kb can be readily mapped by FISH, which covers clones from most large-insert libraries such as P1s, BACs, and YACs. The major histocompatibility (MHC) genes in rainbow trout, for example, are mapped to chromosomes by mapping BACs that carry the genes (Phillips et al. 2003). BAC libraries are now available in several aquaculture species. They can facilitate future mapping of candidate genes by FISH.

The mapping of large-insert clones is also important for the identification of chromosomes and physical mapping. The characterization, identification, and labeling of all individual chromosomes are essential for cytogenetic mapping and analyses. Chromosome identification using traditional banding techniques has been difficult in aquaculture species. Most aquaculture species such as carp, salmonids, and shrimp have a large haploid number. High-quality metaphases needed for banding are difficult to obtain. FISH can map large-insert clones as chromosome-specific probes. By mapping clones that are part of the physical or linkage maps, FISH can anchor the genetic maps to chromosomes (Phillips et al. 2006). It can contribute greatly to the assembly of physical maps (Weier 2001).

P1 clones have been successfully mapped by FISH in the eastern oyster, *Crassostrea virginica* (Wang et al. 2005). P1 clones with an average size of 75 kb were labeled with DIG-11-dUTP using nick translation. Nine of the 21 P1 clones tested produced clear and consistent FISH signals when Cot-1 DNA was used as a blocking agent against repetitive sequences. The nine P1 clones were mapped to 7 of the 10 chromosomes of

the eastern oyster, producing a preliminary cytogenetic map (Figure 17.2). Five of the 9 P1 clones were sequenced at both ends, providing sequence-tagged sites that can be used to integrate linkage and cytogenetic maps.

In humans, chromosome-specific paint probes are available for all chromosomes, and all chromosomes can be simultaneously identified by multicolor FISH (Schröck et al. 1996, Macville et al. 1997). Chromosome-specific paint probes are made from sorted or microdissected chromosomes. Chromosome and chromosome-arm specific paint probes have been produced for salmonids by microdissection (Reed et al. 1995, Phillips 2001).

Localizing Gene Expression and Pathogens

In situ hybridization is widely used to localize gene expression or pathogens in cells and tissues. In these studies, the target is usually large, and detection by immunocytochemistry, which is less expensive than fluorescence-based detection, is usually sufficient. *In situ* hybridization has been used for gene expression studies in aquaculture species (Wong et al. 2004, Takle et al. 2006). Advanced multicolor FISH technology allows simultaneous detection and localization of multiple gene transcripts. Multiplex FISH have been used to detect up to seven different transcripts in a single *Drosophila* embryo (Kosman et al. 2004).

Conventional diagnosis of pathogen infection is often time consuming and unreliable. With a highly specific DNA probe, *in situ* hybridization provides rapid and reliable identification and localization of pathogens. *In situ* hybridization has been used to detect pathogens in aquaculture species (Holzer et al. 2003, Sri Widada et al. 2003, Erickson et al. 2005) as well as microorganisms from aquatic environments such as the bacterial pathogen of the Caribbean coral white plague (Richardson et al. 2005).

Challenges and Future Perspectives

Clearly, FISH is a powerful tool in genomic research and has great potential in aquaculture species. During the past 2 decades, a large number of FISH studies have been conducted in aquaculture species. FISH has contributed to our understanding of the genomes of aquaculture species. Comparative genomic analysis using FISH has led to some interesting findings about genome evolution in fish (Phillips et al. 2003, 2005) and molluscs (Wang et al. 2004, Wang and Guo 2004). However, the full potential of FISH in aquaculture genomics has yet to be realized. Although a large number of FISH studies have been reported, most of the studies are on chromosomal assignment of repetitive genes and sequences. The mapping of large-insert clones is limited to a few species, and the production of chromosome-specific paint probes remains a challenge. No genetic linkage map has been completely anchored to cytogenetic maps. The challenges for using FISH in aquaculture species include the lack of resources and expertise. The development of large-insert libraries and cultured cell lines are needed in many species. Spreading FISH expertise through training and collaboration would be helpful. With the rapid development of new technologies, FISH will become more powerful and easily accessible. Some of the new techniques have already shown great

promise. Now multitarget visualization and quantitative analysis are possible (Levsky and Singer 2003). DNA fiber-FISH and powerful detection systems now permit routine mapping fragments of a few hundred bps at near-kb resolution (Weier 2001, Zwirglmaier 2005). FISH will undoubtedly bring advanced applications to aquaculture genomics in the near future. It will greatly contribute to gene mapping, genomic mapping, and comparative genomic analysis in aquaculture species.

Acknowledgments

Some of the protocols presented here were developed under support from the U.S. Department of Agriculture (96-35205-3854), the NOAA Sea Grant (B/T-9801; NJMSC-6520-0011 & 12), and the New Jersey Commission on Science and Technology (02-2042-007-11).

References

Alcivar-Warren A, D Meehan-Meola, Y Wang, X Guo, L Zhou, J Xiang, S Moss, S Arce, W Warren, Z Xu, and K Bell. 2006. Isolation and mapping of telomeric pentanucleotide (TAACC)n repeats of the pacific whiteleg shrimp, *Penaeus vannamei,* using fluorescence *in situ* hybridization. Mar Biotechnol, 8, pp. 467–480.

Blackburn EH. 1990. Telomeres and their synthesis. Science, 249, pp. 489–490.

Clabby C, U Goswami, F Flavin, NP Wilkins, JA Houghton, and R Powell. 1996. Cloning, characterization and chromosomal location of a satellite DNA from the Pacific oyster, *Crassostrea gigas.* Gene, 168, pp. 205–209.

Colomba MS, R Vitturi, L Castriota, R Bertoni, and A Libertini. 2002. FISH mapping of 18S–28S and 5S ribosomal DNA, $(GATA)_n$ and $(TTAGGG)_n$ telomeric repeats in the periwinkle *Melarhaphe neritoides* (Prosobranchia, Gastropoda, Caenogastropoda). Heredity, 88, pp. 381–384.

Cross I, E Diaz, I Sanchez, and L Rebordinos. 2005. Molecular and cytogenetic characterization of *Crassostrea angulata* chromosomes. Aquaculture, 247, pp. 135–144.

Cross I, L Vega, and L Rebordinos. 2003. Nucleolar organizing regions in *Crassostrea angulata:* chromosomal location and polymorphism. Genetica, 119, pp. 65–74.

De Klein A, AG van Kessel, G Grosveld, CR Bartram, A Hagemeijer, D Bootsma, NK Spurr, N Heisterkamp, J Groffen, and JR Stephenson. 1982. A cellular oncogene is translocated to the Philadelphia chromosome in chronic myelocytic leukemia. Nature, 300, pp. 765–767.

Erickson HS, BT Poulos, KF Tang, D Bradley-Dunlop, and DV Lightner. 2005. Taura syndrome virus from Belize represents a unique variant. Dis Aquat Organisms, 64, pp. 91–98.

Feinberg P and B Vogelstein. 1984. A technique for radiolabeling DNA restriction enzyme fragments to high specific activity. Anal Biochem, 137.

Fernández-Tajes J, A González-Tizón, A Martínez-Lage, and J Méndez. 2003. Cytogenetics of the razor clam *Solen marginatus* (Mollusca : Bivalvia : Solenidae). Cytogenet Genome Res, 101, pp. 43–46.

Fujiwara A, C Nishida-Umehara, T Sakamoto, N Okamoto, I Nakayama, and S Abe. 2001. Improved fish lymphocyte culture for chromosome preparation. Genetica, 111, pp. 77–89.

Gallardo-Escarate C, J Alvarez-Borrego, M del Rio-Portilla, E von Brand-Skopnik, I Cross, A Merlo, and L Rebordinos. 2005. Karyotype analysis and chromosomal localization by fish of ribosomal DNA, telomeric (TTAGGG)(N) and (GATA)(N) repeats in *Haliotis fulgens* and *H. corrugata* (Archeogastropoda : Haliotidae). J Shellfish Res, 24, pp. 1153–1159.

Gonzàlez-Tizón A, A Martínez-Lage, L Mariñas, R Feire, L Cornudella, and J Méndez. 1998. Cytogenetic characterization of *Donax trunculus* (Mollusca, Bivalvia). Proc 13th international chromosome conference: Abstract, p. 109.

Gonzàlez-Tizón AM, A Martínez-Lage, I Rego, J Ausió, and J Méndez. 2000. DNA content, karyotypes, and chromosomal location of 18S–5.8S–28S ribosomal loci in some species of bivalve molluscs from the Pacific Canadian coast. Genome, 43, pp. 1065–1072.

Gornung E, I Gabrielli, and L Sola. 1998. Localization of the (TTAGGG)n telomeric sequence in zebrafish chromosomes. Genome, 41, pp. 136–138.

Guo Jr X and SK Allen. 1997. Fluorescence *in situ* hybridization of the vertebrate telomere sequence to chromosome ends of the Pacific oyster, *Crassostrea gigas* Thunberg. J Shellfish Res, 16, pp. 87–89.

Hafen E and M Levine. 1986. Drosophila, A practical approach. In: Roberts DB, Ed. IRL Press, Oxford, pp. 139–174.

Harley CB, AB Futcher, and CW Greider. 1990. Telomeres shorten during ageing of human fibroblasts. Nature (London), 345, pp. 458–460.

Hastie ND, M Dempster, MG Dunlop, and AM Thompson. 1990. Telomere reduction in human colorectal carcinoma and with ageing. Nature (London), 346, pp. 866–868.

Holzer AS, C Sommerville, and R Wootten. 2003. Tracing the route of Sphaerospora truttae from the entry locus to the target organ of the host, *Salmo salar* L., using an optimized and specific *in situ* hybridization technique. J Fish Diseases, 26, pp. 647–655.

Hurtado NS and JJ Pasantes. 2005. Surface spreading of synaptonemal complexes in the clam *Dosinia exoleta* (Mollusca, Bivalvia). Chromosome Res, 13, pp. 575–580.

Insua A, R Freire, and J Méndez. 1999. The 5S rDNA of the bivalve *Cerastoderma edule:* nucleotide sequence of the repeat unit and chromosomal location relative to 18S–28S rDNA. Genet Sel Evol (Paris), 31, pp. 509–518.

Insua A, R Freire, J Rios, and J Méndez. 2001. The 5S rDNA of mussels *Mytilus galloprovincialis* and *M. edulis:* sequence variation and chromosomal location. Chromosome Res, 9, pp. 495–505.

Insua A, MJ López-Pinón, R Freire, and J Méndez. 2006. Karyotype and chromosomal location of 18S–28S and 5S ribosomal DNA in the scallops *Pecten maximus* and *Mimachlamys varia* (Bivalvia: Pectinidae). Genetica, 126, pp. 291–301.

Insua A, MJ Lopez-Pinon, and J Méndez. 1998. Characterization of *Aequipecten opercularis* (Bivalvia: Pectinidae) chromosomes by different staining techniques and fluorescent *in situ* hybridization. Genes Genet Syst, 73, pp. 193–200.

Insua A and J Méndez. 1998. Physical mapping and activity of ribosomal RNA genes in mussel *Mytilus galloprovincialis.* Hereditas, 128, pp. 189–194.

John H, M Birntiel, and K Jones. 1969. RNA;DNA hybrids at the cytogenetical level. Nature, 223, pp. 582–587.

King M. 1993. Species Evolution: The Role of Chromosome Change. Cambridge University Press, Cambridge.

Kosman D, CM Mizutani, D Lemons, WG Cox, W McGinnis, and E Bier. 2004. Multiplex detection of RNA expression in Drosophila embryos. Science, 305, p. 846.

Langer PR, AA Waldrop, and DC Ward. 1981. Enzymatic synthesis of biotin-labeled polynucleotides: novel nucleic acid affinity probes. Proc Natl Acad Sci USA, 78, pp. 6633–6637.

Leary JL, DJ Brigati, and DC Ward. 1983. Rapid and sensitive colorimetric method for visualizing biotin-labeled DNA probes hybridized to DNA or RNA immobilized on nitrocellulose: Bio-blots. Proc Natl Acad Sci USA, 80, pp. 4045–4049.

Levsky JM and RH Singer. 2003. Fluorescence *in situ* hybridization: past, present and future. J Cell Sci, 116, pp. 2833–2838.

Lichter P and T Cremer. 1992. Human Cytogenetics, A Practical Approach. In: Rooney DE, Czepulkowski BH, Eds. Chromosome analysis by non-isotopic *in situ* hybridization. IRL Press, New York, pp. 157–192.

Liu Z, D Zhang, D Hong, and X Wang. 2002. Chromosomal localization of 5S and 18S–5.8S–25S ribosomal DNA sites in five Asian pines using fluorescence *in situ* hybridization. Theor Appl Genet, 122, pp. 1007–1024.

López-Pinón MJ, A Insua, and J Méndez. 2005. Chromosome analysis and mapping of ribosomal genes by one- and two-color fluorescent in situ hybridization in *Hinnites distortus* (Bivalvia : Pectinidae). J Hered, 96, pp. 52–58.

Macville M, T Veldman, H Padilla-Nash, D Wangsa, P O'Brien, E Schrock, and T Ried. 1997. Spectral karyotyping, a 24-colour FISH technique for the identification of chromosomal rearrangements. Histochem Cell Biol, 108, pp. 299–305.

Martínez A, L Mariñas, A González-Tizón, and J Méndez. 2002. Cytogenetic characterization of Donax trunculus (Bivalvia : Donacidae) by means of karyotyping, fluorochrome banding and fluorescent *in situ* hybridization. J Mollus Stud, 68, pp. 393–396.

Martínez-Expósito MJ, J Méndez, and JJ Pasantes. 1997. Analysis of NORS and NOR-associated heterochromatin in the mussel *Mytilus galloprovincialis* Lmk. Chromosome Res, 5, pp. 268–273.

Martins C and PMJ Galetti. 2001. Two 5S rDNA arrays in neotropical fish species: Is it a general rule for fishes? Genetica, 111, pp. 439–446.

Monk AJ. 1992. Microscopy, photography, and computerized image analysis. In: Rooney DE, Czepulkowski BH, Eds. Human Cytogenetics: A Practical Approach. Oxford University Press, Oxford, pp. 223–249.

Nomoto Y, M Hirai, and R Ueshima. 2001. Cloning of molluscan telomere DNA with (TTAGGG)n repeat and its chromosomal location in the freshwater snail *Biwamelania habei.* Zoolog Sci, 18, pp. 417–422.

Odierna G, G Aprea, M Barucca, A Canapa, T Capriglione, and E Olmo. 2006. Karyology of the Antarctic scallop *Adamussium colbecki,* with some comments on the karyological evolution of pectinids. Genetica, 127, pp. 341–349.

Pardue ML and JG Gall. 1969. Molecular hybridization of radioactive DNA to the DNA of cytological preparations. Proc Natl Acad Sci USA, 64, pp. 600–604.

Pascoe PL, SJ Patton, R Critcher, and DR Dixon. 1996. Robertsonian polymorphism in the marine gastropod, *Nucella lapillus:* Advances in karyology using rDNA loci and NORs. Chromosoma, 104, pp. 455–460.

Phillips RB. 2001. Application of fluorescence *in situ* hybridization (FISH) to fish genetics and genome mapping. Mar Biotechnol, 3, pp. 145–152.

Phillips RB, A Amores, MR Morasch, C Wilson, and JH Postlethwait. 2006. Assignment of zebrafish genetic linkage groups to chromosomes. Cytogenet Genome Res, 114, pp. 155–162.

Phillips RB and KM Reed. 1996. Application of fluorescence *in situ* hybridization (FISH) techniques to fish genetics: a review. Aquaculture, 140, pp. 197–216.

Phillips RB, A Zimmerman, MA Noakes, Y Palti, MW Morasch, L Eiben, SS Ristow, GH Thorgaard, and JD Hansen. 2003. Physical and genetic mapping of the rainbow trout major histocompatibility regions: evidence for duplication of the class I region. Immunogenetics, 55, pp. 561–569.

Plohl M, E Prats, A Martínez-Lage, A González-Tizén, J Méndez, and L Cornudella. 2002. Telomeric localization of the vertebrate-type hexamer repeat, (TTAGGG)n, in the wedgeshell clam *Donax trunculus* and other marine invertebrate genomes. J Biol Chem, 277, pp. 19839–19846.

Reed KM, SK Bohlander, and RB Phillips. 1995. Microdissection of the Y chromosome and FISH analysis of the sex chromosomes of lake trout, *Salvelinus namaycush.* Chromosome Res, 3, pp. 221–226.

Richardson LL, DK Mills, ER Remily, and JD Voss. 2005. Development and field application of a molecular probe for the primary pathogen of the coral disease white plague type II. Revista de Biologia Tropical, 53, pp. 1–10 Suppl.

Rigby PWJ, M Dieckmann, C Rhodes, and P Berg. 1977. Labeling deoxyribonucleic acid to high specific activity *in vitro* by nick translation with DNA polymerase I. J Mol Evol, 113, pp. 237–241.

Sakai M, SI Okumura, and K Yamamori. 2005. Telomere analysis of Pacific abalone *Haliotis discus* hannai chromosomes by fluorescence in situ hybridization. J Shellfish Res, 24, pp. 1149–1151.

Schmitz GG, T Walter, R Seibl, and C Kessler. 1990. Nonradioactive labeling of oligonucleotides in vitro with the hapten digoxigenin by tailing with terminal transferase. Anal Biochem, 192, pp. 222–231.

Schriml LM, HM Padilla-Nash, A Coleman, P Moen, WG Nash, J Menninger, G Jones, T Ried, and M Dean. 1999. Tyramide signal amplification(TSA)-FISH applied to mapping PCR labeled probes less than 1 kb in size. Biotechniques, 27, pp. 608–613.

Schröck E, SD Manoir, T Veldman, B Schoell, J Wienberg, MA Ferguson-Smith, Y Ning, DH Ledbetter, I Bar-Am, D Soenksen, Y Garini, and T Ried. 1996. Multicolor spectral karyotyping of human chromosomes. Science, 273, pp. 494–497.

Singer RH and DC Ward. 1982. Actin gene expression visualized in chicken muscle tissue culture by using *in situ* hybridization with a biotinated nucleotide analog. Proc Natl Acad Sci USA, 79, pp. 7331–7335.

Sri Widada J, S Durand, I Cambournac, D Qian, Z Shi, E Dejonghe, V Richard, and JR Bonami. 2003. Genome-based detection methods of Macrobrachium rosenbergii nodavirus, a pathogen of the giant freshwater prawn, Macrobrachium rosenbergii dot-blot, *in situ* hybridization and RT-PCR. J Fish Diseases, 26, pp. 583–590.

Takle H, A McLeod, and O Andersen. 2006. Cloning and characterization of the executioner caspases 3, 6, 7 and Hsp70 in hyperthermic Atlantic salmon (*Salmo salar*) embryos. Comp Biochem & Physiol Part B, 144, pp. 188–198.

Tautz D, M Hulskamp, and RJ Sommer. 1992. In situ hybridization—A practical approach. In: Wilkinson DG, Ed. IRL Press, Oxford, pp. 63–64.

Torreiro A, MJ Martínez-Expósito, MI Trucco, and JJ Pasantes. 1999. Cytogenetics in *Brachidontes rodriguezi* d'Orb (Bivalvia, Mytilidae). Chromosome Res, 7, pp. 49–55.

Vitturi R, M Colomba, L Castriota, AM Beltrano, A Lannino, and N Volpe. 2002. Chromosome analysis using different staining techniques and fluorescent in situ hybridization in *Cerithium vulgatum* (Gastropoda : Cerithiidae). Hereditas, 137, pp. 101–106.

Vitturi R, MS Colomba, P Gianguzza, and AM Pirrone. 2000a. Chromosomal location of ribosomal DNA (rDNA), (GATA)(n) and (TTAGGG)(n) telomeric repeats in the neogastropod *Fasciolaria lignaria* (Mollusca : Prosobranchia). Genetica, 108, pp. 253–257.

Vitturi R, P Gianguzza, MS Colomba, KR Jensen, and S Riggio. 2000b. Cytogenetic characterization of *Brachidontes pharaonis* (Fisher P, 1870): karyotype, banding and fluorescence *in situ* hybridization(FISH) (Mollusca: Bivalvia: Mytilidae). Ophelia, 52, pp. 213–220.

Vitturi R, P Gianguzza, MS Colomba, KR Jensen, and S Riggio. 2000c. Cytogenetics in the sacoglossan *Oxynoe olivacea* (Mollusca: Opisthobranchia): karyotype, chromosome banding and fluorescent *in situ* hybridization. Mar Biol, 137, pp. 577–582.

Vitturi R, A Libertini, L Sineo, I Sparacio, A Lannino, A Gregorini, and M Colomba. 2005. Cytogenetics of the land snails *Cantareus aspersus* and *C. mazzullii* (Mollusca : Gastropoda : Pulmonata). Micron, 36, pp. 351–357.

Vitturi R, L Sineo, N Volpe, A Lannino, and M Colomba. 2004. Repetitive DNAs in the slug *Milax nigricans:* association of ribosomal (18S–28S and 5S rDNA) and (TTAGGG)(n) telomeric sequences in the slug *Milax nigricans* (Mollusca : Gastropoda : Pulmonata). Micron, 35, pp. 255–260.

Wang Y and X Guo. 2001. Chromosomal mapping of the vertebrate telomeric sequence (TTAGGG)n in four bivalve molluscs by fluorescence *in situ* hybridization. J Shellfish Res, 20, pp. 1187–1190.

Wang Y and X Guo. 2004. Chromosomal Rearrangement in Pectinidae Revealed by rRNA Loci and Implications in Bivalve Evolution. Biol Bull, 207, pp. 247–256.

Wang Y and X Guo. 2007. Chromosomal mapping of major ribosomal rRNA genes in the hard clam (*Mercenaria mercenaria*) using fluorescence *in situ* hybridization. Mar Biol, In press.

Wang Y, Z Xu, and X Guo. 2001. A centromeric satellite sequence in the Pacific oyster (*Crassostrea gigas* Thunberg) identified by fluorescence *in situ* hybridization. Mar Biotechnol, 3, pp. 486–492.

Wang Y, Z Xu, and X Guo. 2004. Differences in the rDNA-bearing chromosome divide the Asian-Pacific and Atlantic species of *Crassostrea* (Bivalvia, Mollusca). Biol Bull, 206, pp. 46–54.

Wang Y, Z Xu, and X Guo. 2005. Chromosomal mapping of 5s ribosomal RNA genes in the eastern oyster, *Crassostrea virginica* Gmelin by fluorescence *in situ* hybridization. J Shellfish Res, 24, pp. 959–964.

Wang Y, Z Xu, JC Pierce, and X Guo. 2005. Characterization of Eastern Oyster (*Crassostrea virginica* Gmelin) Chromosomes by Fluorescence *in situ* Hybridization with Bacteriophage P1 Clones. Mar Biotechnol, 7, pp. 207–214.

Weier HG. 2001. DNA fiber mapping techniques for the assembly of high-resolution physical maps. J Histochem Cytochem, 49, pp. 939–948.

White MJD. 1978. Modes of Speciation. WH Freeman, San Francisco.

Wong TT, Y Gothilf, N Zmora, KE Kight, I Meiri, A Elizur, and Y Zohar. 2004. Developmental expression of three forms of gonadotropin-releasing hormone and ontogeny of the hypothalamic-pituitary-gonadal axis in gilthead seabream (*Sparus aurata*). Biol Reprod, 71, pp. 1026–1035.

Xu Z, X Guo, PM Gaffney, and JC Pierce. 2001. Chromosomal location of the major ribosomal RNA genes in *Crassostrea virginica* and *Crassostrea gigas.* Veliger, 44, pp. 79–83.

Zakian VA. 1995. Telomeres: beginning to understand the end. Science, 270, pp. 1601–1607.

Zhang QY, G Yu, RK Cooper, and TR Tiersch. 1999. Chromosomal location by fluorescence *in situ* hybridization of the 28S ribosomal RNA gene of the eastern oyster. J Shellfish Res, 18, pp. 431–435.

Zwirglmaier R. 2005. Fluorescence *in situ* hybridisation (FISH)—the next generation. FEMS Microbiol Lett, 246, pp. 151–158.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 18
Radiation Hybrid Mapping in Aquatic Species

Caird E. Rexroad III

Introduction

Understanding the mechanisms underlying the inheritance of traits has always been of great interest to humanity. The first documented studies in inheritance were conducted in the mid-1800s by the Austrian monk Gregor Von Mendel (Mendel 1951). Mendel's observations of the inheritance of traits in peas caused him to form a set of laws describing mechanisms of inheritance. These laws included the Laws of Segregation and Independent Assortment stating that for each trait, an organism receives one characteristic from each parent, and traits are inherited independently of other traits. Although originally presented in 1865, the significance of Mendel's research was not fully realized until the early 1900s. In 1909, Thomas Hunt Morgan observed that inheritance of the *white-eye* mutation in the fruit fly *Drosophila* was correlated to the inheritance of the male sex chromosome. Morgan and colleagues continued to examine the inheritance of additional mutations in *Drosophila* and eventually published their Chromosome Theory of Inheritance stating that the inheritance of traits can be explained by a linear order of genes on chromosomes (Morgan et al. 1915). In addition to sex-linked mutations, they observed that in this species having four chromosomes there were four groups of mutations (or traits) that did not sort independently throughout multiple generations. With this information, they were able to develop linear maps of chromosomes establishing the order of markers representing mutations based on the biological phenomenon of homologous recombination.

Somatic Cell Hybrid Genetics

Construction of genetic maps for species such as *Drosophila* succeeded because of the number of well-characterized traits (distinction of mutation versus wild type) and the ability to observe the inheritance of those mutations in large numbers of progeny produced from controlled mating in a laboratory. The development of genetic maps for the human genome obviously could not use this strategy; therefore, progress was hindered for many years by the lack of mapping populations where the inheritance of distinct traits could be observed. Consequently, early efforts to construct maps of the human genome were not based on observations of inheritance; rather they were maps based on the physical properties of chromosomes. One such strategy was the use of Somatic Cell Hybrid (SCH) genetics (Green 1969). The major distinctions between the initial genetic maps of *Drosophila* and the initial physical maps of humans constructed using the SCH approach follow:

1. Biochemical molecules—protein or DNA molecules that can be uniquely observed—replaced mutations as markers.
2. The inheritance of these markers was not observed in families, but instead their retention was observed in a panel of hybrid cell lines.
3. Mapping resolution was limited to grouping of markers that are on the same chromosome, therefore no linear marker orders or genetic distances between markers are produced.

The first step in SCH mapping is creating a panel of hybrid cell lines produced by cell fusion of donor cells from a species of interest with host recipient cells to form a panel of hybrid cell lines (see Figure 18.1). Each hybrid cell line retains a unique subset of chromosomes from the donor cell line, including a chromosome containing a biochemical selectable marker that ensures that only hybrid cell lines are produced from the fusion (Littlefield 1964).

The next step is to develop biochemical markers that can be distinguished between the donor and recipient species. Initially, the availability of markers was a limiting factor in constructing chromosome maps. Over the first decade, the number of genes mapped to chromosomes increased from 2 to more than 300 (Walter and Goodfellow 1993). Researchers did not have access to the abundance of messenger RNA (mRNA) and DNA sequences that exists today, including more than 100 GB pairs available through public databases such as GenBank (www.ncbi.nlm.nih.gov). The most important criterion for marker development is that markers need to be scoreable in hybrid cell lines. If a homologous marker from the recipient species is present, it must somehow be

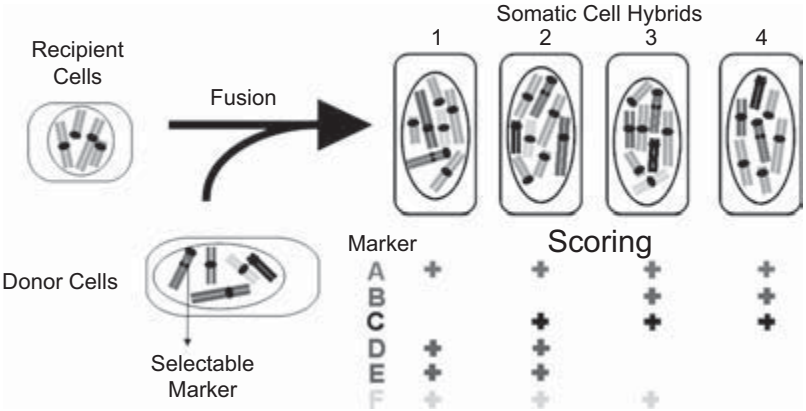


Figure 18.1. A schematic representing a fusion event between a donor cell line and a recipient cell line in the formation of a Somatic Cell Hybrid panel. The hybrid cell lines retain the recipient cell genome, the donor cell chromosome containing the biochemical selectable marker, and other donor cell chromosomes retained randomly. Each chromosome has a unique retention pattern in the panel. For instance, markers observed in all of the hybrid cell lines would be on the green chromosome, which is always retained because it contains the selectable marker. Markers D and E are both retained in cell lines 1 and 2, therefore they are syntenic, or on the same chromosome. Increasing the number of cell lines in a panel and the number of markers observed increases the statistical confidence in the map. (Also see color plate.)

distinguishable from the donor marker. Markers that are not distinguishable are not informative and cannot be included in mapping analyses. To distinguish between host and recipient markers, researchers exploit differences in sequence and/or marker length. This requires the comparison of donor and recipient sequences, which historically required quite an effort in itself. The 3′ untranslated (UTR) or introns of genes are typically examined when creating markers for genes. If base pair (bp) differences are observed, then assays similar to restriction fragment length polymorphisms (Botstein et al. 1980) can be developed that permit scoring the panel. This requires that the sequence variation occur in the cleavage site of a restriction enzyme, therefore not all variation can be used for marker development. After restriction digest, the homologous markers can be distinguished based on length using a Southern blot technique (Southern 1975). More modern approaches use the polymerase chain reaction (PCR) to amplify the marker to a quantity that would permit visualization using agarose gel electrophoresis (Abbott 1992, Ephrussi and Weiss 1969, Mullis and Faloona 1987). Even more recently, today's technologies would employ capillary electrophoresis or mass spectrometry to observe markers developed based on single nucleotide polymorphisms (Sobrino et al. 2005).

Included in the types of markers scored on SCH panels are Type I and Type II loci as described by O'Brien and colleagues (1993). Type I loci represent genes that are coded for in DNA and transcribed into RNA, and are most often translated into proteins, that is, genes. These markers are often conserved between species, having similar DNA/protein sequences and physiological functions. Interest in mapping Type I loci stems from the use of chromosome maps to identify genes affecting traits of interest (Collins 1995, Lander and Botstein 1989). Constructing chromosome maps representing the linear order of genes can be used to identify regions of conserved synteny, retention of genes on chromosome segments across species, and observe evolution at the chromosomal level (Wakefield and Graves 1996). Taken one step further, candidate genes, hypothesized to affect a trait of interest, can be selected based on comparative map information from a species with a better developed map for evaluation in a species of interest (Womack and Kata 1995). Type II loci are highly variable and not necessarily associated with genes. The value of Type II loci is their ability to be placed on both physical and genetic maps, therefore providing a mechanism for integrating their information. This class of markers includes microsatellites (Weber and May 1989), which have been used extensively in the construction of chromosome maps and other genetic analyses because of their high degree of polymorphism and codominant inheritance. See Chapter 5.

The concept of assigning genetic markers to chromosomes through SCH mapping is based on the scoring (positive or negative) of each marker in each hybrid cell line. Markers that are retained in the same hybrid cell lines are likely located on the same chromosome (see Table 18.1).

Radiation Hybrids

Using the genetic map, Morgan could observe and predict the inheritance of mutations in *Drosophila* based on the linear orders of mutations and the genetic distances between them. Although SCH are excellent resources for mapping loci to

Table 18.1. A simplified example of a concordance table containing data from 8 markers (A–H) scored on a panel of 9 somatic cell hybrids. Based on their co-retention in the panel, the marker pairs A & B, C & D, E & F, and G & H are syntenic, meaning they map to the same chromosome. A & B have been scored positive for every SCH cell line; therefore, they must be located on the chromosome containing the selectable marker.

	Somatic Cell Hybrid Lines								
Marker	1	2	3	4	5	6	7	8	9
A	+	+	+	+	+	+	+	+	+
B	+	+	+	+	+	+	+	+	+
C		+	+		+	+			
D		+	+		+	+			
E	+		+		+		+		
F	+		+		+		+		
G				+		+			+
H				+		+			+

chromosomes, their limitation is that their level of resolution is the chromosome, therefore linear maps of chromosomes cannot be constructed. In 1975, Goss and Harris reported the use of "irradiation and fusion gene transfer" to map four genes on the long arm of the human X chromosome (Goss and Harris 1975). This strategy built upon SCH genetics, with the additional step of irradiating the donor cell line prior to fusion. Donor chromosome fragments, including those not containing a selectable marker, were retained in the recipient cell line because they incorporated into the host genome as insertions or translocations or as partial chromosomes that retained their replication machinery. The development of linear maps was based on the assumption that irradiation would cause breaks in chromosomes with equal probability between any two markers. At this point, mapping resolution was increased to localization of large chromosomal segments.

In 1990, Cox and colleagues resurrected irradiation and fusion gene transfer technology by using a somatic cell hybrid containing only human chromosome 21 as a donor cell line (Cox et al. 1990). Cox and colleagues were able to establish a linear order of 14 DNA markers by observing their coretention in 103 hybrid cell lines. Once again, resolution was increased. Along with this new technology, new sophisticated statistical analyses were developed to construct maps from RH datasets (Barrett 1992, Boehnke et al. 1991). Similar to SCH mapping, RH mapping strategies are based on the concept that markers that are close together on chromosomes will frequently be coretained in the same hybrids; the probability that irradiation will induce a chromosome break between two markers decreases as the physical distance between the two markers decreases. To provide adequate statistical support for mapping, marker retention frequencies, which is the percentage of times a marker is scored positive in an RH panel, is critical. Optimal retention is 20–50% (Walter and Goodfellow 1993). RH mapping is calculated based on the coretention of markers in fragments across the panel. The estimated frequency of breakage between two markers is θ, which ranges from 0 to 1 and is analogous to recombination frequencies (r) used in genetic mapping. A θ value of 0 means two markers are always coretained, a value of 1 means they are coretained at random. This raw value is then included in multipoint analyses and

transformed into centiRays (cR)–the RH map unit–using map functions similar to those of Haldane or Kosambi, which are used in genetic map construction. Hence, observation of chromosome breaks between two markers in RH mapping is analogous to observing recombination between two markers in genetic mapping. In fact, the term "linkage" is often used in RH mapping. The frequency of chromosome breaks between two markers is not only due to their physical distance, but also to the intensity of the radiation used to create the panel. Siden and colleagues (1992) conducted experiments to observe the effects of different dosages of radiation on a segment of the human X chromosome. At 5,000 rad 10% of the clones retained the entire chromosome arm, 40% had fragments of 3–30 megabase (Mb), and 50% had fragments less than 3 Mb. At 25,000 rad only 6% had fragments larger than 3 Mb. Therefore, the radiation hybrid map-distance unit is annotated with a subscript stating the dosage used to create the panel in rads (i.e., cR_{3000}). Retention of multiple fragments from a single chromosome in a hybrid cell line complicates analyses; therefore, 100–300 cell lines must be scored for a panel to construct statistically significant maps.

Using Cox's strategy, constructing a map of every human chromosome would require prescreening of an SCH panel to identify the chromosome containing the markers followed by subsequent screening of the appropriate chromosome specific radiation hybrid panel. To simplify this procedure, Walter and colleagues (1994) reported the development of whole genome radiation hybrid panels WG-RH (see Figure 18.2). The

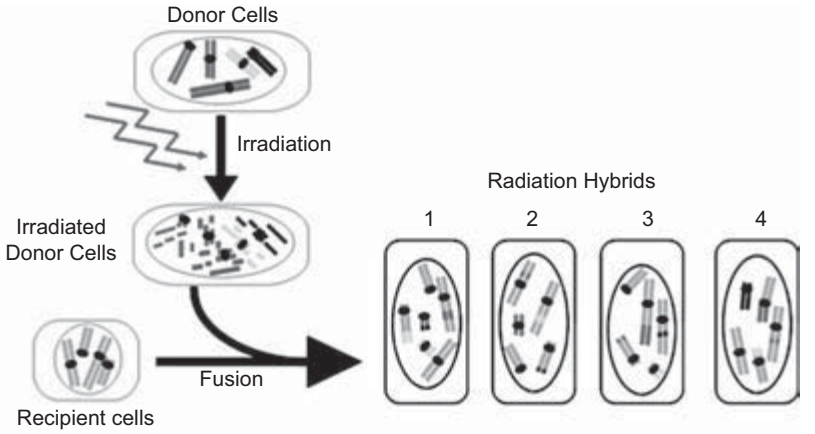


Figure 18.2. A schematic representing irradiation of donor cells and fusion to recipient cells to form a radiation hybrid cell panel. Donor cell fragments are retained in the hybrid cell lines as translocations, insertions, or partial chromosomes retaining replication machinery. High doses of radiation results in more chromosome breaks, increasing the mapping resolution of the panel. Notice that multiple fragments from the aqua donor cell chromosome are retained in hybrid cell 1. This complicates mapping analyses, therefore panels of ~100 hybrid cell lines are typically used to construct statistically significant linear marker orders. Similar to somatic cell hybrid mapping, cell lines are scored as positive or negative for each marker. The higher the frequency of co-retention for a set of markers, the closer they are expected to be to one another on the chromosome. Multiple strategies exist for creating donor cell selectable markers. (Also see color plate.)

benefit of this strategy is that screening all markers on a single panel of ~100 radiation hybrid cell lines can produce high-resolution maps of all chromosomes. Initially, WG-RH panels were only developed to construct genome maps for humans (Stewart et al. 1997, Schuler et al. 1996, Gyapay et al. 1996) and mammalian model organisms including mouse (Schmitt et al. 1996), rat (Watanabe et al. 1999), cat (Murphy et al. 1999), dog (Priat et al. 1998), and rhesus macaque (Murphy et al. 2001). In 1997, Womack and colleagues (1997) developed the first WG-RH panel for an agriculturally important species, the cow. Since then WG-RH panels have been constructed for mapping agriculturally important animal genomes including the pig (Yerle et al. 1998), horse (Chowdhary et al. 2002), chicken (Morisson et al. 2002), and deer (Ramsdell et al. 2006). RH panels have also been constructed for several agriculturally important plant species including cotton (Gao et al. 2004) and maize (Kynast et al. 2004).

Radiation Hybrid Mapping in Aquatic Species

In addition to mammalian model organisms, RH panels have also been produced to construct genome maps for the zebrafish, an aquatic model organism used to study the genetics of development, growth, reproduction, and disease resistance (Kwok et al. 1999, Geisler et al. 1999, Hukriede et al. 1999). The Goodfellow T51 panel was created by irradiating zebrafish donor fibroblasts with 3,000 rad and fusing them with rodent recipient cells. To date more than 18,000 markers having an average retention of 20–25% have been mapped using this resource. The public availability of this panel has facilitated the development of high-density maps. Researchers worldwide are able to purchase DNA from the panel, screen their markers, and analyze their new data with previously compiled data for a large number of markers spanning the entire genome.

To date the only RH panel and map reported for an aquacultured organism is for the gilthead seabream (Senger et al. 2006). Similar to the zebrafish panel, this WG-RH panel was created by irradiating a fibroblast cell line from seabream fin tissue with 3,000 rad of gamma rays and fusing it with a rodent recipient cell line. Although 170 radiation hybrid cell lines were produced, only 93 were necessary to construct a map comprised of 440 markers, including 288 microsatellites, 82 genes, and 70 sequence-tag sites. The marker retention average was ~30% and ranged from 8–59%. The markers were distributed throughout the genome, assembling into 29 linkage groups representing the n = 24 chromosomes in the species. On average each linkage group has 15 markers, ranging from 2–29; only 5/440 markers were unlinked. Based on an estimated genome size of 800 Mb and total RH map length of 5,683 cR_{3000}, the resolution of this map is estimated to be less than 2 Mb. Comparative analyses with the genome sequence of *Tetraodon* resulted in identification of 109 assignments of homology. This map is being used to study chromosomal evolution and to identify traits affecting aquaculture production efficiency.

Future Perspectives of RH Mapping for Aquaculture Species

In recent years the availability of genome research tools for aquaculture species has increased dramatically. However, these efforts have targeted the development of

genome tools and reagents such as genetic maps, construction of bacterial artificial chromosome (BAC) libraries (Shizuya et al. 1992) for the development of physical maps (Marra et al. 1997), and the identification of expressed sequence tags (EST) (Adams et al. 1993) for functional genomics. Species of great interest include catfish, tilapia, rainbow trout, Atlantic salmon, oyster, and shrimp. One reason that RH panels have not been developed is the increase in the number of species for which whole-genome sequences are available, including *Tetraodon*, Fugu, and zebrafish. This has led to the widespread expectation that technological advances will facilitate possibilities for assembling whole-genome sequences; therefore, physical maps using other technologies such as radiation hybrid mapping will become obsolete. Also, one limitation to creating RH panels is that up to 1 year must be spent in cell culture. This is quite an investment, especially since a large number of markers must be scored before the panel's quality can be determined. However, in spite of many community efforts to develop genome-sequencing projects for aquaculture species, to date there has been no success. As a result, more RH map resources for aquaculture species may be developed in the next few years. For many of these species, sufficient genetic markers are already available to produce high-density maps including ESTs, BAC-end sequences, and microsatellites. Once developed and characterized, RH maps will serve to integrate maps including multiple genetic maps, BAC physical maps, and whole-genome scaffold sequences. High-density RH maps have been used in other agriculture species to aid in the assembly of whole-genome sequences in an effort to identify genes affecting agriculturally important production traits (Weikard et al. 2006). As demonstrated with seabream, RH maps for additional aquaculture species will serve as comparative maps to be integrated with chromosome maps and/or genome sequences of other species.

References

Abbott C. 1992. Characterization of mouse-hamster somatic cell hybrids by PCR: a panel of mouse-specific primers for each chromosome. Mamm Genome, 2, pp. 106–109.

Adams MD, MB Soares, AR Kerlavage, C Fields, and JC Venter. 1993. Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. Nat Genet, 4, pp. 373–380.

Barrett JH. 1992. Genetic mapping based on radiation hybrid data. Genomics, 13, pp. 95–103.

Boehnke M, K Lange, and DR Cox. 1991. Statistical methods for multipoint radiation hybrid mapping. Am J Hum Genet, 49, pp. 1174–1188.

Botstein D, RL White, M Skolnick, and RW Davis. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet, 32, pp. 314–331.

Chowdhary BP, T Raudsepp, D Honeycutt, EK Owens, F Piumi, G Guerin, TC Matise, SR Kata, JE Womack, and LC Skow. 2002. Construction of a 5000 (rad) whole-genome radiation hybrid panel in the horse and generation of a comprehensive and comparative map for ECA11. Mamm Genome, 13, pp. 89–94.

Collins FS. 1995. Positional cloning moves from perditional to traditional. Nat Genet, 9, pp. 347–350.

Cox DR, M Burmeister, ER Price, S Kim, and RM Myers. 1990. Radiation hybrid mapping: a somatic cell genetic method for constructing high-resolution maps of mammalian chromosomes. Science, 250, pp. 245–250.

Ephrussi B and MC Weiss. 1969. Hybrid somatic cells. Sci Am, 220, pp. 26–35.

Gao W, ZJ Chen, JZ Yu, D Raska, RJ Kohel, JE Womack, and DM Stelly. 2004. Wide-cross whole-genome radiation hybrid mapping of cotton (*Gossypium hirsutum L.*). Genetics, 167, pp. 1317–1329.

Geisler R, GJ Rauch, H Baier, F van Bebber, L Bross, MP Dekens, K Finger, C Fricke, MA Gates, H Geiger, S Geiger-Rudolph, D Gilmour, S Glaser, L Gnugge, H Habeck, K Hingst, S Holley, J Keenan, A Kirn, H Knaut, D Lashkari, F Maderspacher, U Martyn, S Neuhauss, C Neumann, T Nicolson, F Pelegri, R Ray, JM Rick, H Roehl, T Roeser, HE Schauerte, AF Schier, U Schonberger, HB Schonthaler, S Schulte-Merker, C Seydler, WS Talbot, C Weiler, C Nusslein-Volhard, and P Haffter. 1999. A radiation hybrid map of the zebrafish genome. Nat Genet, 23, pp. 86–89.

Goss SJ and H Harris. 1975. New method for mapping genes in human chromosomes. Nature, 255, pp. 680–684.

Green H. 1969. Prospects for the chromosomal localization of human genes in human-mouse somatic cell hybrids. Wistar Inst Symp Monogr, 9, pp. 51–59.

Gyapay G, K Schmitt, C Fizames, H Jones, N Vega-Czarny, D Spillett, D Muselet, JF Prud'homme, C Dib, C Auffray, J Morissette, J Weissenbach, and PN Goodfellow. 1996. A radiation hybrid map of the human genome. Hum Mol Genet, 5, pp. 339–346.

Hukriede NA, L Joly, M Tsang, J Miles, P Tellis, JA Epstein, WB Barbazuk, FN Li, B Paw, JH Postlethwait, TJ Hudson, LI Zon, JD McPherson, M Chevrette, IB Dawid, SL Johnson, and M Ekker. 1999. Radiation hybrid mapping of the zebrafish genome. Proc Natl Acad Sci USA, 96, pp. 9745–9750.

Kwok C, R Critcher, and K Schmitt. 1999. Construction and characterization of zebrafish whole genome radiation hybrids. Methods Cell Biol, 60, pp. 287–302.

Kynast RG, RJ Okagaki, MW Galatowitsch, SR Granath, MS Jacobs, AO Stec, HW Rines, and RL Phillips. 2004. Dissecting the maize genome by using chromosome addition and radiation hybrid lines. Proc Natl Acad Sci USA, 101, pp. 9921–9926.

Lander ES and D Botstein. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics, 121, pp. 185–199.

Littlefield JW. 1964. Selection of hybrids from matings of fibroblasts in vitro and their presumed recombinants. Science, 145, pp. 709–710.

Marra MA, TA Kucaba, NL Dietrich, ED Green, B Brownstein, RK Wilson, KM McDonald, LW Hillier, JD McPherson, and RH Waterston. 1997. High throughput fingerprint analysis of large-insert clones. Genome Res, 7, pp. 1072–1084.

Mendel GV. 1951. Experiments in plant hybridization. Journal of Heredity, 42, pp. 3–4.

Morgan TH, AH Sturtevant, HJ Muller, and CB Bridges. 1915. The mechanism of Mendelian heredity. Henry Holt and Company.

Morisson M, A Lemiere, S Bosc, M Galan, F Plisson-Petit, P Pinton, C Delcros, K Feve, F Pitel, V Fillon, M Yerle, and A Vignal. 2002. ChickRH6: a chicken whole-genome radiation hybrid panel. Genet Sel Evol, 34, pp. 521–533.

Mullis KB and FA Faloona. 1987. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods Enzymol, 155, pp. 335–350.

Murphy WJ, M Menotti-Raymond, LA Lyons, MA Thompson, and SJ O'Brien. 1999. Development of a feline whole genome radiation hybrid panel and comparative mapping of human chromosome 12 and 22 loci. Genomics, 57, pp. 1–8.

Murphy WJ, JE Page, C Smith Jr, RC Desrosiers, and SJ O'Brien. 2001. A radiation hybrid mapping panel for the rhesus macaque. J Hered, 92, pp. 516–519.

O'Brien SJ, JE Womack, LA Lyons, KJ Moore, NA Jenkins, and NG Copeland. 1993. Anchored reference loci for comparative genome mapping in mammals. Nat Genet, 3, pp. 103–112.

Priat C, C Hitte, F Vignaux, C Renier, Z Jiang, S Jouquand, A Cheron, C Andre, and F Galibert. 1998. A whole-genome radiation hybrid map of the dog genome. Genomics, 54, pp. 361–378.

Ramsdell CM, EL Thames, JL Weston, and MJ Dewey. 2006. Development of a deer mouse whole-genome radiation hybrid panel and comparative mapping of Mus chromosome 11 loci. Mamm Genome, 17, pp. 37–48.

Schmitt K, JW Foster, RW Feakes, C Knights, ME Davis, DJ Spillett, and PN Goodfellow. 1996. Construction of a mouse whole-genome radiation hybrid panel and application to MMU11. Genomics, 34, pp. 193–197.

Schuler GD, MS Boguski, EA Stewart, LD Stein, G Gyapay, K Rice, RE White, P Rodriguez-Tome, A Aggarwal, E Bajorek, S Bentolila, BB Birren, A Butler, AB Castle, N Chiannilkulchai, A Chu, C Clee, S Cowles, PJ Day, T Dibling, N Drouot, I Dunham, S Duprat, C East, C Edwards, JB Fan, N Fang, C Fizames, C Garrett, L Green, D Hadley, M Harris, P Harrison, S Brady, A Hicks, E Holloway, L Hui, S Hussain, C Louis-Dit-Sully, J Ma, A MacGilvery, C Mader, A Maratukulam, TC Matise, KB McKusick, J Morissette, A Mungall, D Muselet, HC Nusbaum, DC Page, A Peck, S Perkins, M Piercy, F Qin, J Quackenbush, S Ranby, T Reif, S Rozen, C Sanders, X She, J Silva, DK Slonim, C Soderlund, WL Sun, P Tabar, T Thangarajah, N Vega-Czarny, D Vollrath, S Voyticky, T Wilmer, X Wu, MD Adams, C Auffray, NA Walter, R Brandon, A Dehejia, PN Goodfellow, R Houlgatte, JR Hudson Jr, SE Ide, KR Iorio, WY Lee, N Seki, T Nagase, K Ishikawa, N Nomura, C Phillips, MH Polymeropoulos, M Sandusky, K Schmitt, R Berry, K Swanson, R Torres, JC Venter, JM Sikela, JS Beckmann, J Weissenbach, RM Myers, DR Cox, MR James, D Bentley, P Deloukas, ES Lander, and TJ Hudson. 1996. A gene map of the human genome. Science, 274, pp. 540–546.

Senger F, C Priat, C Hitte, E Sarropoulou, R Franch, R Geisler, L Bargelloni, D Power, and F Galibert. 2006. The first radiation hybrid map of a perch-like fish: The gilthead seabream (Sparus aurata L). Genomics, 87, pp. 793–800.

Shizuya H, B Birren, UJ Kim, V Mancino, T Slepak, Y Tachiiri, and M Simon. 1992. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. Proc Natl Acad Sci USA, 89, pp. 8794–8797.

Siden TS, J Kumlien, CE Schwartz, and D Rohme. 1992. Radiation fusion hybrids for human chromosomes 3 and X generated at various irradiation doses. Somat Cell Mol Genet, 18, pp. 33–44.

Sobrino B, M Brion, and A Carracedo. 2005. SNPs in forensic genetics: a review on SNP typing methodologies. Forensic Sci Int, 154, pp. 181–194.

Southern EM. 1975. Detection of specific sequences among DNA fragments separated by gel electrophoresis. J Mol Biol, 98, pp. 503–517.

Stewart EA, KB McKusick, A Aggarwal, E Bajorek, S Brady, A Chu, N Fang, D Hadley, M Harris, S Hussain, R Lee, A Maratukulam, K O'Connor, S Perkins, M Piercy, F Qin, T Reif, C Sanders, X She, WL Sun, P Tabar, S Voyticky, S Cowles, JB Fan, DR Cox, et al. 1997. An STS-based radiation hybrid map of the human genome. Genome Res, 7, pp. 422–433.

Wakefield MJ and JA Graves. 1996. Comparative maps of vertebrates. Mamm Genome, 7, pp. 715–734.

Walter MA and PN Goodfellow. 1993. Radiation hybrids: irradiation and fusion gene transfer. Trends Genet, 9, pp. 352–356.

Walter MA, DJ Spillett, P Thomas, J Weissenbach, and PN Goodfellow. 1994. A method for constructing radiation hybrid maps of whole genomes. Nat Genet, 7, pp. 22–28.

Watanabe TK, MT Bihoreau, LC McCarthy, SL Kiguwa, H Hishigaki, A Tsuji, J Browne, Y Yamasaki, A Mizoguchi-Miyakita, K Oga, T Ono, S Okuno, N Kanemoto, E Takahashi, K Tomita, H Hayashi, M Adachi, C Webber, M Davis, S Kiel, C Knights, A Smith, R Critcher, J Miller, T Thangarajah, PJ Day, JR Hudson Jr, Y Irie, T Takagi, Y Nakamura, PN Goodfellow, GM Lathrop, A Tanigami, and MR James. 1999. A radiation hybrid map of the rat genome containing 5,255 markers. Nat Genet, 22, pp. 27–36.

Weber JL and PE May. 1989. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am J Hum Genet, 44, pp. 388–396.

Weikard R, T Goldammer, P Laurent, JE Womack, and C Kuehn. 2006. A gene-based high-resolution comparative radiation hybrid map as a framework for genome sequence assembly of a bovine chromosome 6 region associated with QTL for growth, body composition, and milk performance traits. BMC Genomics, 7, p. 53.

Womack JE, JS Johnson, EK Owens, CE Rexroad III, J Schlapfer, and YP Yang. 1997. A whole-genome radiation hybrid panel for bovine gene mapping. Mamm Genome, 8, pp. 854–856.

Womack JE and SR Kata. 1995. Bovine genome mapping: evolutionary inference and the power of comparative genomics. Curr Opin Genet Dev, 5, pp. 725–733.

Yerle M, P Pinton, A Robic, A Alfonso, Y Palvadeau, C Delcros, R Hawken, L Alexander, C Beattie, L Schook, D Milan, and J Gellin. 1998. Construction of a whole-genome radiation hybrid panel for high-resolution gene mapping in pigs. Cytogenet Cell Genet, 82, pp. 182–188.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 19
Comparative Genomics and Positional Cloning

Bo-Young Lee and Thomas D. Kocher

Genetic linkage maps of molecular markers have facilitated the identification of quantitative trait loci (QTL) for economically important traits in a variety of cultured fish species (Tao and Boulding 2003, Cnaani et al. 2004, Reid et al. 2005). In rainbow trout, several QTLs have been identified for disease resistance (Palti et al. 1999, Ozaki et al. 2001), environmental stress tolerance (Danzmann et al. 1999, Perry et al. 2001), and spawning dates (Sakamoto et al. 1999, O'Malley et al. 2003). In tilapia, QTLs for sex determination and stress response have been identified in several different species (Lee et al. 2003, Lee et al. 2004, Cnaani et al. 2004).

The next obvious goal is to identify the causative genes underlying these QTLs. By identifying these genes, and studying their function, we will gain insight into the physiological mechanisms underlying traits such as growth rate, salinity tolerance, and disease resistance. Identification of the genes also lays the groundwork for marker-assisted selective breeding to improve commercial stocks.

Positional cloning to identify the genes underlying QTL is still a very challenging task, and requires a convergence of three genetic resources. The first of these is a large population segregating for the QTL. Fortunately, most aquaculture species are fecund, and rearing large F2 or backcross families is not difficult. The second resource is a collection of closely spaced molecular markers to allow fine-scale genetic mapping of the QTL. These are typically microsatellite markers, but increasingly single-nucleotide polymorphisms (SNP) are being used. The third essential resource is an ordered collection of physical clones, or even better, the DNA sequence, spanning the region containing the QTL. Because research funds are finite, it is not likely that complete genome sequences for the great diversity of fish species used in aquaculture will be available soon. Only a few biomedical model species (e.g., Fugu, *Tetraodon*, *Oryzias*, *Danio*) have been sequenced to date. Even if genome sequencing is initiated for important aquaculture species, it will likely progress only as far as a relatively incomplete draft assembly. For the foreseeable future, genome resources for these species will be limited to moderate density genetic linkage maps of microsatellite markers, and physical maps based on restriction fingerprints of large-insert bacterial artificial chromosome (BAC) clones (Table 19.1). Nevertheless, even with this minimal set of resources, it should be possible to positionally clone genes for important traits.

Positional cloning is the process of mapping QTL to ever-smaller regions of the chromosome until the causative gene is identified. Each time the region is narrowed; new markers are identified and mapped, until a continuous path of clones or DNA sequence are identified across the region of the QTL. Experiments then can be designed to test the effects of mutations in particular genes within the region. Because

Table 19.1. Genomic resources of aquaculture fish.

Scientific name	Common name	Genome size (Mb)	Resources	References	Web resources for analysis
Danio rerio	Zebrafish	1,700	Genetic map	Postlethwait et al. 1994	http://zfin.org/cgi-bin/webdriver?MIval=aa-ZDB_home.apg
			RH map	Hukriede et al. 1999	
			ESTs	Kelly et al. 2000	
			Physical map		http://www.sanger.ac.uk/Projects/D_rerio/webFPC/zebrafish/small.shtml
			Genome project		http://www.sanger.ac.uk/Projects/D_rerio/ http://www.ensembl.org/Danio_rerio/index.html
Ictalurus punctatus	Channel catfish	1,000	Genetic map	Liu et al. 2003	
			ESTs	He et al. 2003	http://www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=catfish http://morag.umsmed.edu/
Oncorhynchus mykiss	Rainbow trout	2,600	Genetic map	Nichols et al. 2003	
			ESTs	Rexroad et al. 2003	http://www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=r_trout
Salmo salar	Atlantic salmon	3,000	Genetic map	Gilbey et al. 2004, Moen et al. 2004	
			ESTs	Rise et al. 2004	http://www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=salmon http://web.uvic.ca/cbr/grasp/ http://www.salmongenome.no/cgi-bin/sgp.cgi
Gasterosteus aculeatus	Three-spined stickleback	675	Genetic map	Peichel et al. 2001	http://cegs.stanford.edu/Linkage_Maps.jsp
			Physical map		http://www.bcgsc.ca/platform/mapping/data http://cegs.stanford.edu/Physical_map.jsp

			ESTs		http://cegs.stanford.edu/stickleback_EST_sequences2.jsp
Oreochromis niloticus	Nile tilapia	1,000	Genetic map	Kocher et al. 1998, Lee et al. 2005	http://www.thearkdb.org/ http://hcgs.unh.edu/comp/
			Physical map	Katagiri et al. 2005	http://hcgs.unh.edu/ushl/
Several species of	East African cichlids		Genetic map	Albertson et al. 2003,	http://hcgs.unh.edu/comp/index.php?choice=2
				Streelman et al. 2003	http://hcgs.unh.edu/comp/index.php?choice=2
			ESTs		http://www.tigr.org/tdb/tgi/
Takifugu rubripes	Torafugu	400	Genetic map	Kai et al. 2005	http://www.se.a.u-tokyo.ac.jp/Fugu-Map/map.html
			ESTs	Clark et al. 2003	http://www.ensembl.org/Fugu_rubripes/index.html http://www.fugu-sg.org/
			Genome project	Aparicio et al. 2002	http://www.ensembl.org/Takifugu_rubripes/index.html http://genome.jgi-psf.org/Takru4/Takru4.home.html
Tetraodon nigroviridis	Spotted green pufferfish	385	Genome project	Jaillon et al. 2004	http://www.genoscope.cns.fr/externe/tetranew/
Oryzias latipes	Japanese medaka	1,000	Genetic map	Naruse et al. 2000	http://mbase.bioweb.ne.jp/~dclust/ml_base.html
			ESTs		http://mbase.bioweb.ne.jp/~dclust/me_base.html http://ani.embl.de:8080/mepd/
			Physical map	Khorasani et al. 2004	
			Genome project		http://dolphin.lab.nig.ac.jp/medaka/ http://shigen.lab.nig.ac.jp/medaka/genome/top.jsp

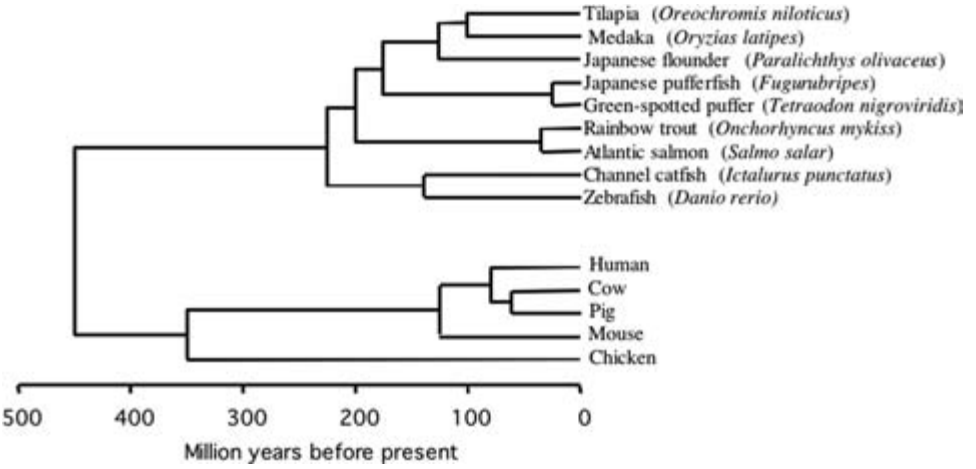


Figure 19.1. Phylogenetic relationships of important aquaculture species relative to species with sequenced genome. This diagram is a synthesis of molecular and paleontological work, especially Kumar and Hedges (1998) and ongoing work by Miya and colleagues (2003).

of the large size of eukaryotic genomes and the typically low rates of recombination within a particular interval, positional cloning can be challenging and tedious.

The speed at which a positional cloning project 'walks' along a chromosome can be enhanced significantly through comparative mapping. Teleost fishes are a closely related group, so comparative genomics can provide some insight into genome content and structure of species that have not been sequenced. Conservation of gene order among fishes at scales of several megabases (Mb) allows us to use the relatively complete sequences of model fish species to accelerate gene discovery and positional cloning of economically important traits in cultured fish. Figure 19.1 shows the phylogenetic relationship of important aquaculture species relative to species with sequenced genomes. Creative bioinformatics approaches allow us to take advantage of the genome sequences of Fugu, *Tetraodon*, and *Danio* to predict the sequence of genes in unsequenced species such as catfish, rainbow trout, and tilapia.

This chapter outlines a general strategy for leveraging the genome sequences of model species to infer the gene content of QTL regions in aquaculture species. Our focus is on tilapia, one of the most popular fish in aquaculture, for which extensive genetic linkage maps and physical maps of fingerprinted BAC clones have been developed. Using these resources, we demonstrate the development of a comparative map across a region contributing to sex determination in the Nile tilapia (*Oreochromis niloticus*).

Genetic Linkage Maps as Resources for Comparative Mapping

Genetic maps containing a few hundred polymorphic markers provide a starting point for determining the chromosomal location of cloned genes or markers (see Chapter 10), and are also used as a framework for QTL analysis (see Chapter 11). Several linkage maps of amplified fragment length polymorphism (AFLP) and microsatellite

DNA markers have been developed using intra- and interspecific crosses among tilapia species (Kocher et al. 1998, Agresti et al. 2000, McConnell et al. 2000). The most recent map is based on the segregation of 525 microsatellite and 21 gene-based markers in an F2 cross between *O. niloticus* and *O. aureus.* This map spans 1,131 centiMorgans (cM) in 24 linkage groups, and has an average marker spacing of 2.4 cM (Lee et al. 2005).

Expressed Sequence Tags as Resources for Comparative Mapping

Sequencing of complementary DNA (cDNA) clones to generate expressed sequence tags (EST) is the most efficient method for gene discovery. ESTs are good anchors for comparative mapping because the homology of particular sequences among species can be readily assessed through sequence analysis. ESTs are also readily included in genetic linkage maps by identifying SNPs or microsatellites in the ESTs. Microsatellite markers developed from EST sequences are particularly useful because they serve as both highly polymorphic markers for genetic mapping and highly conserved markers for comparative mapping. Because they are located near genes, the sequences flanking the microsatellite are highly conserved, making it relatively easy to recover the homologous sequences from related species by PCR amplification. For these reasons, microsatellites associated with ESTs have become a favorite marker for comparative mapping among fish species (Serapion et al. 2004, Rexroad et al. 2005).

Although relatively few tilapia ESTs are available in the public databases, large numbers of ESTs have been developed for closely related species of haplochromine cichlid from East Africa (Renn et al. 2004, Watanabe et al. 2004). Because these species diverged from the tilapias only about 20 million years ago, they provide a rich source of sequence information for tilapia genomics.

BAC Libraries and Physical Maps as Resources for Comparative Mapping

Libraries of large cloned fragments, contained in bacterial or yeast artificial chromosomes (BAC or YAC), are useful for constructing the physical resource of cloned DNA spanning a particular gene or QTL (see Chapters 13 and 14). Several high-quality BAC libraries have been produced for tilapia (Katagiri et al. 2001).

A BAC-based physical map of the tilapia genome has been developed by characterizing the restriction fragment fingerprints of about 35,000 BAC clones (Katagiri et al. 2005). Overlapping clones were identified from the fingerprint data using the computer program FPC (version 6.0). This physical map consists of 3,621 contigs of overlapping clones, and is estimated to span a total of 1.752 Gb in physical length (Katagiri et al. 2005).

Several methods are available to create physical maps at larger spatial scales. Methods for mapping cloned DNA to chromosomes by fluorescent *in situ* hybridization (FISH) are well developed for tilapia, as described in Chapter 17 (Harvey et al. 2003, Ezaz et al. 2004). Radiation hybrid maps (see Chapter 18) are also very useful, but a radiation hybrid panel is not yet available for tilapia.

Identification of BAC Clones Containing Genetically Mapped Markers

To relate the genetic and physical maps, it is necessary to screen the BAC libraries for clones containing genetically mapped markers. We have used two methods for screening the BAC libraries.

The first method uses PCR to identify clones containing a particular marker sequence. Rather than test each clone separately, the BACs are pooled in various ways to reduce the number of PCR reactions required. Pools were constructed from a total of 24,000 clones that provide approximately 2× coverage from the two tilapia libraries with the largest inserts (Katagiri et al. 2005). These clones are stored in 252 96-well plates, and the initial step in pooling was to pool the clones from each row and column. The row pools were then pooled to produce a single pool for each plate. Groups of 25 plates were arranged in a 5 × 5 matrix, and 'super-row' and 'super-column' pools made from the plates in the rows and columns of this matrix. Finally, 'superpools' were constructed from the 25 plates in each group. These pooled DNAs allowed us to identify positive clones with a sequence of just four cycles of PCR screening. We first screened the 10 'superpools' to identify which groups of 25 plates contained the marker. We then screened the 'super-row' and 'super-column' pools for those groups to identify the particular plate containing the marker. Next, we screened the row pools of a single 96-well plate to identify the positive row. A final round of PCR confirmed the identity of the positive clone.

BAC libraries can also be screened by filter hybridization (Asakawa and Shimizu 1998, Foster et al. 2001). Nylon filters can be spotted with as many as 18,432 independent clones. These high-density filters can be hybridized using either radioactive or nonradioactive probes. We used filter hybridization to identify BACs containing some EST markers in the sex-determining region.

Shotgun Sequencing

Once a clone containing a particular marker was identified, we used the database of restriction fingerprints to identify overlapping clones. After confirming the marker content of the clones in this contig, one or more clones were selected for shotgun sequencing.

Initially a small sample of 100–200 sequences from each BAC were used to identify the synteny between the BAC clone and the sequence of a reference genome. Each BAC clone was sheared to small fragments and cloned into pGEM-T vector. Inserts were amplified by PCR using the universal primer T7 or M13R. PCR products were purified and cycle sequenced. Sequences were analyzed on an ABI 377 automated DNA sequencer.

With the introduction of new sequencing techniques, it is now feasible to completely sequence each BAC. We contracted with 454 Life Sciences (Branford, CT) to completely sequence a BAC from each of the three contigs in the sex-determining region. We identified many genes and developed several new markers from the sequence of these BAC clones.

Identification of Sequence Similarities

Sequences from shotgun libraries were analyzed to identify homologous regions of sex determination between tilapia and other fish species. We used the Basic Local Alignment Search Tool (BLAST) algorithm to search the various databases of linkage maps and genome sequences for fish species.

Our first efforts at comparative mapping used the Fugu genome sequence. Shotgun sequences from the sex-determining region in tilapia found several homologous Fugu scaffolds. However, the largest of these scaffolds covered only 190 kilobase (kb), which is too short to generate predictions useful for chromosome walking.

Medaka is the biomedical model most closely related to tilapia. Genome resources for this species include an extensive genetic linkage map based on EST sequences, and a preliminary genome assembly. The linkage map provides useful information about conserved synteny at the largest spatial scales, but there is currently little overlap between the genes known from tilapia and the genes placed on the medaka linkage map. The preliminary medaka genome assembly provides useful information at the scale of sequence scaffolds (<8 Mbs).

The genome sequence of *Tetraodon* proved to be the most useful for our project, because the sequence contigs have been joined into chromosome-sized scaffolds. Our shotgun sequences identified a homologous region almost 2 Mb long on *Tetraodon* chromosome 5. This information provided a good starting point for fine mapping and chromosome walking in the sex-determining region of tilapia.

Reciprocal BLAST hits can provide good anchors for comparative maps, but raw BLAST results must be used with care. High sequence similarities among nonorthologous repetitive elements must be excluded from the analysis. Misleading BLAST results are also common. For example, the rod opsin gene, which was derived by reverse transcription from an early vertebrate green cone opsin gene, frequently gives a higher BLAST score than the true orthologue of a cone opsin EST simply because the sequence similarities form a longer continuous block.

The Major Sex-determining Region of the Nile Tilapia

Oreochromis niloticus has been mapped to LG1 (Lee et al. 2003). The sex-determining locus lies between markers GM201 and UNH995 with an interval of 11 cM. Because we had no additional markers in this interval, comparative information was needed to develop new markers.

Fine Mapping

We started by polymerase chain reaction (PCR) screening the BAC library pools for marker UNH995. We identified four positive BAC clones, and recovered the corresponding BAC contig from the fingerprint database. We found a reliable contig of 22 BAC clones (FPC tolerance = 5, significance = 1e − 07). A minimum tiling path of four clones across the contig was chosen for shotgun sequencing of 100–200 reads per

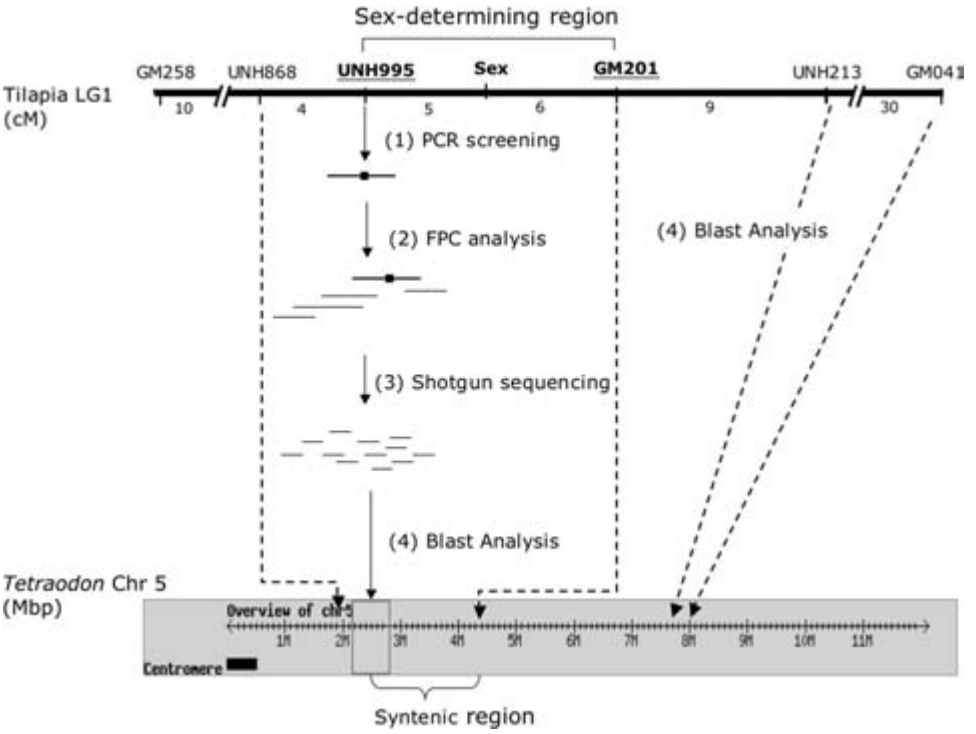


Figure 19.2. Identification of the synteny of sex-determining region between tilapia and *Tetraodon.* (1) Marker UNH995 was used for PCR probe to screen BAC pools and to identify BAC clones containing the marker. (2) Using fingerprinting data and FPC analysis, a contig containing the clones was identified. (3) A minimum tiling path of clones across the contig was shotgun sequenced. (4) Shotgun sequences were blasted against *Tetraodon* genome database. Square box in the *Tetraodon* chromosome presents a region that tilapia shotgun sequences hit by BLAST. Dotted arrows indicated where the tilapia microsatellites marker hit in *Tetraodon* chr5. The sex-determining region of LG1 seems to be syntenic with the 2.7~4.8 Mb region of *Tetraodon* chr5.

clone. The shotgun sequences were searched using BLAST against the several fish genome databases. Several of the sequences matched annotated genes and identified homologous regions of other fish genomes. In particular, the BLAST results consistently hit a block of genes *Tetraodon* chromosome 5. The sequence of other microsatellite markers on tilapia linkage group 1 (UNH868, UNH213, and GM041) also hit *tetraodon* chr5. The sex-determining region of LG1 seems to be syntenic with the region of *Tetraodon* chr5 from 2.7 Mb to 4.8 Mb. The order of these markers in this region is consistent between the two species (Figure 19.2).

We next sought to identify other gene sequences that might map to the same region and that might be used to develop additional genetic markers across the interval. We used BLAST to identify sequence similarities between cichlid ESTs and the corresponding region of *Tetraodon* chr5. A microsatellite marker (BJ702072) was developed from one of the ESTs, and SNPs were identified in seven others (BJ692919, BJ687805, BJ678620, AB004170, BJ675743, BJ690768, and BJ684727).

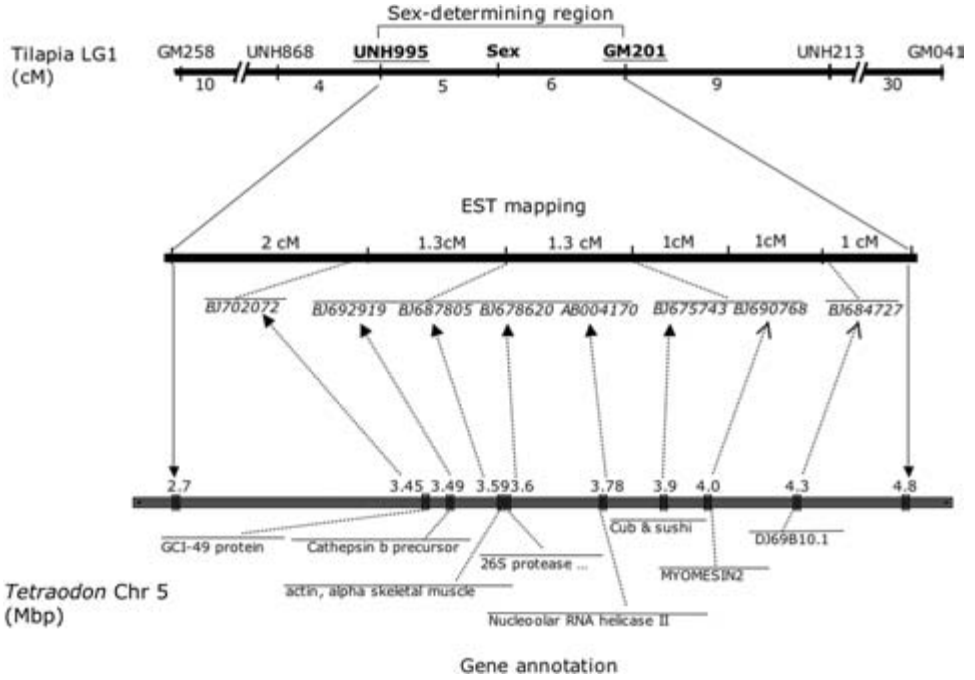


Figure 19.3. Fine mapping of cichlid ESTs in the sex-determining region of tilapia. Using BLAST analysis, some cichlid ESTs were collected by identifying sequence similarities between cichlid ESTs and the corresponding region of *Tetraodon* chr5. Eight of them were successfully mapped in the sex-determining region in tilapia. Markers started with BJ or AB present cichlid EST. Recombinants were found for BJ702072 and BJ675743, which narrow the sex-determining region to 2.6 cM. This region represents about 400 kb in *Tetraodon* genome, and corresponds to about 1 Mb in the tilapia genome.

These eight markers allowed us to further narrow the sex-determining region to 2.6 cM interval between markers BJ702072 and BJ675743 (Figure 19.3). This region represents about 400 kb in the *Tetraodon* genome, and corresponds to about 1 Mb in the tilapia genome. Although we have four informative EST markers within this region, we have not yet identified any additional recombinants to narrow the interval. We are currently genotyping several thousand additional fish to find recombinants. As these new EST markers were mapped, we also used them to screen the BAC libraries to complete the physical map across the region. BAC clones containing these markers assembled into four separate contigs. We determined the orientation of three of these contigs on the *Tetraodon* sequence by determining the marker content of individual BAC clones (Figure 19.4). One BAC from each of these three contigs was shotgun sequenced at 454 Life Sciences. These sequences confirmed the synteny between tilapia and *Tetraodon* in this region. The three remaining gaps in the physical map are being closed by chromosome walking in the BAC libraries. The complete gene content of this region of the tilapia will not be known until the physical map and sequencing are complete.

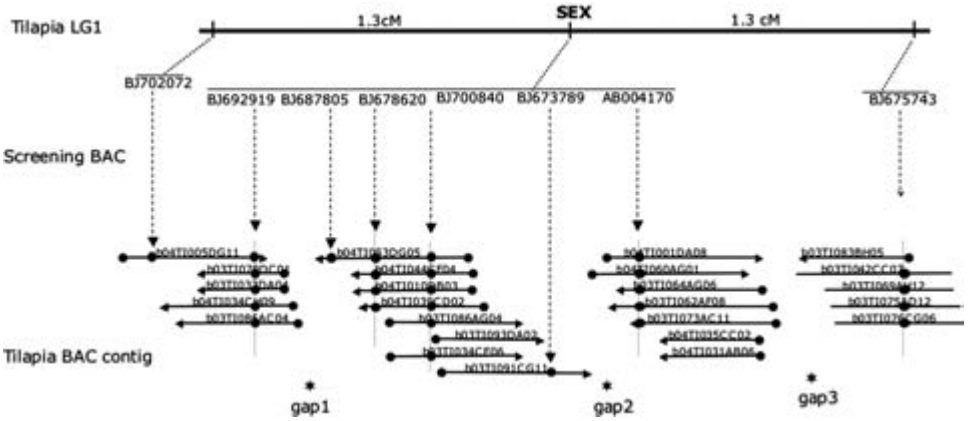


Figure 19.4. BAC-based physical mapping across the sex-determining region in tilapia. BAC libraries were screened using new EST markers. BAC clones containing these markers assembled into four separate contigs. The marker content of individual BAC clones determined the orientation of three contigs from left. The top clones of the three contigs were shotgun sequenced using the 454 sequencing technology. The three remaining gaps in the physical map can be closed by chromosome walking in the BAC libraries.

Conclusion

Many QTL for economically important traits are being identified in aquaculture species including tilapia. Positional cloning to identify the genes underlying these QTLs is still a very challenging exercise because high-density linkage maps or highly redundant physical maps are not yet available for most of these taxa. Conservation of gene order among fish at scales of several Mbs allows us to use the relatively complete sequences of model fish species to accelerate gene discovery and positional cloning of these genes. Comparative mapping has greatly accelerated our chromosome walks for the genes controlling both sex determination and red skin color in tilapia.

Creative bioinformatics approaches will allow us to take advantage of the genome sequences of Fugu, *Tetraodon*, and *Danio* to predict the sequence of genes in important aquaculture species such as catfish, rainbow trout, and tilapia. We anticipate that these informatic tools will enable a strategy that will allow the development of a comprehensive set of genetic, physical, and comparative maps for a new species for less than $1 million, thereby extending the number of fish taxa that can be studied with genomic methodologies.

References

Agresti JJ, S Seki, A Cnaani, S Poompuang, EM Hallerman, N Umiel, G Hulata, GAE Gall, and B May. 2000. Breeding new strains of tilapia: development of an artificial center of origin and linkage map based on AFLP and microsatellite loci. Aquaculture, 185, pp. 43–56.

Albertson RC, JT Streelman, and TD Kocher. 2003. Directional selection has shaped the oral jaws of Lake Malawi cichlid fishes. Proc Natl Acad Sci USA, 100, pp. 5252–5257.

Aparicio S, J Chapman, E Stupka, N Putnam, JM Chia, P Dehal, A Christoffels, S Rash, S Hoon, A Smit, MD Gelpke, J Roach, T Oh, IY Ho, M Wong, C Detter, F Verhoef, P Predki, A Tay, S Lucas, P Richardson, SF Smith, MS Clark, YJ Edwards, N Doggett, A Zharkikh, SV Tavtigian, D Pruss, M Barnstead, C Evans, H Baden, J Powell, G Glusman, L Rowen, L Hood, YH Tan, G Elgar, T Hawkins, B Venkatesh, D Rokhsar, and S Brenner. 2002. Whole genome shotgun assembly and analysis of the genome of *Fugu rubripes.* Science, 297, pp. 1301–1310.

Asakawa S and N Shimizu. 1998. High-fidelity digital hybridization screening. Genomics, 49, pp. 209–217.

Clark MS, YJ Edwards, D Peterson, SW Clifton, AJ Thompson, M Sasaki, Y Suzuki, K Kikuchi, S Watabe, K Kawakami, S Sugano, G Elgar, and SL Johnson. 2003. *Fugu* ESTs: new resources for transcription analysis and genome annotation. Genome Res, 13, pp. 2747–2753.

Cnaani A, N Zilberman, S Tinman, G Hulata, and M Ron. 2004. Genome-scan analysis for quantitative trait loci in an F2 tilapia hybrid. Mol. Genet Genomics, 272, pp. 162–72.

Danzmann RG, TR Jackson, and MM Ferguson. 1999. Epistasis in allelic expression at upper temperature tolerance QTL in rainbow trout. Aquaculture, 173, pp. 45–58.

Ezaz MT, SC Harvey, C Boonphakdee, AJ Teale, BJ McAndrew, and DJ Penman. 2004. Isolation and physical mapping of sex-linked markers in Nile tilapia (*Oreochromis niloticus*). Mar Biotechnol, 6, pp. 435–445.

Foster JM, IH Kamal, J Daub, MC Swan, JR Ingram, M Ganatra, J Ware, D Guiliano, A Aboobaker, L Moran, M Blaxter, and BE Slatko. 2001. Hybridization to high-density filter arrays of a Brugia malayi BAC library with biotinylated oligonucleotides and PCR products. Biotechniques, 30, pp. 1216–1218.

Gilbey J, E Verspoor, A McLay, and D Houlihan. 2004. A microsatellite linkage map for Atlantic salmon (*Salmo salar*). Anim Genet, 35, pp. 98–99.

Harvey SC, JY Kwon, and DJ Penman. 2003. Physical mapping of the brain and ovarian aromatase genes in the Nile tilapia, *Oreochromis niloticus*, by fluorescence in situ hybridization. Anim Genet, 34, pp. 62–64.

He C, L Chen, M Simmons, P Li, S Kim, and ZJ Liu. 2003. Putative SNP discovery in interspecific hybrids of catfish by comparative EST analysis. Anim Gent, 34, pp. 445–448.

Hukriede NA, L Joly, M Tsang, J Miles, P Tellis, JA Epstein, WB Barbazuk, FN Li, B Paw, JH Postlethwait, TJ Hudson, LI Zon, JD McPherson, M Chevrette, IB Dawid, SL Johnson, and M Ekker. 1999. Radiation hybrid mapping of the zebrafish genome. Proc Natl Acad Sci USA, 96, pp. 9745–9750.

Jaillon O, JM Aury, F Brunet, JL Petit, N Stange-Thomann, E Mauceli, L Bouneau, C Fischer, C Ozouf-Costaz, A Bernot, S Nicaud, D Jaffe, S Fisher, G Lutfalla, C Dossat, B Segurens, C Dasilva, M Salanoubat, M Levy, N Boudet, S Castellano, V Anthouard, C Jubin, V Castelli, M Katinka, B Vacherie, C Biemont, Z Skalli, L Cattolico, J Poulain, V De Berardinis, C Cruaud, S Duprat, P Brottier, JP Coutanceau, J Gouzy, G Parra, G Lardier, C Chapple, KJ McKernan, P McEwan, S Bosak, M Kellis, JN Volff, R Guigo, MC Zody, J Mesirov, K Lindblad-Toh, B Birren, C Nusbaum, D Kahn, M Robinson-Rechavi, V Laudet, V Schachter, F Quetier, W Saurin, C Scarpelli, P Wincker, ES Lander, J Weissenbach, and H Roest Crollius. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature, 431, pp. 946–957.

Kai W, K Kikuchi, M Fujita, H Suetake, A Fujiwara, Y Yoshiura, M Ototake, B Venkatesh, K Miyaki, and Y Suzuki. 2005. A genetic linkage map for the tiger pufferfish, *Takifugu rubripes.* Genetics, 171, pp. 227–238.

Katagiri T, S Asakawa, S Minagawa, N Shimizu, I Hirono, and T Aoki. 2001. Construction and characterization of BAC libraries for three fish species; rainbow trout, carp and tilapia. Anim Genet, 32, pp. 200–204.

Katagiri T, C Kidd, E Tomasino, J Davis, C Wishon, JE Stern, K Carleton, AE Howe, and TD Kocher. 2005. A BAC-based physical map of the Nile tilapia genome. BMC Genomics, 6, 89.

Kelly PD, F Chu, IG Woods, P Ngo-Hazelett, T Cardozo, H Huang, F Kimm, L Liao, YL Yan, Y Zhou, SL Johnson, R Abagyan, AF Schier, JH Postlethwait, and WS Talbot. 2000. Genetic linkage mapping of zebrafish genes and ESTs. Genome Res, 19, pp. 558–567.

Khorasani MZ, S Hennig, G Imre, S Asakawa, S Palczewski, A Berger, H Hori, K Naruse, H Mitani, A Shima, H Lehrach, J Wittbrodt, H Kondoh, N Shimizu, and H Himmelbauer. 2004. A first generation physical map of the medaka genome in BACs essential for positional cloning and clone-by-clone based genomic sequencing. Mech Dev, 121, pp. 903–913.

Kocher TD, WJ Lee, H Sobolewska, D Penman, and B McAndrew. 1998. A genetic linkage map of a cichlid fish, the tilapia (*Oreochromis niloticus*). Genetics, 148, pp. 1225–1232.

Kumar S and SB Hedges. 1998. A molecular timescale for vertebrate evolution. Nature, 392, pp. 917–920.

Lee BY, G Hulata, and TD Kocher. 2004. Two unlinked loci controlling the sex of blue tilapia (*Oreochromis aureus*). Heredity, 92, pp. 543–549.

Lee BY, WJ Lee, JT Streelman, KL Carleton, AE Howe, G Hulata, A Slettan, JE Stern, Y Terai, and TD Kocher. 2005. A second-generation genetic linkage map of tilapia (*Oreochromis spp.*). Genetics, 170, pp. 237–244.

Lee BY, DJ Penman, and TD Kocher. 2003. Identification of a sex-determining region in Nile tilapia (*Oreochromis niloticus*) using bulked segregant analysis. Anim Genet, 34, pp. 379–383.

Liu Z, A Karsi, P Li, D Cao, and R Dunham. 2003. An AFLP-based genetic linkage map of channel catfish (*Ictalurus punctatus*) constructed by using an interspecific hybrid resource family. Genetics, 165, pp. 687–694.

McConnell SKJ, C Beynon, J Leamon, and DOF Skibinski. 2000. Microsatellite marker based genetic linkage maps of *Oreochromis aureus* and *O. niloticus* (*Cichlidae*): extensive linkage group segment homologies revealed. Anim Genet, 31, pp. 214–218.

Miya M, H Takeshima, H Endo, NB Ishiguro, JG Inoue, T Mukai, TP Satoh, M Yamaguchi, A Kawaguchi, K Mabuchi, SN Shirai, and M Nishida. 2003. Major patterns of higher teleostean phylogenies: a new perspective based on 100 complete mitochondrial DNA sequences. Mol Phyl Evol, 26, pp. 121–38.

Moen T, B Hoyheim, H Munch, and L Gomez-Raya. 2004. A linkage map of Atlantic salmon (*Salmo salar*) reveals an uncommonly large difference in recombination rate between the sexes. Anim Genet, 35, pp. 81–92.

Naruse K, S Fukamachi, H Mitani, M Kondo, T Matsuoka, S Kondo, N Hanamura, Y Morita, K Hasegawa, R Nishigaki, A Shimada, H Wada, T Kusakabe, N Suzuki, M Kinoshita, A Kanamori, T Terado, H Kimura, M Nonaka, and A Shima. 2000. A detailed linkage map of medaka, *Oryzias latipes*: comparative genomics and genome evolution. Genetics, 154, pp. 1773–1784.

Nichols KM, WP Young, RG Danzmann, BD Robison, C Rexroad, M Noakes, RB Phillips, P Bentzen, I Spies, K Knudsen, FW Allendorf, BM Cunningham, J Brunelli, H Zhang, S Ristow, R Drew, KH Brown, PA Wheeler, and GH Thorgaard. 2003. A consolidated linkage map for rainbow trout (*Oncorhynchus mykiss*). Anim Genet, 34, pp. 102–115.

O'Malley KG, T Sakamoto, RG Danzmann, and MM Ferguson. 2003. Quantitative trait loci for spawning data and body weight in rainbow trout: testing for conserved effects across ancestrally duplicated chromosomes. J Heredity, 94, pp. 273–284.

Ozaki A, T Sakamoto, S Khoo, K Nakamura, MRM Coimbra, T Akutsu, and N Okamoto. 2001. Quantitative trait loci (QTLs) associated with resistance/susceptibility to infectious pancreatic necrosis virus (IPNV) in rainbow trout (*Oncorhynchus mykiss*). Mol Genet Genomics, 265, pp. 23–31.

Palti Y, JE Parsons, and GH Thorgaard. 1999. Identification of candidate DNA markers associated with IHN virus resistance in backcrosses of rainbow (*Oncorhynchus mykiss*) and cutthroat trout (*O. clarki*). Aquaculture, 173, pp. 81–94.

Peichel CL, KS Nereng, KA Ohgi, BL Cole, PF Colosimo, CA Buerkle, D Schluter, and DM Kingsley. 2001. The genetic architecture of divergence between three spine stickleback species. Nature, 414, pp. 901–905.

Perry GML, RG Danzmann, MM Ferguson, and JP Gibson. 2001. Quantitative trait loci for upper thermal tolerance in outbred strains of rainbow trout (*Oncorhynchus mykiss*). Heredity, 86, pp. 333–341.

Postlethwait JH, SL Johnson, CN Midson, WS Talbot, M Getes, EW Ballinger, et al. 1994. A genetic linkage map for the zebrafish. Science, 264, pp. 699–703.

Reid DP, A Szanto, B Glebe, RG Danzmann, and MM Ferguson. 2005. QTL for body weight and condition factor in Atlantic salmon (*Salmo salar*): comparative analysis with rainbow trout (*Onchrynchus mykiss*) and Arctic charr (*Salvelinus alpinus*). Heredity, 94, pp. 166–172.

Renn SCP, N Aubin-Horth, and H Hofmann. 2004. Biologically meaningful expression profiling across species using heterologous hybridization to a cDNA microarray. BMC Genomics, 5, pp. 42–54.

Rexroad III CE, Y Lee, JW Keele, S Karamycheva, G Brown, B Koop, SA Gahr, Y Palti, and J Quackenbush. 2003. Sequence analysis of a rainbow trout cDNA library and creation of a gene index. Cytogenet Genome Res, 102, pp. 347–354.

Rexroad III CE, MF Rodriguez, I Coulibaly, K Gharbi, RG Danzmann, J DeKoning, R Phillips, and Y Palti. 2005. Comparative mapping of expressed sequence tags containing microsatellite in rainbow trout (*Oncorhynchus mykiss*). BMC Genomics, 6, pp. 54–61.

Rise ML, KR von Schalburg, GD Brown, MA Mawer, RH Devlin, N Kuipers, M Busby, M Beetz-Sargent, R Alberto, AR Gibbs, P Hunt, R Shukin, JA Zeznik, C Nelson, SR Jones, DE Smailus, SJ Jones, JE Schein, MA Marra, YS Butterfield, JM Stott, SH Ng, WS Davidson, and BF Koop. 2004. Development and application of a salmonid EST database and cDNA microarray: data mining and interspecific hybridization characteristics. Genome Res, 14, pp. 478–490.

Sakamoto T, RG Danzmann, N Okamoto, MM Ferguson, and PE Ihssen. 1999. Linkage analysis of quantitative trait loci associated with spawning time in rainbow trout (*Oncorhynchus mykiss*). Aquaculture, 173, pp. 33–43.

Serapion J, H Kucuktas, J Feng, and Z Liu. 2004. Bioinformatic mining of Type I microsatellites from expressed sequence tags of channel catfish (*Ictalurus punctatus*). Mar Biotechnol, 6, pp. 364–477.

Tao WJ and EG Boulding. 2003. Association between single nucleotide polymorphisms in candidate genes and growth rate in arctic charr (*Salvelinus alpinus* L.). Heredity, 92, pp. 60–69.

Watanabe M, N Kobayashi, T Shin-i, T Horiike, Y Tateno, Y Kohara, and N Okada. 2004. Extensive analysis of ORF sequences from two different cichlid species in Lake Victoria provides molecular evidence for a recent radiation event of the Victoria species flock: identity of EST sequences between *Haplochromis chilotes* and *Haplochromis sp.* "Redtail sheller." Gene, 243, pp. 263–269.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Part 3
Analysis of Genome Expression and Function

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 20
Transcriptome Characterization Through the Analysis of Expressed Sequence Tags

Zhanjiang Liu

A good discussion of transcriptome analysis requires us first to review the genetic central dogma and define our terms. The entire genetic material of an organism is defined as its genome. The DNA of an organism is transcribed into RNA and then translated into proteins as the final biologically active molecules. The entire RNA composition of an organism thus is defined as its transcriptome while the complete protein component of the organism is defined as its proteome. Although the genome is relatively stable for a given organism, the transcriptome is dynamic depending on changes in development, physiological conditions, and the environment. The RNA expressed in a cell at any given moment, both the classes of genes expressed and their level of expression, depends on the biological state of the cell.

Sequencing of expressed sequence tags (EST) has been a primary approach in the characterization of the transcriptome. In addition, several approaches such as expression profiling using microarrays, as described in Chapters 21–23, serial analysis of gene expression (SAGE), and massively parallel signature sequencing (MPSS), as described in Chapter 22, have also been used for the characterization of transcriptomes. ESTs are single pass sequences of random cDNA clones. They are partially complementary DNA (cDNA) sequences corresponding to messenger RNAs (mRNA) generated from randomly selected cDNA library clones. Therefore, knowledge and skill in cDNA library construction, library normalization, and analysis of generated EST sequences are essential for a good understanding of the transcriptome. In this chapter, I intend to provide information on construction of normalized cDNA libraries for efficient generation and analysis of ESTs, and the significance of EST analysis for the genome research of species lacking a sequenced genome. A successful EST project can provide the sequence tools needed for not only gene discovery, but also for physical, linkage, and comparative mapping, analysis of alternative splicing and gene duplication, and microarray development.

Construction of Normalized cDNA Libraries for Efficient EST Analysis

EST analysis has traditionally been conducted by sequencing random cDNA clones from cDNA libraries. Such an approach is efficient at initial stages of gene discovery, but has proven to be inefficient in the gene discovery of rarely expressed genes. Theoretically, a typical fish or shellfish species expresses perhaps no more than 30,000

Table 20.1. The top 100 expressed genes account for 33.9% of the transcriptome of the channel catfish head kidney.

No. of Genes	Clones	% sequenced
10	100	9.4
20	302	14.3
30	389	18.5
40	462	21.9
50	520	24.7
60	570	27.1
70	615	29.2
80	655	31.1
90	685	32.5
100	715	33.9

genes, while the human genome contains only some 25,000 genes. At initial stages of EST analysis, the gene discovery rate is almost linear to the EST sequences generated. Thus, it appears, at first glance, that even a small laboratory could complete gene discovery of the entire transcriptome of a species in less than a year. For instance, if one can sequence 100 ESTs a day, it only takes 300 days to complete 30,000 ESTs. However, the rate of gene discovery usually drops precipitously soon after reaching a level of several hundred ESTs. This phenomenon can be easily explained when one considers that several hundred of the most abundantly expressed genes make up the vast majority of the mRNA mass in cells. The chances of encountering cDNAs representing rarely expressed genes, therefore, are small without normalizing the cDNA library. For example, only 100 genes accounted for 33.9% of the transcriptome in the head kidney of channel catfish (Table 20.1). Only 23 genes, sequenced five or more times, accounted for 250 clones of the 1,093 clones (22.8%) sequenced from the catfish liver (Figure 20.1). It is estimated that levels of mRNA can range from 200,000 copies to 1 or fewer copies per cell (Galau et al. 1977). By using regular cDNA libraries, the most abundantly expressed genes would have been sequenced 200,000 times before the most rarely expressed genes are sequenced just once. Clearly, EST sequencing from non-normalized libraries is inefficient for gene discoveries of rarely expressed genes. Normalization decreases the prevalence of clones representing abundant transcripts and dramatically increases the efficiency of random sequencing and rare gene discovery.

Normalized cDNA libraries are cDNA libraries that have been equalized in representation to reduce the representation of abundantly expressed genes and to increase the representation of rarely expressed genes. This concept was initially proposed by Dr. Soares and others (1994) and further modified by Bonaldo and others (1996). In the original protocol, partial extension products of the cDNA library were used as drivers. Single-stranded (SS) libraries were prepared. The SS-library was then extended from poly A for 20–200 bases. The partially extended products were used to hybridize to the cDNA. After hybridization, the double-stranded (DS) fraction is removed by using hydroxyapatite columns, and the remaining SS fraction is converted to DS cDNA and is included in the normalized cDNA library (Soares et al. 1994). The protocol was then

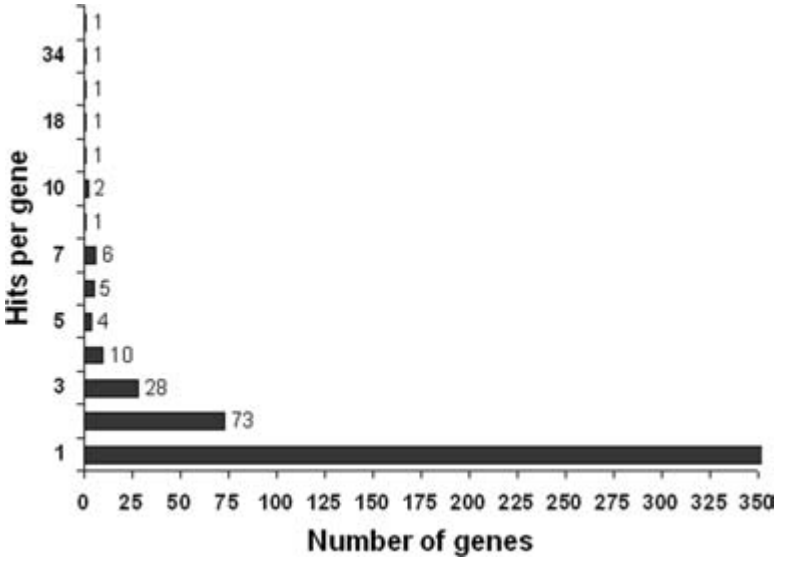


Figure 20.1. Expression profiles and sequencing redundancy among known genes in the analysis of ESTs from the channel catfish liver tissue, adopted from Liu (2006).

modified to use polymerase chain reaction (PCR) amplified cDNA inserts as drivers (Bonaldo et al. 1996). In some other cases, mRNA has also been used as drivers. In these cases, the corresponding mRNAs used for cDNA synthesis were biotinylated with photobiotin and used in excess for hybridization with the cDNA. The hybridization was then treated with streptavidin, followed by phenol extraction to remove the cDNAs that hybridized to the biotinylated mRNA. The cDNA-mRNA hybrids and mRNA are removed by phenol extractions. Only the unhybridized cDNA remain in the aqueous phase for use in construction of normalized libraries. While the details of how the subtraction is conducted may differ greatly, the basic principles behind normalization are the same (i.e., they all depend on the faster hybridization kinetics of abundantly expressed genes to form DS complexes that can be removed by various means, whereas it takes a long time for the rarely expressed genes to reassociate).

We have used the Creator™ SMART™ cDNA Library Construction Kit for the synthesis of cDNA (Zhu et al. 2001). Use of this kit is purely based on the experience of the author whose intention is not to provide a thorough list of methodologies, but to provide a procedure with personal experience. (For alternative methods, readers are referred to Chapter 22.) In our experience, this kit provides high-quality, full-length, directionally cloned cDNA libraries from nanograms of total or poly A^+ RNA. This system has two unique characteristics. The first is provided by the SMART (Switching Mechanism At 5′ end of RNA Transcript) that offers the ability to synthesize full-length cDNA (Figure 20.2). Most commonly used cDNA synthesis methods rely on the ability of reverse transcriptase (RT) to transcribe mRNA into SS DNA in the first strand reaction. In some cases, RT terminates before transcribing the complete mRNA sequence. This is particularly true for long mRNAs, especially if the first strand synthesis is primed with oligo(dT) primers only or if the mRNA contains abundant secondary

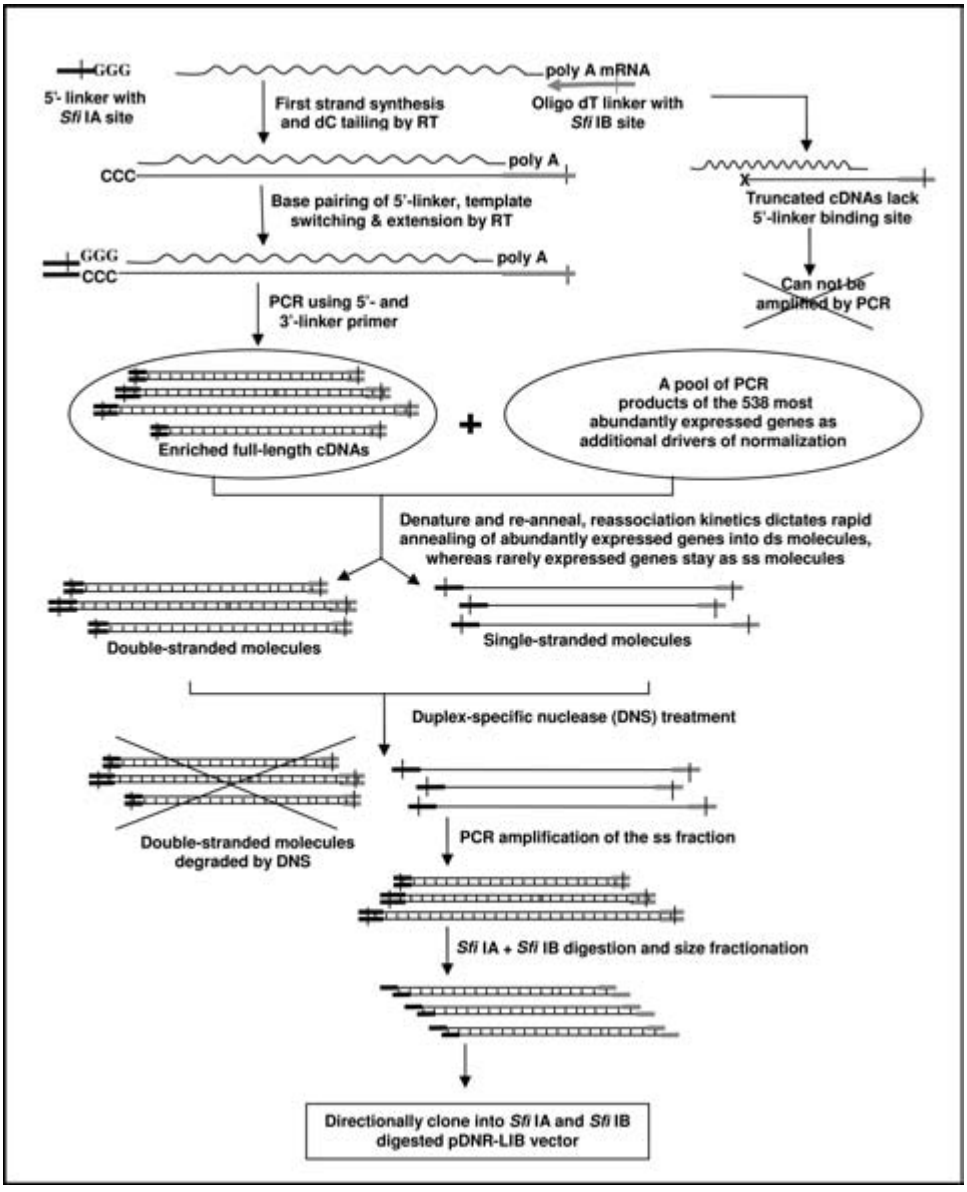


Figure 20.2. Schematic presentation of the procedures for the construction of normalized cDNA libraries, adopted from Liu (2006).

structures. The SMART system is designed to preferentially enrich for full-length cDNAs, while eliminating adaptor ligation. The mechanism for the enrichment of full-length cDNA is the use of a 5′-linker with 3′-GGG tails. RT has terminal transferase activity that preferentially adds three additional Cs at the end of first-strand cDNA. As a result, the first strand cDNA is able to base pair with the 5′-linker with 3′-GGG tails. Once base paired, the RT would switch the template and extend into the linker

sequences allowing PCR amplification of full-length cDNA. Truncated cDNAs are not able to base pair with the 5′-linker, and therefore, get lost in the PCR amplification of the full-length cDNA. The second feature is provided by the Creator system. The Creator system allows the transfer of a target gene from a single donor vector directly into multiple acceptor expression vectors using Cre-loxP recombination. Using this method, any gene cloned into a specialized cloning vector (such as pDNR-LIB) can be transferred into any acceptor vector for functional analysis without the need for subcloning. Since this feature is less relevant to EST analysis, interested readers are referred to the Web site of BD Biosciences (http://www.clontech.com/clontech/techinfo/manuals/PDF/PT3577-1.pdf).

cDNA Synthesis

Depending on the availability of biological samples, different protocols should be adopted. If biological samples are limiting (e.g., 50 nanograms [ng] of total RNA), cDNA synthesis can be accomplished by long-distance PCR. In this protocol, a modified oligo(dT) primer (CDS III/3′ PCR Primer) primes the first-strand synthesis reaction, and the 5′-linker serves as a short, extended template at the 5′ end of the mRNA (Figure 20.2). When the RT reaches the 5′ end, the enzyme's terminal transferase activity adds a few additional nucleotides, primarily deoxycytidine, to the 3′ end of the cDNA. The 5′-linker, which has an oligo (G) sequence at its 3′ end, base pairs with the C stretch, creating an extended template. RT then switches templates and continues DNA synthesis to the end of the 5′-linker. The resulting full-length ss cDNA contains the complete 5′ end of the mRNA, as well as the sequence complementary to the 5′-linker. Now long distance PCR is used to amplify the cDNAs using the 3′ PCR primer and the 5′-primer. Only those SS cDNAs having a 5′-linker sequence at the 5′ end and the 3′-linker sequence at the 3′ end can serve as templates for PCR. Incomplete cDNAs lacking the linker sequences will not be amplified, thus allowing construction of cDNA libraries with a high percentage of full-length cDNAs. If biological samples are not limiting (e.g., 1 μg or more poly A^+ RNA is available), after the first cDNA is made, direct primer extension is conducted using the 5′-linker to generate the second-strand cDNA.

The quality of mRNA or total RNA (in case of small amounts of starting material) is the key to the construction of high quality cDNA libraries or the normalized libraries thereafter. We generally check the quality of RNA by running a gel. Typically, the 28S and 18S RNA should form tight bands and the mRNA appears as a smear (Figure 20.3). The ratio of the two bands should be roughly 2:1 (28S:18S). If the prominence of the 28S RNA is decreased, it is a reflection of partially degraded RNA.

Normalization and Subtraction

Several strategies have been developed for the normalization of cDNA libraries. The fundamental principles behind all the normalization procedures are the same, and they all depend on the differential hybridization of abundant molecules over rare molecules. We have used a strategy using the Evrogen TRIMMER DIRECT Kit (http://www.evrogen.com/p3_2.shtml). This system is specially developed to normalize

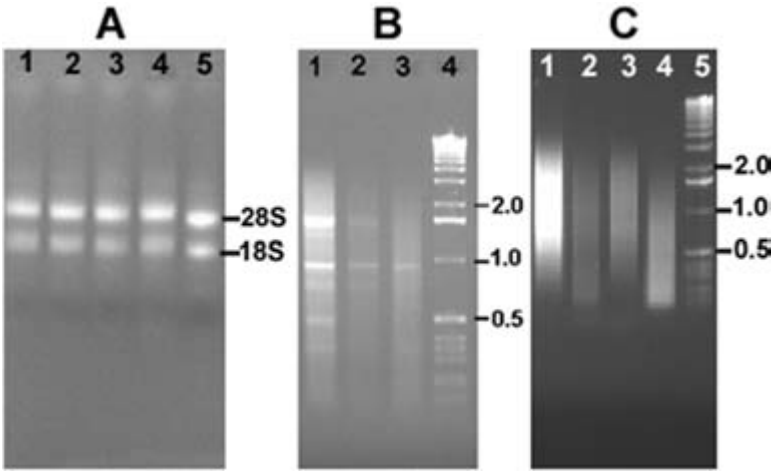


Figure 20.3. Quality checking for RNA (A), cDNA synthesis (B), and normalized cDNA (C). Lanes 1–5 of (A) were RNA from head kidney, liver, skin, muscle, and gill; lane 1–3 of (B) were cDNAs made from muscle, liver, and head kidney. Note the prominent bands over cDNA smear. Lane 4 of (B) is 1 kb molecular weight standard; Lanes 1–4 of (C) were normalized cDNAs from head kidney, liver, and gill with lane 5 of (C) as 1 kb molecular weight standard. Figure was adopted from Liu (2006).

cDNA enriched with full-length sequences (Zhulidov et al. 2004). The method involves denaturation-reassociation of cDNA, degradation of DS-fraction formed by abundant transcripts, and PCR amplification of the equalized SS-DNA fraction. The key element of this method is degradation of the DS-fraction formed during reassociation of cDNA using Duplex-Specific Nuclease (DSN) enzyme (Shagin et al. 2002). A number of specific features of DSN make it ideal for removing DS-DNA from complex mixtures of nucleic acids. DSN displays a strong preference for cleaving DS-DNA in both DNA-DNA and DNA-RNA hybrids, compared to SS-DNA and RNA, regardless of the sequence length. Moreover, the enzyme remains stable over a wide range of temperatures and displays optimal activity at 55–65°C. Consequently, degradation of the DS DNA-containing fraction by this enzyme occurs at elevated temperatures, thereby decreasing loss of transcripts due to the formation of secondary structures and nonspecific hybridization involving adapter sequences.

In addition to the normalization procedures, we have also used additional drivers for subtraction. The subtraction drivers were determined from our previous EST analysis of 40,000 ESTs of catfish. Cluster analysis of the catfish ESTs allowed the identification of abundantly expressed genes. We decided to subtract any message whose representation was more than 2 out of 10,000. In other words, if the transcript was found more than 8 times in the 40,000 ESTs, they were defined as abundantly expressed. This led to the identification of 538 genes for subtraction. The inserts of the EST clones containing these 538 genes were amplified and used as additional drivers for subtraction.

The remaining procedures for the construction of normalized cDNA libraries are straightforward. However, in our experience with the CloneTech's vector pDNR-Lib, a control of "vector only ligation" appeared to be important. The reason is that the

cloning vector is prepared by isolation of the vector backbone after restriction digestion using *Sfi* I endonuclease. In the vector, two *Sfi* I sites exist with different sequences that are referred to as *Sfi* IA and *Sfi* IB. If the vector is digested at only one site, then the vector self ligation can lead to the very high rate of insertless clones. As the quality of the vector preparation may vary from batch to batch, a control ligation of vector DNA only should reveal if the background is a problem immediately.

Significance of EST Analysis, Its By-products, and Applications of EST Resources

The significance of EST analysis has been recognized ever since the first EST analysis experiment was conducted (Adams et al. 1991). However, most scientists who are not familiar with EST analysis may only partially recognize its value. In this chapter, I will provide information concerning objectives of genome research that can be reached by EST analysis.

EST Analysis Is One of the Most Rapid Methods for Gene Discovery and Identification

EST analysis is one of the most rapid approaches for gene discovery. A small collection of ESTs in a species without any genome information can result in the rapid identification of a large number of genes. Gene discovery and identification is, therefore, the primary function of EST analysis.

Back in the 1980s, to clone a gene or a cDNA was very difficult. Ph.D. students were put to the hard task of gene cloning for 3 to 6 months to clone a cDNA. In most cases, the cloned cDNAs were quite highly expressed. Adams and others (1991) put forth a procedure defined as the analysis of ESTs that technically involves direct sequencing of random cDNA clones. However, the outcome was extremely significant. Rather than conducting cDNA cloning through traditional screening, rescreening, purification, and sequencing, direct sequencing of cDNA clones allows rapid gene discovery that often times includes the cDNAs for the genes of interest. At the same time, many other cDNAs are identified that are also of high scientific interest. For instance, during our EST work back in 1997 using manual sequencing, analysis of just 100 clones from the pituitary cDNA library led to the identification of growth hormone, gonadotropins, prolactin, and proopiomelanocortin cDNAs, all of which were of great interest, and the cloning of which would have otherwise taken us a much longer period of time and greater effort and resources (Karsi et al. 1998). Sequencing of 2,228 EST clones from the head kidney tissue allowed identification of 753 distinct known genes plus 739 unknown gene clones (Cao et al. 2001). Sequencing of 1,201 clones from the brain led to the identification of 330 known genes plus 330 unknown genes (Ju et al. 2000). In 2001, the catfish EST collection reached a historical 10,000 clones that allowed the identification of 5,905 genes. These ESTs were the basis for the first aquaculture species to be listed under the TIGR Gene Index in 2002. As of 2005, our catfish EST collection reached 44,000 that represented 25,000 unique gene

sequences. Clearly, it is the high gene discovery rate of EST analysis that has allowed the identification of such a large number of genes. To the best of my knowledge, there is no other method that can provide an equal gene discovery rate as EST analysis while also providing long-term genome resources for many other applications.

Because of the exceptionally high gene discovery rate of the EST approach, EST analysis has been extremely popular. The EST database dbEST has been one of the fastest growing databases at the National Center for Biotechnology Information (NCBI). As of January 20, 2006, there are 32,889,225 entries in the NCBI's public EST database dbEST (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html).

ESTs Provide a First Glance at Gene Expression Profiles

EST sequencing from non-normalized cDNA libraries should reveal the true representation of the mRNAs in the tissues or cells from which the cDNA libraries were made. Large-scale EST analysis is, therefore, a direct way to conduct expression profiling (Franco et al. 1995, Azam et al. 1996, Lee et al. 2000). It offers a rapid and valuable first look at genes expressed in specific tissue types, under specific physiological conditions, or during specific developmental stages (Ju et al. 2002, Cao et al. 2001, Karsi et al. 2002, Kocabas et al. 2002). To demonstrate this point, once again, let me use the example of our EST work from the pituitary of catfish. Sequencing of just 100 EST clones revealed that the majority of clones sequenced from the pituitary were in the category of hormones (Karsi et al. 1998). As we all know, the pituitary is the organ whose main function is production of hormones involved in various physiological regulations. When large numbers of ESTs are sequenced, a relatively accurate picture can be obtained for gene expression profiling, as has been demonstrated in various species (Kurobe et al. 2005, Song et al. 2004, Kimura et al. 2004, Lo et al. 2003).

When EST analysis is conducted using cDNA libraries constructed from various tissues, the generated ESTs can provide a comprehensive comparison of gene expression in different tissues. In most recent cases, such tissue expression profiling has been mostly accomplished by microarray analysis. However, direct sequencing of ESTs from tissue libraries can provide very similar information. In addition, EST analysis using cDNA libraries from various tissues provides a greater control to investigators as to how deep each library is sequenced. Alternatively, some investigators prefer to attach a molecular tag on each of the tissue cDNAs by using different adaptors attached to the oligo dT primers during the first-strand cDNA synthesis. Such molecular tags allow tracking of the cDNAs for the tissues from which they were derived. This value of EST analysis is undoubtedly signified by the greater power of emerging sequencing technologies (see Chapter 25).

ESTs Provide a Robust Approach for the Study of Alternative Splicing and Differential Polyadenylation

ESTs provide a good deal of information about alternatively spliced and polyadenylated transcripts. Although the number of genes now seems to be quite smaller than

we once thought, the number of distinct transcripts can be much larger. As summarized in a recently published special issue of *Science,* the total number of distinct transcripts can be one order of magnitude larger than the number of genes. (A series of reviews on RNA can be found in the September 2, 2005 issue of *Science,* Vol. 309, No. 5740). Alternative splicing and differential polyadenylation are probably widespread. Different transcripts probably exist in nature for many, if not most, genes. However, using traditional molecular biology approaches, one would have no way of knowing what other transcripts are expressed in the cells, other than the few that have been accidentally identified. In contrast, a wide variety of transcripts are sequenced in EST projects. If primary cDNA libraries are used, the chances of finding various transcripts are proportional to the representation of the transcripts in the mRNA pool. However, if normalized libraries are used, the chances of finding rare types of transcripts are greatly increased.

EST Analysis Is the Most Efficient Way for the Identification of Type I Polymorphic Markers

ESTs provide great opportunities for the identification of Type I polymorphic markers. Markers can be divided into Type I markers and Type II markers. Type I markers are markers associated with genes of known functions, whereas Type II markers are markers developed from anonymous genomic regions. Obviously, when polymorphism is the major interest, sequence variations within gene coding regions are lower as restricted by functional constraints, and thus development of Type I markers is relatively more difficult. However, Type I markers are of greater value because genes are highly conserved across a wide spectrum of evolution. In addition to their value for linkage mapping, Type I markers are also useful for comparative mapping. EST analysis can provide two types of Type I markers: single nucleotide polymorphism within transcripts, and microsatellites associated with ESTs. As a matter of fact, these approaches to developing Type I markers are among the most effective approaches. For instance, a single nucleotide polymorphisms (SNP) rate of 1.32% was found during analysis of 161 genes in catfish, making comparative EST analysis one of the most efficient approaches for the identification of SNP markers within genes (He et al. 2003). Analysis of 43,033 catfish ESTs led to the identification of 4,855 EST clones containing microsatellites (11.2%). Further cluster analysis revealed that the majority (4,103) clones represent unique genes (Serapion et al. 2004). Obviously, not all microsatellites identified in EST analysis are directly useful as markers. Several situations may be encountered. First, the microsatellites may exist at the very upstream of ESTs; second, the microsatellites may exist in the downstream immediately before poly A; and third, the microsatellites may be flanked by simple sequences. In all of these three cases, the identified microsatellites cannot be directly used until further sequences are obtained by genomic sequencing. In addition, flanking sequences used for primer design may amplify introns that are not known from EST projects, making allele prediction difficult. Nonetheless, as a by-product of EST sequencing, EST-associated microsatellites are by far the most efficient way for the development of Type I microsatellite markers.

Related ESTs Are Useful for the Identification of Duplicated Genes

EST analysis provides one of the most efficient ways for the identification of duplicated genes. Gene duplication is a widespread phenomenon in fish. Despite the great debate about the origins of duplicated genes, whether through entire genome duplication followed by differential gene retention and gene loss, or through duplication that does not involve duplication of the entire genome, studies of duplicated genes are still limited by technical difficulties. Even for entirely sequenced genomes, genome assembly can even be hindered by duplicated segments. EST analysis should produce sequences of related transcripts that can be analyzed through phylogenetic approaches. For instance, if two related transcripts are found from channel catfish and one of them is more closely related to a transcript from blue catfish than related to the other transcript from channel catfish, these two transcripts are likely encoded by two duplicated genes, rather than products of allelic variation. The rationale is that allelic variation of the same species should be smaller than the variation between species. In this case, if the two transcripts are more closely related between the two species, the transcripts are likely orthologs. In contrast, if the transcripts are more distantly related, even though it is encoded by the same species, it is likely a paralog (gene duplication products in evolution). Therefore, through large-scale EST analysis followed by phylogenetic analysis, large-scale differentiation of orthologs and paralogs becomes possible.

ESTs Serve as the Basis for Comparative Mapping

The basic concept of comparative mapping is based on the assumption that the genomes of closely related species are highly conserved in the organization of their genes (see Chapter 19). Thus, ESTs can be used as a great resource for comparative mapping. First, EST resources from aquaculture species can be analyzed to determine on which chromosomes their corresponding genes reside in completely sequenced genomes such as zebrafish and *Tetraodon nigroviridis.* We have found that the *Tetraodon* genome sequence is of good quality and works well for comparative genome analysis. For instance, a large number of catfish ESTs have been identified to correspond to genes from all 21 chromosomes of *Tetraodon* (see Table 20.2 for an example). The idea is that if the genome organization is highly conserved between the catfish and *Tetraodon,* then many of the chromosome-specific ESTs should also be located on the same chromosomes in catfish. Testing of this hypothesis should be much easier than mapping blindly without any information. Second, once conserved syntenies are identified, the genes between the conserved genes can be easily verified by using the EST resources. For instance, in a recent study involving BAC-end sequencing, more than 100 mate-paired genes were identified on both ends of BAC-end sequences (Xu et al. 2006). Direct Basic Local Alignment Search Tool (BLAST) searches allowed identification of conserved syntenies with the zebrafish and *Tetraodon* genomes. Once the conserved syntenies are identified, the internal genes between the mate-paired genes on the BAC-ends can be inferred by determining the genes between these genes in the zebrafish or *Tetraodon* genomes. Once the gene identities are determined, sequencing primers or hybridization overgo probes can be designed using the EST resource. Direct BAC

Table 20.2. An example of the utility of ESTs for comparative mapping. Here, *Tetraodon* chromosome-specific catfish EST hits are shown along with their microsatellite repeats for future mapping.

Tetraodon Chromosome	*Tetraodon* Query	Catfish Hit	E-value	Repeat Location (bp)	Motif	# of repeats
1	CAG04525	BM028034	2E-20	1	gt	11
1	CAF91213	BM425454	6E-19	461	tg	8
1	CAG07953	CB936804	8E-27	527	ca	23
1	CAG07953	CB936804	8E-27	575	ca	9
1	CAG01395	CB938863	3E-74	635	tc	10
1	CAG08047	CF262362	3E-20	18	ac	14
1	CAG05652	CF971555	2E-31	429	tga	15
1	CAF93405	CK403365	1E-37	403	aata	5
1	CAG01184	CK410393	1E-73	587	gt	14
1	CAG06269	CK411617	2E-55	152	gt	16
1	CAG01390	CK412611	2E-11	108	gatg	5
1	CAG01390	CK412611	2E-11	425	gttt	5
1	CAG08382	CK414367	2E-83	833	ac	14
1	CAG11623	CK414895	2E-29	43	ag	9
1	CAG11623	CK414895	2E-29	61	aaag	5
1	CAG10993	CK415834	9E-17	616	tg	9
1	CAG09843	CK417836	4E-64	575	atg	10
1	CAG10133	CK422325	4E-15	491	at	10
1	CAF99087	CV987962	3E-15	430	agc	6
1	CAG01185	CV988882	3E-63	574	gac	6
1	CAG01185	CV988882	3E-63	595	gac	6

sequencing or overgo hybridization should verify if the genes existing between the mate-paired genes in zebrafish or *Tetraodon* indeed also exist in the same location in catfish. Such an approach has been demonstrated to be highly efficient for comparative mapping (Xu et al. 2006).

ESTs Provide the Basis for Integration of Genetic and Physical Maps

For efficient genome research, linkage maps constructed by genotyping of a resource family or families using polymorphic DNA markers need to be integrated with the physical maps. In most cases these days, particularly with aquaculture species, physical maps are constructed using BAC-based contigs. Mapping of Type I markers derived from EST analysis on both the linkage map and the physical map would effectively integrate the two maps. Although mapping of Type I microsatellites and SNP on linkage maps is straightforward by traditional linkage mapping, mapping of the same set of Type I markers to BACs requires hybridization. Traditionally, isolation of a large number of probes from cDNA clones and hybridization would involve a large amount of work. However, use of overgo probes (Figure 20.4) can significantly reduce the workload for the isolation of purified probe fragments. In this strategy, an oligonucleotide primer can be selected in the coding region of the cDNA, and an antisense primer is

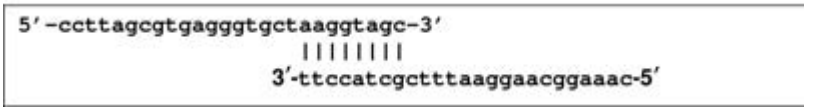


Figure 20.4. The heteroduplex formed between overgo probe primers. This structure serves as template for filling in reactions of polymerase, during which the newly synthesized probes can be labeled. Figure was adopted from Liu (2006).

then designed to partially overlap with the sense primer by 8 base pairs (bp). The primers would form a duplex upon annealing to each other, forming a structure that would be perfectly primed for "filling in" reactions by polymerase. Radioactive or fluorescent nucleotides can be used during "filling in" reactions to label the overgo probes.

ESTs Provide the Basis for the Development of Microarray Technology

Lastly, ESTs provide the material basis for the development of cDNA microarrays. cDNA microarrays are constructed by amplification of the inserts from a pool of unique ESTs. Therefore, in most cases, after EST sequencing, ESTs are analyzed through clustering analysis using various software packages. Once the cluster analysis is completed, the investigator should know how many unique genes the ESTs represent. The inserts of the distinct and unique EST clones are then amplified using PCR, and the PCR products printed on microarray slides. As a note, ESTs are also very important for the verification of gene models for completely sequenced genomes. As we all know, the Human Genome Project was initially announced to be complete in 2000. Politicians such as President Bill Clinton and Prime Minister Tony Blair declared the historical completion of the project at this date, the Human Genome Project was reannounced to be complete in 2002, and once again was announced to be truly completed in 2004. Between 2000 and 2004, the total number of genes included in the human genome was reduced from 40,000 to approximately 25,000. Many wondered what could account for the large reduction of 15,000 genes because the world's most intelligent and highest caliber scientists were involved in the mapping and assembly of the human genome. These scientists had learned that, without the hard evidence of gene products such as ESTs, it was not possible to have an accurate picture of how many genes humans encode. The 40,000 gene number came from computational predictions in 2000, 15,000 (37.5%) of which were later found to be unsupported by EST evidence and other expression data.

Criteria for Successful EST Projects

The factors one needs to consider for a successful EST project depend on the objectives of the project. However, a set of criteria should be established at the inception of an EST project and used to measure its subsequent level of success. These criteria include a high gene discovery rate, a high gene identification rate, a high capacity for the identification of gene-associated Type I markers, a high rate of full length cDNAs, the ability to differentiate orthologs and paralogs, the ability to identify alternatively processed

transcripts, and an organized inventory so that the ESTs can be used to develop cDNA-based microarrays or be distributed to other researchers upon request.

Before the start of an EST project, the researchers should evaluate the entire purpose of the project in relation to their genome program. For instance, if interspecific hybrid systems are used for linkage mapping, then development of ESTs from both species in parallel may be of interest because many Type I markers developed from the EST project should be directly useful for mapping, as we have done with channel catfish and blue catfish (He et al. 2003).

The number of individuals used for construction of cDNA libraries is important for the identification of SNPs within the cDNAs. For the most part, the majority of aquaculture species are either diploid organisms or tetraploid organisms. Therefore, there should be some level of allelic variation in cDNA sequences. However, including multiple individuals in the library should increase the possibility of SNP identification. In the construction of the catfish cDNA libraries, we used 10 individuals that were required for the collection of sufficient biological samples, but also for the consideration of SNP discoveries.

If EST analysis is a part of broader transcriptome analysis, inclusion of biological samples containing various genetic backgrounds, tissues, developmental stages, and physiological conditions should be considered. For instance, if disease defense genes are of interest, it is best to include tissue samples after infection because some of the defense-related genes may be expressed only after infection. Inclusion of various tissues should allow the capture of various tissue-specific transcripts in sequenced ESTs. Many genes are developmental-stage specific. For instance, genes expressed at early stages of life can only be captured by using embryonic samples at various stages. Similarly, certain genes are expressed only in certain physiological conditions. For instance, genes related to reproductive processes likely are expressed when the animals are sexually mature.

The most important factor for a successful EST project is the quality of cDNA libraries. Of the important criteria for EST projects, most of them are correlated with the quality of cDNA libraries, for example, gene discovery rate, gene identification rate, percentage of known genes with complete open reading frames, etc. The most significant consideration is the quality of starting RNA samples, the production of full-length cDNAs, and efficient normalization.

Several quality standards pertain to EST sequencing or postsequencing analysis. Of these, important factors to consider include average EST length, quality scores of ESTs, usefulness and accessibility of ESTs, timely GenBank submission, and proper level of annotation. EST length and sequence quality is often related to plasmid quality. Standard plasmid kits such as those from Qiagen provide high quality plasmid DNA that allows long, high quality sequencing reads.

After completion of EST sequencing, proper annotation is crucial to make the ESTs useful. If the ESTs are properly annotated, they will be searchable in the public databases. Simply "dumping" the raw EST sequences into public databases makes the EST sequences much less useful not only for other researchers, but also for the "dumpers" themselves. Finally, a good clone inventory is required for successful EST projects because the investigators need to be able to provide any clone upon request. Scientists are excellent in discovery and innovation, but are relatively poor when it comes to inventory. A telling comparison can be made between a scientist's ability to find proper inventories of EST clones with that of a supermarket to find retail items in the warehouse. It

likely takes far less time for Wal-Mart workers to find any of the tens of thousands of items in their warehouse than for a scientist to find a cDNA clone from his/her freezer. ESTs are valuable long-term resources that require careful inventory practices to ensure their availability in the future. In fact, EST resources often grow in value to the researcher over time. For example, as the genome resources of catfish have grown, ESTs initially used for simple gene discovery and expression analysis are now useful in marker development, comparative mapping, and gene duplication studies.

Acknowledgments

Our catfish genome research is currently supported by a grant from the USDA NRI Animal Genome Tools and Resources Program (Award No. 2006-35616-16685). We are grateful for an equipment grant from the National Research Initiative Competitive Grant No. 2005-35206-15274 from the USDA Cooperative State Research, Education, and Extension Service. I appreciated critical reading of the manuscript by Eric Peatman. The author is grateful to his staff and hard working graduate students.

References

Adams MD, JM Kelley, JD Gocayne, M Dubnick, MH Polymeropoulos, H Xiao, CR Merril, A Wu, B Olde, RF Moreno, AR Kerlavage, WR McCombie, and JC Venter. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. Science, 252, pp. 1651–1656.

Azam A, J Paul, D Sehgal, J Prasad, S Bhattacharya, and A Bhattacharya. 1996. Identification of novel genes from *Entamoeba histolytica* by expressed sequence tag analysis. Gene, 181(1–2), pp. 113–116.

Bonaldo MF, G Lennon, and MB Soares. 1996. Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res, 6(9), pp. 791–806.

Cao D, A Kocabas, Z Ju, A Karsi, P Li, A Patterson, and Z Liu. 2001. Transcriptome of channel catfish (*Ictalurus punctatus*): initial analysis of genes and expression profiles of the head kidney. Anim Genet, 32(4), pp. 169–188.

Franco GR, MD Adams, MB Soares, AJ Simpson, JC Venter, and SD Pena. 1995. Identification of new *Schistosoma mansoni* genes by the EST strategy using a directional cDNA library. Gene, 152(2), pp. 141–147.

Galau GA, WH Klein, RJ Britten, and EH Davidson. 1977. Significance of rare mRNA sequences in liver. Arch Biochem Biophys, 179, pp. 584–599.

He C, L Chen, M Simmons, P Li, S Kim, and ZL Liu. 2003. Putative SNP discovery in interspecific hybrids of catfish by comparative EST analysis. Anim Genet, 34, pp. 445–448.

Ju Z, A Karsi, A Kocabas, A Patterson, P Li, D Cao, R Dunham, and Z Liu. 2002. Transcriptome analysis of channel catfish (*Ictalurus punctatus*): genes and expression profile from the brain. Gene, 261(2), pp. 373–382.

Karsi A, D Cao, P Li, A Patterson, A Kocabas, J Feng, Z Ju, KD Mickett, and Z Liu. 2002. Transcriptome analysis of channel catfish (*Ictalurus punctatus*): initial analysis of gene expression and microsatellite-containing cDNAs in the skin. Gene, 285(1–2), pp. 157–168.

Karsi A, P Li, RA Dunham, and ZL Liu. 1998. Transcriptional activities in the pituitaries of channel catfish before and after induced ovulation by injection of carp pituitary extract as revealed by expressed sequence tag analysis. J Mol Endocrinol, 21(2), pp. 121–129.

Kimura T, T Jindo, T Narita, K Naruse, D Kobayashi, T Shin-i, T Kitagawa, T Sakaguchi, H Mitani, A Shima, Y Kohara, and H Takeda. 2004. Large-scale isolation of ESTs from medaka embryos and its application to medaka developmental genetics. Mech Dev, 121(7–8), pp. 915–932.

Kocabas AM, P Li, D Cao, A Karsi, C He, A Patterson, Z Ju, RA Dunham, and Z Liu. 2002. Expression profile of the channel catfish spleen: analysis of genes involved in immune functions. Mar Biotechnol (NY), 4, pp. 526–536.

Kurobe T, M Yasuike, T Kimura, I Hirono, and T Aoki. 2005. Expression profiling of immune-related genes from Japanese flounder *Paralichthys olivaceus* kidney cells using cDNA microarrays. Dev Comp Immunol, 29(6), pp. 515–523.

Lee EK, SB Seo, TH Kim, SK Sung, G An, CH Lee, and YJ Kim. 2000. Analysis of expressed sequence tags of *Porphyra yezoensis.* Mol Cells, 10(3), pp. 338–342.

Liu ZJ. 2006. Transcriptome characterization through the generation and analysis of expressed sequence tags: Factors to consider for a successful EST project. Israel Journal of Aquaculture-Bamidgeh, 58, 328–341.

Lo J, S Lee, M Xu, F Liu, H Ruan, A Eun, Y He, W Ma, W Wang, Z Wen, and J Peng. 2003. 15000 unique zebrafish EST clusters and their future use in microarray for profiling gene expression patterns during embryogenesis. Genome Res, 13, pp. 455–466.

Serapion J, H Kucuktas, J Feng, and Z Liu. 2004. Bioinformatic mining of type I microsatellites from expressed sequence tags of channel catfish (*Ictalurus punctatus*). Mar Biotechnol (NY), 6(4), pp. 364–377.

Shagin DA, DV Rebrikov, VB Kozhemyako, IM Altshuler, AS Shcheglov, PA Zhulidov, EA Bogdanova, DB Staroverov, VA Rasskazov, and S Lukyanov. 2002. A novel method for SNP detection using a new duplex_specific nuclease from crab hepatopancreas. Genome Res, 12, pp. 1935–1942.

Soares MB, MF Bonaldo, P Jelene, L Su, L Lawton, and A Efstratiadis. 1994. Construction and characterization of a normalized cDNA library. Proc Natl Acad Sci, 91, pp. 9228–9232.

Song HD, XJ Sun, M Deng, GW Zhang, Y Zhou, XY Wu, Y Sheng, Y Chen, Z Ruan, CL Jiang, HY Fan, LI Zon, JP Kanki, TX Liu, AT Look, and Z Chen. 2004. Hematopoietic gene expression profile in zebrafish kidney marrow. Proc Natl Acad Sci, 101, pp. 16240–16245.

Xu P, S Wang, L Liu, E Peatman, B Somridhivej, J Thimmapuram, G Gong, and Z Liu. 2006. Generation of channel catfish BAC end sequences for marker development and assessment of syntenic conservation with fish model species. Anim Genet, 37, pp. 321–326.

Zhu YY, EM Machleder, A Chenchik, R Li, and PD Siebert. 2001. Reverse transcriptase template switching: a SMART approach for full length cDNA library construction. Biotechniques, 30, pp. 892–897.

Zhulidov PA, EA Bogdanova, AS Shcheglov, LL Vagner, GL Khaspekov, VB Kozhemyako, MV Matz, E Meleshkevitch, LL Moroz, SA Lukyanov, and DA Shagin. 2004. Simple cDNA normalization using kamchatka crab duplex-specific nuclease. Nucleic Acid Res, 32, p. e37.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 21
Microarray Fundamentals: Basic Principles and Application in Aquaculture

Eric Peatman and Zhanjiang Liu

Researchers have long harnessed the basic molecular principle of nucleic acid hybridization to study the expression patterns of cell transcripts. Transcript studies allow a valuable assessment of the genetic response to environmental changes (i.e., infection, temperature, feeding rates). Incremental progress over the last 2 decades has been made from radioactively labeled probing of one gene to tens of genes to nylon-filter-based macroarrays containing hundreds of genes. In early years, progress in transcript detection techniques largely corresponded to strides in gene sequencing and discovery. However, as gene sequencing grew exponentially in the early 1990s and genomic approaches revolutionized molecular biology, a similarly radical leap forward was needed to bring transcript studies into the "–omics" era. Microarray technology provided such a boost by combining simple nucleic acid hybridization with high-density spotting robots, fluorescence-based signal detection, high-resolution laser scanners, and bioinformatic tools, allowing simultaneous expression analysis of thousands of genes (Schena et al. 1995). *In situ* oligonucleotide synthesis through photolithography (Fodor et al. 1991) as developed by Affymetrix, offered even greater gene densities than spotted arrays, albeit at a much higher cost. A decade of refinements of both spotted and *in situ* microarray technologies have resulted in further capacity increases and widened array applications without altering the fundamentals of either approach. Microarray technology is now widely accessible in biomedical and agricultural genetics research. Only within the last several years, however, have researchers in aquaculture species generated sufficient expressed sequence tags (EST) to justify using transcriptomic approaches for expression analysis. Furthermore, the high cost of microarray technology has hindered its implementation in underfunded species groups. Of the aquaculture species cultured most prevalently around the world, the majority still lacks microarray resources. A current review of microarray technology, therefore, from an aquaculture perspective should be helpful for the many species groups only now beginning to harness the potential of genomic approaches to research. In this chapter, we will review the basic principles of microarrays, present and compare the two prevalent array technologies, and discuss important factors to consider before beginning microarray research. Because of the complicated and ever-changing nature of microarray research, it is beyond the scope of this chapter to provide detailed protocols or exhaustive lists of microarray techniques and tools. Readers looking for more technical details after reading this chapter are referred to Chapter 22 concerning applications of microarrays in aquaculture and to the cited references of this chapter and Chapter 22.

Principles of Microarrays

Although microarrays use several recent technological innovations, they are, at their core, simply a high-density dot blot. In both spotted microarrays and *in situ* arrays, DNA is anchored or spotted onto a surface and then probed with labeled molecules. Hybridization and subsequent signal detection depends on the presence of complementary nucleotide sequences between the probes and the spotted sample. Microarrays achieve higher gene feature densities and, therefore, greater power for expression analysis by applying new tools to this old process. High-density spotting robots and photolithography allow each feature to be placed accurately within nanometers of the next feature on a glass slide, clearly an impossible task with the human hand. Furthermore, fluorescence-based probe labeling provides a cleaner and clearer signal than the radiation traditionally used in blotting. Finally, laser scanners facilitate the resolution of such tremendous feature densities and provide accurate fluorescent signal quantification. With an understanding of the basic principles of microarrays, we can now examine the two prevalent microarray platforms: the spotted arrays and the *in situ* arrays.

Spotted Array Design and Construction

There are two primary approaches to microarrays, differing in both their construction and their sample labeling. Spotted arrays are constructed by spotting long oligos or complementary DNAs (cDNA) using a printing robot, whereas *in situ* arrays are constructed by synthesizing short oligos directly onto the slide by photolithography. The terms spotted array and *in situ* array will be used throughout this chapter to refer to these two approaches.

Spotted array technology encapsulates the printing of either PCR products or long oligos (60–70 mers). Traditionally referred to as cDNA arrays, spotted arrays are today just as likely to be long oligos, as the cost of synthesizing oligos continues to decline, and because the parallel polymerase chain reaction (PCR) required to prepare for cDNA arrays is labor-intensive, costly, and requires having clones on hand. While these cDNA-associated difficulties can be overcome through hard work and collaboration among members of a species group, the printing of long oligos offers advantages in start-up time, the purity of commercial oligo synthesis, easier clone tracking, and the ability to use all available sequences in public genetic databases for array construction. Regardless of whether cDNA or synthesized oligos are used, most steps in array construction are similar for all spotted arrays (Figure 21.1).

Many research groups have combined EST sequencing projects with microarray construction, a logical, time-saving approach given that many of the steps of EST analysis are also needed for probe selection. Clustering of ESTs by sequence similarity allows an assessment of the number of unique sequences available for array construction and, in the case of repeated sequences, allows one to choose the best clone for the array. A unique gene (Unigene) list is a starting place for picking the sequences that will constitute the microarray. Depending on array design and experimental goals, clones can be picked to bias them toward the 3′ end, ensure complete inserts, maximize unique sequence stretches, and/or maximize genes (or features) included

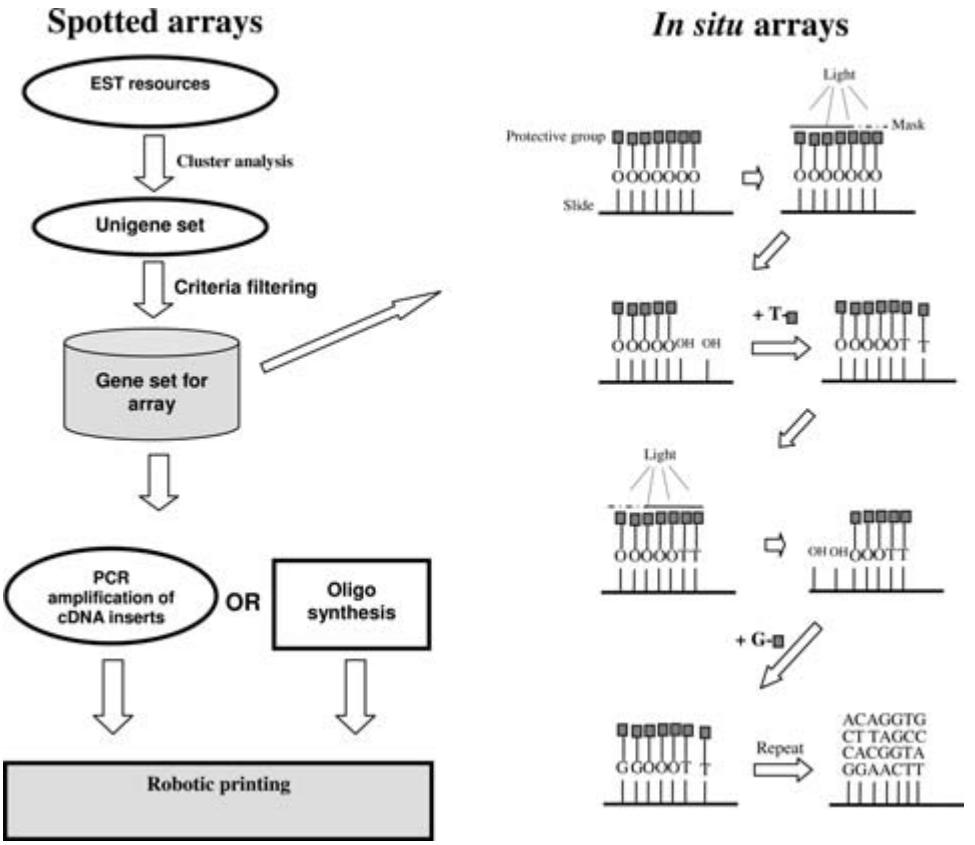


Figure 21.1. Microarray design and construction for spotted arrays and *in situ* arrays. Array design, deciding which genes to include on the array, is similar for both platforms. Construction of the physical microarray differs between robotic printing (spotted arrays) and photolithography (*in situ*). For photolithography, oligonucleotides are synthesized directly on the surface of the array, one base at a time. Unique physical lithographic masks are created for each array design, to either block or allow light to reach the slide. In the places the mask does not cover, light deprotects, or converts the protective group to a hydroxyl group, allowing the binding of single oligo at that specific site by its phosphate group. This oligo also bears a protective group that must be deprotected before an additional oligo can be coupled to it. Through repeated cycles of deprotection and coupling, 25-mer oligos are synthesized directly on the slide.

on the slide. Good examples of spotted microarray design considerations can be found in the literature, and the specific approach chosen by each group may differ. Interested readers are referred to Chapter 22, and to Whitfield and others' (2002) EST sequencing and microarray research on the honey bee using spotted cDNAs; Rise and others (2004a) and von Schalburg and others (2005a) describe considerations taken in construction of salmonid spotted cDNA arrays; and Zhao and others (2005) report validation of a porcine spotted oligo array. Operon Biotechnologies (http:// www.operon.com/) is a leading provider of sets of synthetic oligos for microarray spotting, and their Web site provides an excellent resource for criteria used in gene selection and long oligo design. Additionally, the Institute for Genomic

Research (TIGR), well known for its EST indices, provides 70 mer oligo predictions for genes in each of its indices that have been used by some groups (Zhao et al. 2005). Researchers also should decide in the design phase their array layout, feature duplication, and controls to be spotted on the slide (Whitfield et al. 2002, Smyth et al. 2005).

The clone set selected for inclusion in the microarray must now be amplified by PCR (in the case of cDNA arrays), or probes based on their sequences synthesized (long oligos). For large array designs, this is best done using robotic handling to avoid mistakes and to simplify clone tracking throughout the process. PCR products, after purification, or long oligos are usually rearrayed into 384 well plates in preparation for printing.

A variety of microarray slides are available for printing, most are poly-L-lysine and amino silane-coated. See Hessner and others (2004) for surface-chemistry comparisons. Telechem (http://www.arrayit.com/Products/Substrates/) and Erie Scientific (http://www.eriemicroarray.com/index.aspx) are leading providers of microarray slides. The actual robotic printing of microarrays is increasingly being outsourced to large university core labs or private companies, which now have years of experience in the field. For groups that anticipate printing multiple array designs and batches and want increased printing flexibility, purchasing a spotting robot may be a good choice. Perkin Elmer (http://las.perkinelmer.com/) and Genomic Solutions (http://www.genomicsolutions.com) offer popular printing systems.

In Situ Array Design and Construction

In situ array technology relies on photolithography for microarray construction (Lipshutz et al. 1999), a technique often used in computer chip fabrication. In contrast to spotting nucleotide products on the slide surface, oligonucleotides are synthesized directly on the surface of the array, one base at a time. To achieve sufficient feature densities, unique physical lithographic masks are created for each array design, to either block or allow light to reach the slide (Figure 21.1). In the places the mask does not cover, light deprotects, or converts a special protective group to a hydroxyl group. This allows the binding of single oligo at that specific site by its phosphate group. This oligo also bears a protective group that must be deprotected before an additional oligo can be coupled to it. Through repeated cycles of deprotection and coupling, 25-mer oligos are synthesized directly on the slide at densities currently as high as 1.3 million features per array. Affymetrix (http://www.affymetrix.com/) is recognized as the developer and industry leader for *in situ* arrays. Although their technology made genome-wide arrays a reality for model species and continues to expand the horizons of microarray research in biomedical fields, the technology has been prohibitively expensive for the smaller species groups including aquaculture species. Nimblegen Systems (http://www.nimblegen.com) has recently developed a "maskless" version of the Affymetrix technology that uses digital mirrors to achieve the same effect (Nuwaysir et al. 2002) at a significantly lower start-up cost, making *in situ* arrays now a feasible choice for aquaculture genomic research.

The majority of array design considerations for *in situ* arrays overlap with those of spotted arrays. EST analysis, clustering, quality control, and probe selection are still necessary steps to arrive at the set of genes that will be synthesized on the array. The higher feature density allowed with *in situ* arrays means that more genes, duplicates,

and/or controls may be included on the array, if desired. Because the per array cost is significantly higher for *in situ* arrays and project flexibility considerably less than for spotted arrays, researchers usually attempt to maximize the information that can be gained from each slide. Usually, desired sequences for the array are sent electronically to the company, which then carries out oligo probe selection (23–25 mers) and designs the array layout. Both Affymetrix and Nimblegen use a perfect match (PM) and mismatch (MM) system that accounts for the majority of the features on *in situ* arrays. Mismatch probes, as their name suggests, contain one or more mismatched nucleotides in the PM probe sequence and are used to detect and screen out false background fluorescence resulting from nonspecific cross-hybridization. Commonly, 10 PM and 10 MM probes are synthesized for each gene included on the array, and are believed to significantly increase the accuracy and sensitivity of gene expression detection. See Chen and others (2005), Han and others (2004), and Irizarry and others (2003) for more information on PM and MM probe theory.

Experimental Sample Labeling and Hybridization

Spotted and *in situ* microarrays differ not only in their array construction but also in the procedures used to label and hybridize experimental samples (probes in the traditional sense) to them (Figure 21.2). Both array platforms require that you start with RNA sources. The RNA is extracted from the samples of interest. For the sake of simplicity, we will use the example of a control sample and a treatment sample. Each RNA sample is reverse transcribed to cDNA, after quantification and quality checking by spectrophotometer measurement and agarose gel electrophoresis. From this step, differences in the procedure arise between the two microarray platforms. We will follow the spotted array protocol first, before returning to the *in situ* procedure.

The cDNA samples for spotted arrays are labeled with two different fluorescent dyes, Cy3 and Cy5, which fluoresce "green" and "red," respectively, under two different wavelengths of light (633 nanometer [nm] and 543 nm). The control sample is labeled with one dye and the treatment sample with the other. Dye assignments should be swapped in replicates to avoid dye-associated bias of hybridization (Churchill 2002). Dye labeling is most commonly done either directly or through indirect aminoallyl labeling. See Manduchi and others (2002) and Badiee and others (2003) for a comparison of labeling methods. The two labeled samples are hybridized simultaneously in equal amounts to the same array for 16–20 hours. The hybridized array, after washing to remove unhybridized probes, is scanned under a laser scanner (e.g., Molecular Devices/Axon Instruments' Axon 4000B) at both fluorescent wavelengths (or channels) for the two dyes. A digital image is acquired for both channels, and, by overlaying the two images, a fluorescent signal ratio for each array feature is obtained. This fluorescent signal ratio indicates gene expression levels. Using the Cy3/Cy5 labeling system, yellow spots indicate approximately equal levels of mRNA from both the control and treatment samples (equal signals from the green Cy3 and the red Cy5). Features that appear red or green have hybridized a majority of mRNA from only one sample. Fluorescent intensity data for each feature are recorded, and the scanned image and data can be linked back to gene feature identities through programs such as Molecular Devices/Axon Instruments' GenePix Pro software. Background subtraction and

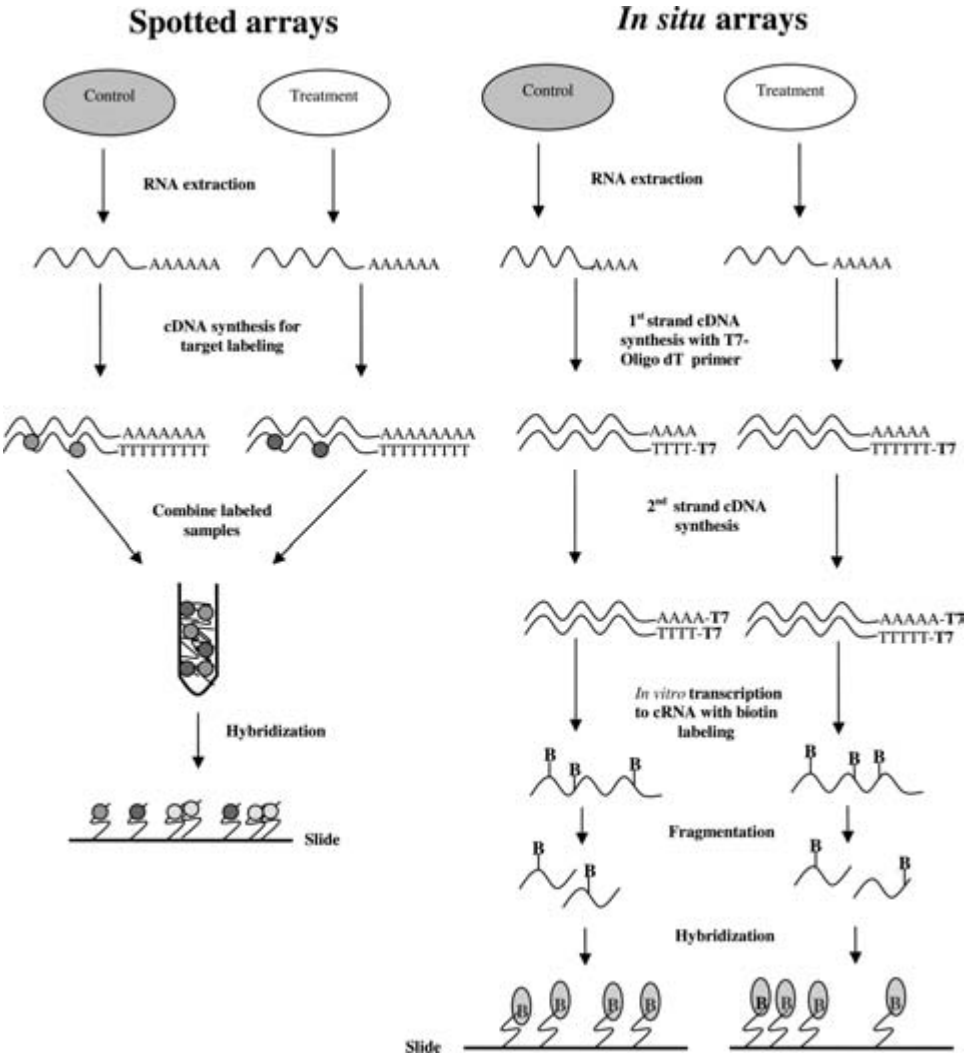


Figure 21.2. Experimental sample labeling and hybridization for spotted and *in situ* microarrays. In both cases, RNA is extracted from control and treatment samples. For spotted arrays, these samples are reverse-transcribed and labeled with two fluorescent dyes, Cy3 and Cy5, as described in the text. These two samples are then mixed equally and hybridized to a single microarray slide. Where the "green" and "red" probes hybridize to the same spot, yellow is seen. In the case of *in situ* arrays, the RNA samples are reverse transcribed using a T7 promoter oligo-dT primer. The resulting cDNA is converted to a double-stranded template by a second strand synthesis reaction, and these double-stranded cDNA samples are converted by *in vitro* transcription to biotin-labeled (B) cRNA using a T7 RNA polymerase. The cRNA from each sample is fragmented and hybridized to its own slide. Streptavidin-phycoerithryn is added as the fluorescent dye for both the control and treatment samples.

normalization are customarily carried out at this point, followed by microarray analysis and validation of genes determined to be significantly differentially expressed after treatment. For more on microarray analysis and validation for both platforms, refer to Chapter 22 in this book as well as numerous excellent research papers and reviews in

Table 21.1. A comparison of several important aspects of *in situ* and spotted array platforms. *Cost/slide can vary significantly from these figures depending on design, quantities ordered, core facility discounts, etc.

	In situ arrays	Spotted arrays
Starting material	DNA sequences	DNA sequences or cDNA
Array fabrication	*In situ* synthesis by photolithography	Robotic spotting
Features	>400,000	<50,000
Spot quality	High	Variable
Oligo length	23–25mer	Usually 70mer
Labeling	Single dye label, e.g., biotin-streptavidin-phycoerithryn	Two dye label—Cy3, Cy5
Cost/slide	>500*	<100*
Probe/slide	One	Two
Dye swapping?	No	Yes
Controls	PM/MM, ±	Duplicates, ±
Providers	Affymetrix, Nimblegen, etc.	Species groups, core facilities, biotech

the literature (e.g., Leung and Cavalieri et al. 2003, Walsh and Henderson 2004, D'Ambrosio et al. 2005).

We return now to *in situ* arrays in order to contrast the labeling and hybridization of their samples with that of spotted arrays. The original RNA samples are reverse transcribed using a T7 promoter oligo-dT primer. The resulting cDNA is converted to a double-stranded template by a second strand synthesis reaction. After purification, these double-stranded cDNA samples (again control and treatment) are converted by *in vitro* transcription to biotin-labeled cRNA using a T7 RNA polymerase. The cRNA from each sample is fragmented and hybridized to its own slide (note: no mixing of samples). Streptavidin-phycoerithryn is added as the fluorescent dye for both the control and treatment samples (Figure 21.2). To clarify, each biological sample for *in situ* arrays is hybridized to a *different* slide and labeled with a *single* dye. Differential expression is measured by comparing the fluorescent intensity measurement of a given gene on the control slide with a separate measurement for the same gene from the treatment slide. Labeling reactions and hybridizations of *in situ* arrays are commonly carried out by the array provider or core lab.

Table 21.1 provides a side-by-side comparison of spotted and *in situ* arrays, summarizing the advantages and drawbacks of each platform as reviewed above. Several groups have experimentally compared the precision and accuracy of the two platforms using the same biological samples. Their studies may prove helpful to those considering which system to implement in their own research. See Woo and others (2004), Yauk and others (2004), and Meijer and others (2005).

Microarray Research in Aquaculture Species

Microarray research has advanced dramatically in recent years in aquaculture species. However, the field is still in its infancy, and distribution of resources remains uneven.

Table 21.2. Some examples of microarray studies in aquaculture species and their pathogens.

Species	Common name	References
Cyprinus carpio	Common carp	Gracey et al. 2004
Ictalurus punctatus	Channel catfish	Ju et al. 2002
Ameiurus catus	White catfish	Kocabas et al. 2004
Paralichthys olivaceus	Japanese flounder	Kurobe et al. 2005; Byon et al. 2005, 2006
Platichthys flesus	European flounder	Williams et al. 2003
Salmo salar	Atlantic salmon	Morrison et al. 2006; Martin et al. 2006; Jordal et al. 2005; von Schalburg et al. 2005a; Aubin-Horth et al. 2005; Ewart et al. 2005; Rise et al. 2004a, 2004b
Oncorhynchus mykiss	Rainbow trout	Purcell et al. 2006; MacKenzie et al. 2006; von Schalburg et al. 2006, 2005b; Tilton et al. 2005; Krasnov et al. 2005a, 2005b, 2005c; Vornanen et al. 2005; Koskinen et al. 2004a, 2004b
Oncorhynchus keta	Chum salmon	Moriya et al. 2004
Astatotilapia burtoni	African cichlid	Renn et al. 2004
WSSV and *Penaeus* sp.	White spot syndrome virus and shrimps	Lan et al. 2006, Marks et al. 2005, Tsai et al. 2004, Dhar et al. 2003, Khadijah et al. 2003
Sparus auratus	Gilthead seabream	Sarropoulou et al. 2005
Aeromonas salmonicida	Furunculosis	Nash et al. 2006
Crassostrea sp.	Oyster	Not published yet

Concerted efforts by researchers working on salmonid species has resulted in the generation of several arrays that are now available to the general research community (see Chapter 22). These arrays have been rapidly integrated into salmonid research, as seen in Table 21.2. The largest salmonid microarray generated to date contains 16,006 cDNAs with 13,421 coming from Atlantic salmon and 2,576 from rainbow trout (von Schalburg et al. 2005a). Table 21.2 lists additional microarray studies conducted on aquaculture species or aquaculture-associated pathogens. With the exception of salmonids, other microarray studies have, for the most part, been small-scale, noncollaborative efforts. A forthcoming microarray for the oyster should also be widely distributed. To date, most published microarray studies have used PCR-amplified spotted cDNA clones to fabricate the array. However, as microarray research typically takes several years from its inception to reach publication, the recent trends toward spotted oligos and *in situ* microarrays may not be reflected in the aquaculture literature for several years. A well-designed microarray can be a valuable asset to an aquaculture species group, especially if the cost per slide can be minimized to the extent that researchers can integrate transcriptomic approaches into their already established research. Microarray studies are most successful when they are just one of several approaches used to answer biological questions. For example, salmonid researchers have implemented array technology in their study of reproductive development, toxicology, physiology, and repeat structures (von Schalburg et al. 2006, Tilton et al. 2005, Vornanen et al. 2005, Krasnov et al. 2005a).

Future Directions of Microarray Research

Due to low funding levels and a relatively small research community, aquaculture genomics stands today where the model species did almost a decade ago. In the same way, microarray research in aquaculture species is only in its infancy. Like researchers of humans and mice 10 years ago, we are currently using microarrays to accelerate gene expression analysis under varied experimental conditions, to reveal novel functions in genes, and to discover possible gene interactions and networking through cluster analysis. To find future directions for microarray research in aquaculture species, we need only to observe microarray studies in model species today. The future looks especially promising for using microarrays for single nucleotide polymorphisms (SNP) analysis and quantitative trait loci (QTL) mapping to make tangible progress toward widespread marker-assisted selection (MAS) in aquaculture (see Chapter 12, and also Walsh and Henderson (2004) and Li and Burmeister (2005) for reviews). In particular, merging positional candidate genes with expression candidate genes from microarray information may reveal QTL genes responsible for important performance traits (see Drake et al. 2006). Microarrays have, furthermore, evolved to allow studies of metabolomics and proteomics that will be important in development of fish vaccines (Cretich et al. 2006). A Veterinary Immune Reagent Network has already been established in the U.S. toward development of a set of antibodies for use in agricultural research including aquaculture (http://www.avma.org/onlnews/ javma/jun06/060615b.asp). Microarrays are also being used in livestock disease diagnostics, a use easily adapted for detection of outbreaks of aquaculture pathogens (Schmitt and Henderson 2005, Baxi et al. 2006). Much of the groundwork for practical microarray research has already been laid. It is up to the aquaculture community to exploit and adapt these advances for the advantage of their respective species.

Considerations When Starting Microarray Research

The aquaculture researcher is faced with several important considerations when starting microarray research. Decisions made in the early planning phases may critically impact research for years to come, and, therefore, should be made carefully. Three broad, interrelated areas of consideration should be addressed. First, the investigators should decide why and how they plan on using microarrays in their research. Approaches will differ based on whether the investigator wants to answer a single biological question or use microarrays in several different experiments over an extended period of time. Smaller "subset" arrays of just immune-related genes or developmental genes may be more appropriate for some research. Will the arrays be used in hypothesis-driven experiments or as hypothesis-generating experiments (candidate gene fishing), or both? Second, the investigator must assess the resources available for constructing a microarray including both financial and genomic resources. While it is simple to make a decision based on financial limitations, many questions should be answered with regard to the availability of genomic resources. Some questions to ask follow: How many ESTs are available for the species of interest, and will this number increase significantly in the near future? If a spotted cDNA approach is taken to

array fabrication, are the clones available for amplification and printing? Is sequence information publicly available to allow construction of a spotted oligo array or an *in situ* array? How many researchers are interested in an array for this species? Can a collaborative, cost-sharing agreement be reached, or will a single lab have to pay all costs? What are the current costs per slide for the different platforms? Will per array costs be low enough to use the microarray as part of regular research, or is this a "one-shot" approach? Can a previously constructed array be used or adapted from a different species, a cross-species approach (Renn et al. 2004)? Third, the investigator must give intensive consideration to experimental design; this design must be statistically rigorous and stand up to scientific scrutiny (Churchill 2002, Dobbin and Simon 2002, Simon et al. 2002, Shih et al. 2004). Some design considerations include the levels of biological replication and technical replication needed to allow vigorous statistical analysis. Furthermore, the design should maximize existing resources and the ability to answer biologically meaningful questions (Chen et al. 2004). The design should also meet the standards of the Microarray Gene Expression Data Society (MGED). Compliance with their standard Minimum Information About a Microarray Experiment (MIAME) is required for publication in most journals (http://www.mged.org/Workgroups/MIAME/miame.html). Researchers should plan beyond the initial microarray experiment to likely follow up experiments to avoid generating only a list of gene names. Adequate consideration of each of these areas before starting work will help to ensure that the constructed microarrays can be used as a powerful genomic tool for long-term research on aquaculture species.

Researchers of aquaculture species may benefit from the several points of consensus that have only recently emerged after more than a decade of debate over microarray experimental design, inference, and validation. The mountain of literature describing new methods of microarray design and analysis has continued to grow rapidly, overwhelming the ability of the average scientist to keep up. For example, the journal *Bioinformatics* alone has published more than 500 papers to date on microarray analysis, many proposing novel algorithms. Therefore, the emergence of consensus points (as outlined by Allison et al. 2006) should provide some welcome and much-needed clarity to researchers beginning microarray projects. Of greatest impact to investigators is the emerging consensus that "pooling biological samples can be useful" for microarray analysis (Allison et al. 2006). Real-world financial and RNA sample constraints often make pooling a necessity for microarray research in underfunded agricultural species, but pooling may also have the added benefit of reducing variability among arrays so that broad, global expression changes can be meaningfully assessed. However, multiple pools must be used to estimate variance for inference testing. Three technical replicates of a single RNA pool is not statistically equivalent to three distinct RNA pools each hybridized to a separate array. A second consensus point likely to influence many microarray researchers is that false-discovery rate (FDR) is a better scale with which to quantify confidence than the standard p-value (Benjamini and Hochberg 1995, Pawitan et al. 2005, Allison et al. 2006). The false-discovery rate is the expected proportion of false positives among the results declared significant and is better suited to testing the tens of thousands of individual hypotheses associated with microarrays than is a too lenient p-value or the too conservative Bonferroni correction. Inferences based on expression fold change and FDR as implemented in microarray analysis programs such as Significance Analysis of Microarrays (SAM) (Tusher et al. 2001) allow users to decide the level of acceptable

Type I error suitable for their needs when evaluating differentially expressed gene lists. By adapting these, and other, newly emerged points of consensus into their research, aquaculture researchers can avoid some of the pitfalls and controversies that have hindered past microarray projects.

Acknowledgments

Research in our laboratory is supported by grants from the USDA NRI Animal Genome and Genetic Mechanisms Program, the USDA NRI Basic Genome Reagents and Tools Program, the Mississippi-Alabama Sea Grant Consortium, the Alabama Department of Conservation, the USAID, and the BARD.

References

Allison DB, X Cui, GP Page, and M Sabripour. 2006. Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet, 7, pp. 55–65.

Aubin-Horth N, CR Landry, BH Letcher, and HA Hofmann. 2005. Alternative life histories shape brain gene expression profiles in males of the same population. Proc Biol Sci, 272, pp. 1655–1662.

Badiee A, HG Eiken, VM Steen, and R Lovlie. 2003. Evaluation of five different cDNA labeling methods for microarrays using spike controls. BMC Biotechnol, 3, p. 23.

Baxi MK, S Baxi, A Clavijo, KM Burton, and D Deregt. 2006. Microarray-based detection and typing of foot-and-mouth disease virus. Vet J, 172, pp. 473–481.

Benjamini Y and Y Hochberg. 1995. Controlling the false discovery rate—a practical and powerful approach to multiple testing. J R Stat Soc Ser B, 57, pp. 289–300.

Byon JY, T Ohira, I Hirono, and T Aoki. 2005. Use of a cDNA microarray to study immunity against viral hemorrhagic septicemia (VHS) in Japanese flounder (*Paralichthys olivaceus*) following DNA vaccination. Fish Shellfish Immunol, 18, pp. 135–147.

Byon JY, T Ohira, I Hirono, and T Aoki. 2006. Comparative immune responses in Japanese flounder, *Paralichthys olivaceus* after vaccination with viral hemorrhagic septicemia virus (VHSV) recombinant glycoprotein and DNA vaccine using a microarray analysis. Vaccine, 24, pp. 921–930.

Chen DT, JJ Chen, and SJ Soong. 2005. Probe rank approaches for gene selection in oligonucleotid arrays with a small number of replicates. Bioinformatics, 21, pp. 2861–2866.

Chen YA, DJ McKillen, S Wu, MJ Jenny, R Chapman, PS Gross, GW Warr, and JS Almeida. 2004. Optimal cDNA microarray design using expressed sequence tags for organisms with limited genomic information. BMC Bioinformatics, 5, p. 191.

Churchill GA. 2002. Fundamentals of experimental design for cDNA microarrays. Nat Genet, 32 Suppl, pp. 490–495.

Cretich M, F Damin, G Pirri, and M Chiari. 2006. Protein and peptide arrays: Recent trends and new directions. Biomol Eng, 23, pp. 77–88.

D'Ambrosio C, L Gatta, and S Bonini. 2005. The future of microarray technology: networking the genome search. Allergy, 60, pp. 1219–1226.

Dhar AK, A Dettori, MM Roux, KR Klimpel, and B Read. 2003. Identification of differentially expressed genes in shrimp (*Penaeus stylirostris*) infected with White spot syndrome virus by cDNA microarrays. Arch Virol, 148, pp. 2381–2396.

Dobbin K and R Simon. 2002. Comparison of microarray designs for class comparison and class discovery. Bioinformatics, 18, pp. 1438–1445.

Drake TA, EE Schadt, and AJ Lusis. 2006. Integrating genetic and gene expression data: application to cardiovascular and metabolic traits in mice. Mamm Genome, 17, pp. 466–479.

Ewart KV, JC Belanger, J Williams, T Karakach, S Penny, SC Tsoi, RC Richards, and SE Douglas. 2005. Identification of genes differentially expressed in Atlantic salmon (*Salmo salar*) in response to infection by *Aeromonas salmonicida* using cDNA microarray technology. Dev Comp Immunol, 29, pp. 333–347.

Fodor SP, JL Read, MC Pirrung, L Stryer, AT Lu, and D Solas. 1991. Light-directed, spatially addressable parallel chemical synthesis. Science, 251, pp. 767–773.

Gracey AY, EJ Fraser, W Li, Y Fang, RR Taylor, J Rogers, A Brass, and AR Cossins. 2004. Coping with cold: An integrative, multitissue analysis of the transcriptome of a poikilothermic vertebrate. Proc Natl Acad Sci USA, 101, pp. 16970–16975.

Han ES, Y Wu, R McCarter, JF Nelson, A Richardson, and SG Hilsenbeck. 2004. Reproducibility, sources of variability, pooling, and sample size: important considerations for the design of high-density oligonucleotide array experiments. J Gerontol A Biol Sci Med Sci, 59, pp. 306–315.

Hessner MJ, L Meyer, J Tackes, S Muheisen, and X Wang. 2004. Immobilized probe and glass surface chemistry as variables in microarray fabrication. BMC Genomics, 5, p. 53.

Irizarry RA, B Hobbs, F Collin, YD Beazer-Barclay, KJ Antonellis, U Scherf, and TP Speed. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, 4, pp. 249–264.

Jordal AE, BE Torstensen, S Tsoi, DR Tocher, SP Lall, and SE Douglas. 2005. Dietary rapeseed oil affects the expression of genes involved in hepatic lipid metabolism in Atlantic salmon (*Salmo salar* L.). J Nutr, 135, pp. 2355–2361.

Ju Z, RA Dunham, and Z Liu. 2002. Differential gene expression in the brain of channel catfish (*Ictalurus punctatus*) in response to cold acclimation. Mol Genet Genomics, 268, pp. 87–95.

Khadijah S, SY Neo, MS Hossain, LD Miller, S Mathavan, and J Kwang. 2003. Identification of white spot syndrome virus latency-related genes in specific-pathogen-free shrimps by use of a microarray. J Virol, 77, pp. 10162–10167.

Kocabas A, R Dunham, and ZJ Liu. 2004. Alterations in gene expression in the brain of white catfish (*Ameirus catus*) in response to cold acclimation. Marine Biotechnology, 6, pp. S431–S438.

Koskinen H, A Krasnov, C Rexroad, Y Gorodilov, S Afanasyev, and H Molsa. 2004a. The 14-3-3 proteins in the teleost fish rainbow trout (*Oncorhynchus mykiss*). J Exp Biol, 207, pp. 3361–3368.

Koskinen H, P Pehkonen, E Vehniainen, A Krasnov, C Rexroad, S Afanasyev, H Molsa, and A Oikari. 2004b. Response of rainbow trout transcriptome to model chemical contaminants. Biochem Biophys Res Commun, 320, pp. 745–753.

Krasnov A, H Koskinen, S Afanasyev, and H Molsa. 2005a. Transcribed Tc1-like transposons in salmonid fish. BMC Genomics, 6, p. 107.

Krasnov A, H Koskinen, P Pehkonen, and CE Rexroad III, S Afanasyev, and H Molsa. 2005b. Gene expression in the brain and kidney of rainbow trout in response to handling stress. BMC Genomics, 6, p. 3.

Krasnov A, H Koskinen, C Rexroad, S Afanasyev, H Molsa, and A Oikari. 2005c. Transcriptome responses to carbon tetrachloride and pyrene in the kidney and liver of juvenile rainbow trout (*Oncorhynchus mykiss*). Aquat Toxicol, 74, pp. 70–81.

Kurobe T, M Yasuike, T Kimura, I Hirono, and T Aoki. 2005. Expression profiling of immune-related genes from Japanese flounder *Paralichthys olivaceus* kidney cells using cDNA microarrays. Dev Comp Immunol, 29, pp. 515–523.

Lan Y, X Xu, F Yang, and X Zhang. 2006. Transcriptional profile of shrimp white spot syndrome virus (WSSV) genes with DNA microarray. Arch Virol, 151, pp. 1723–1733.

Leung YF and D Cavalieri. 2003. Fundamentals of cDNA microarray data analysis. Trends Genet, 19, pp. 649–659.

Li J and M Burmeister. 2005. Genetical genomics: combining genetics with gene expression analysis. Hum Mol Genet, 2, pp. R163–R169.

Lipshutz RJ, SP Fodor, TR Gingeras, and DJ Lockhart. 1999. High density synthetic oligonucleotide arrays. Nat Genet, 21, pp. 20–24.

MacKenzie S, D Iliev, C Liarte, H Koskinen, JV Planas, FW Goetz, H Molsa, A Krasnov, and L Tort. 2006. Transcriptional analysis of LPS-stimulated activation of trout (*Oncorhynchus mykiss*) monocyte/macrophage cells in primary culture treated with cortisol. Mol Immunol, 43, pp. 1340–1348.

Manduchi E, LM Scearce, JE Brestelli, GR Grant, KH Kaestner, and CJ Stoeckert Jr. 2002. Comparison of different labeling methods for two-channel high-density microarray experiments. Physiol Genomics, 10, pp. 169–179.

Marks H, O Vorst, AM van Houwelingen, MC van Hulten, and JM Vlak. 2005. Gene-expression profiling of White spot syndrome virus in vivo. J Gen Virol, 86, pp. 2081–2100.

Martin SA, SC Blaney, DF Houlihan, and CJ Secombes. 2006. Transcriptome response following administration of a live bacterial vaccine in Atlantic salmon (*Salmo salar*). Mol Immunol, 43, pp. 1900–1911.

Meijer AH, FJ Verbeek, E Salas-Vidal, M Corredor-Adamez, J Bussman, AM van der Sar, GW Otto, R Geisler, and HP Spaink. 2005. Transcriptome profiling of adult zebrafish at the late stage of chronic tuberculosis due to *Mycobacterium marinum* infection. Mol Immunol, 42, pp. 1185–1203.

Moriya S, S Urawa, O Suzuki, A Urano, and S Abe. 2004. DNA microarray for rapid detection of mitochondrial DNA haplotypes of chum salmon. Mar Biotechnol (NY), 6, pp. 430–434.

Morrison RN, GA Cooper, BF Koop, ML Rise, AR Bridle, MB Adams, and BF Nowak. 2006. Transcriptome profiling the gills of amoebic gill disease (AGD)-affected Atlantic salmon (*Salmo salar* L.)—A role for tumor suppressor p53 in AGD-pathogenesis? Physiol Genomics, 26, pp. 15–34.

Nash JH, WA Findlay, CC Luebbert, OL Mykytczuk, SJ Foote, EN Taboada, CD Carrillo, JM Boyd, DJ Colquhoun, ME Reith, and LL Brown. 2006. Comparative genomics profiling of clinical isolates of Aeromonas salmonicida using DNA microarrays. BMC Genomics, 7, p. 43.

Nuwaysir EF, W Huang, TJ Albert, J Singh, K Nuwaysir, A Pitas, T Richmond, T Gorski, JP Berg, J Ballin, M McCormick, J Norton, T Pollock, T Sumwalt, L Butcher, D Porter, M Molla, C Hall, F Blattner, MR Sussman, RL Wallace, F Cerrina, and RD Green. 2002. Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. Genome Res, 12, pp. 1749–1755.

Pawitan Y, S Michiels, S Koscielny, A Gusnanto, and A Ploner. 2005. False discovery rate, sensitivity and sample size for microarray studies. Bioinformatics, 21, pp. 3017–3024.

Purcell MK, KM Nichols, JR Winton, G Kurath, GH Thorgaard, P Wheeler, JD Hansen, RP Herwig, and LK Park. 2006. Comprehensive gene expression profiling following DNA vaccination of rainbow trout against infectious hematopoietic necrosis virus. Mol Immunol, 43, pp. 2089–2106.

Renn SC, N Aubin-Horth, and HA Hofmann. 2004. Biologically meaningful expression profiling across species using heterologous hybridization to a cDNA microarray. BMC Genomics, 5, p. 42.

Rise ML, SR Jones, GD Brown, KR von Schalburg, WS Davidson, and BF Koop. 2004b. Microarray analyses identify molecular biomarkers of Atlantic salmon macrophage and hematopoietic kidney response to *Piscirickettsia salmonis* infection. Physiol Genomics, 20, pp. 21–35.

Rise ML, KR von Schalburg, GD Brown, MA Mawer, RH Devlin, N Kuipers, M Busby, M Beetz-Sargent, R Alberto, AR Gibbs, P Hunt, R Shukin, JA Zeznik, C Nelson, SR Jones, DE Smailus, SJ Jones, JE Schein, MA Marra, YS Butterfield, JM Stott, SH Ng, WS Davidson, and BF Koop. 2004a. Development and application of a salmonid EST database and cDNA

microarray: data mining and interspecific hybridization characteristics. Genome Res, 14, pp. 478–490.

Sarropoulou E, G Kotoulas, DM Power, and R Geisler. 2005. Gene expression profiling of gilthead sea bream during early development and detection of stress-related genes by the application of cDNA microarray technology. Physiol Genomics, 23, pp. 182–191.

Schena M, D Shalon, RW Davis, and PO Brown. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 270, pp. 467–470.

Schmitt B and L Henderson. 2005. Diagnostic tools for animal diseases. Rev Sci Tech., 24, pp. 243–250.

Shih JH, AM Michalowska, K Dobbin, Y Ye, TH Qiu, and JE Green. 2004. Effects of pooling mRNA in microarray class comparisons. Bioinformatics, 20, pp. 3318–3325.

Simon R, MD Radmacher, and K Dobbin. 2002. Design of studies using DNA microarrays. Genet Epidemiol, 23, pp. 21–36.

Smyth GK, J Michaud, and HS Scott. 2005. Use of within-array replicate spots for assessing differential expression in microarray experiments. Bioinformatics, 21, pp. 2067–2075.

Tilton SC, LG Gerwick, JD Hendricks, CS Rosato, G Corley-Smith, SA Givan, GS Bailey, CJ Bayne, and DE Williams. 2005. Use of a rainbow trout oligonucleotide microarray to determine transcriptional patterns in aflatoxin B1-induced hepatocellular carcinoma compared to adjacent liver. Toxicol Sci, 88, pp. 319–330.

Tsai JM, HC Wang, JH Leu, HH Hsiao, AH Wang, GH Kou, and CF Lo. 2004. Genomic and proteomic analysis of thirty-nine structural proteins of shrimp white spot syndrome virus. J Virol, 78, pp. 11360–11370.

Tusher VG, R Tibshirani, and G Chu. 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA, 98, pp. 5116–5121.

von Schalburg KR, SP McCarthy, ML Rise, JC Hutson, WS Davidson, and BF Koop. 2006. Expression of morphogenic genes in mature ovarian and testicular tissues: potential stem-cell niche markers and patterning factors. Mol Reprod Dev, 73, pp. 142–152.

von Schalburg KR, ML Rise, GD Brown, WS Davidson, and BF Koop. 2005b. A comprehensive survey of the genes involved in maturation and development of the rainbow trout ovary. Biol Reprod, 72, pp. 687–699.

von Schalburg KR, ML Rise, GA Cooper, GD Brown, AR Gibbs, CC Nelson, WS Davidson, and BF Koop. 2005a. Fish and chips: various methodologies demonstrate utility of a 16,006-gene salmonid microarray. BMC Genomics, 6, p. 126.

Vornanen M, M Hassinen, H Koskinen, and A Krasnov. 2005. Steady-state effects of temperature acclimation on the transcriptome of the rainbow trout heart. Am J Physiol Regul Integr Comp Physiol, 289, pp. R1177–1184.

Walsh B and D Henderson. 2004. Microarrays and beyond: what potential do current and future genomics tools have for breeders? J Anim Sci E-Suppl, pp. E292–299.

Whitfield CW, MR Band, MF Bonaldo, CG Kumar, L Liu, JR Pardinas, HM Robertson, MB Soares, and GE Robinson. 2002. Annotated expressed sequence tags and cDNA microarrays for studies of brain and behavior in the honey bee. Genome Res, 12, pp. 555–566.

Williams TD, K Gensberg, SD Minchin, and JK Chipman. 2003. A DNA expression array to detect toxic stress response in European flounder (*Platichthys flesus*). Aquat Toxicol, 65, pp. 141–157.

Woo Y, J Affourtit, S Daigle, A Viale, K Johnson, J Naggert, and G Churchill. 2004. A comparison of cDNA, oligonucleotide, and Affymetrix GeneChip gene expression microarray platforms. J Biomol Tech, 15, pp. 276–284.

Yauk CL, ML Berndt, A Williams, and GR Douglas. 2004. Comprehensive comparison of six microarray technologies. Nucleic Acids Res, 32, p. e124.

Zhao SH, J Recknor, JK Lunney, D Nettleton, D Kuhar, S Orley, and CK Tuggle. 2005. Validation of a first-generation long-oligonucleotide microarray for transcriptional profiling in the pig. Genomics, 86, pp. 618–625.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 22
Salmonid DNA Microarrays and Other Tools for Functional Genomics Research

Matthew L. Rise, Kristian R. von Schalburg, Glenn A. Cooper, and Ben F. Koop

Introduction: An Overview of Salmonid Genomic Research Projects and Resources

The advent of genomic techniques has revolutionized research in areas such as agriculture, toxicology, and medicine. Large-scale, federally funded salmonid genomic research projects have been initiated by groups in Canada, the U.S., Norway, and France. As a result, there are now large expressed sequence tag (EST) collections for Atlantic salmon and rainbow trout, and steadily evolving salmonid genomic resources such as DNA microarrays and bacterial artificial chromosome (BAC) libraries. This chapter aims to provide a brief overview of the progress to date in the area of salmonid functional genomic research, focusing on the literature pertaining to salmonid ESTs and DNA microarrays. We present current applications of salmonid genomic resources and methods, and ideas for future directions in this area. In addition, we present sample methods related to global gene expression profiling experiments, and links to Web references providing further detail on methods, reagents, and software discussed in the chapter.

Genomic research projects often begin with a gene discovery phase, in which high-complexity complementary DNA (cDNA) libraries are created and sequenced to generate EST databases. Following large-scale gene discovery, a genomic research project may begin building resources, such as DNA microarrays, for global gene expression profiling in functional genomic research. We will briefly introduce salmonid EST collections and DNA microarrays here, and subsequent sections of the chapter will consider these resources in greater depth.

Salmonid ESTs

Links to Web sites providing information for this chapter are given in the Web Site References section at the end of this chapter. Using the Taxonomy Browser of the National Center for Biotechnology Information (NCBI), we identified 30 salmonid species having at least 10 nucleotide sequences each in GenBank. Of the 496,982 nucleotide sequences (as of April 24, 2006), 249,740 (50.3%) are from rainbow trout (*Oncorhynchus mykiss*), and 239,837 (48.3%) are from Atlantic salmon (*Salmo salar*). EST projects contributing greater than 50,000 salmonid nucleotide sequences to GenBank include the Genome Canada-funded Genomic Research on Atlantic Salmon

Project (GRASP, now called cGRASP for the consortium for Genomic Research on All Salmon Project), the Norwegian Salmon Genome Project, the United States Department of Agriculture-funded rainbow trout genomic project of the National Center for Cool and Cold Water Aquaculture, and the French Institut National de la Recherche Agronomique Analyse du Genome des Animaux d'Elevage (INRA AGENAE) rainbow trout program. The Institute for Genomic Research (TIGR) gene indices for rainbow trout and Atlantic salmon provide valuable tools such as the association of assembled ESTs with metabolic pathways. However, TIGR gene indices are not as current as GenBank; the rainbow trout gene index version 5.0 was released January 31, 2005, and contains 199,167 sequences, and the Atlantic salmon gene index version 2.1 was released June 22, 2004, and contains only 87,721 sequences. Large-scale salmonid EST projects have dramatically improved the characterization of salmonid transcriptomes, resulting in the creation of genomic resources (i.e., EST databases and DNA microarrays) that are being used for functional genomics research in areas such as aquaculture, ecology, fish health, and toxicology. The characterization of salmonid transcriptomes also benefits from smaller EST projects, such as those that identify genes responsive to stressors such as pathogens or environmental toxicants. In the gene discovery section of this chapter, we will provide an overview of the literature related to salmonid ESTs and descriptions of related genomic and bioinformatics techniques.

Salmonid DNA Microarrays

DNA microarrays are particularly important genomic research tools because they can be used to reveal the relative expression of thousands of genes simultaneously, thereby allowing the rapid identification of molecular pathways altered during pathological or normal processes. Because of the popularity of the EST/microarray approach for gene discovery and global gene expression profiling, especially within the salmonid research community, the global gene expression profiling and applications for salmonid DNA microarrays sections of this chapter focus on these resources and related applications. Other methods for global gene expression profiling (i.e., Serial Analysis of Gene Expression [SAGE] and Massively Parallel Signature Sequencing [MPSS]) have not yet appeared in the scientific literature related to salmonid research, and therefore will only be briefly described later in the chapter.

Gene Discovery

Gene expression analysis has been compared to a puzzle, where every piece must be present from the beginning of the analysis for a complete picture to emerge (Constans 2002). In studying the overall regulation of a biological process, such as host immune response to a pathogen, it is necessary to first identify all pieces of the puzzle (i.e., immune-relevant genes). For the last several years, the standard method of large-scale gene identification has been through the creation and characterization of high-complexity cDNA libraries. GRASP and the Salmon Genome Project (Norway) have

employed this approach with Atlantic salmon, and INRA and the United States Department of Agriculture (USDA)-funded group at the National Center for Cool and Cold Water Aquaculture have used this approach with rainbow trout. The GRASP EST project focused primarily on 3-prime sequencing of directionally cloned inserts because analysis of poorly conserved 3-prime untranslated regions allows discrimination between duplicated genes. This is an important consideration, particularly in tetraploid organisms such as salmonids. Other salmonid EST projects have focused on 5-prime sequencing to potentially increase the proportion of Basic Local Alignment Search Tool (BLAST)-identifiable ESTs. Midway through the GRASP EST effort they reported that 61,819 Atlantic salmon ESTs (primarily 3′) derived from 23 cDNA libraries assembled into 11,560 contigs (contiguous sequences containing two or more ESTs) and 17,150 singletons, totaling 28,710 different putative transcripts (Rise et al. 2004b). Of these assembled Atlantic salmon ESTs, 10,511 (36.6%) had significant BLASTX hits, 13,459 (46.9%) had significant BLASTN hits, and 11,802 (41.1%) had no significant BLAST hits (E value threshold for significance = 10^{-5}) (Rise et al. 2004b). At that point in the GRASP EST project, 10–20% of sequences from their highest complexity libraries were novel, suggesting that the total number of Atlantic salmon genes may be greater than 30,000.

As of April 2006, GRASP has built and sequenced more than 200 salmonid cDNA libraries (von Schalburg et al. 2005b) (Table 22.1). The libraries were derived from a variety of tissue types, life stages, and histories, and from both healthy and immunologically challenged fish. Various processes (i.e., normalization, subtractive hybridization, and selective cloning of higher molecular weight fractions) were used to increase cDNA library complexity, thereby improving the rate of discovery of rare transcripts (Rise et al. 2004b). GRASP has sequenced 180,955 Atlantic salmon clones, yielding 211,855 high-quality sequences. By including all of GRASP's EST sequence data and all available EST data from NCBI for Atlantic salmon, the total number of high-quality sequences is 251,725 that were assembled into 37,281 contigs and 34,820 singletons (72,101 putative transcripts) (Table 22.2). In addition to its Atlantic salmon EST database, cGRASP has generated smaller collections of rainbow trout, chinook salmon (*O. tshawytscha*), sockeye salmon (*O. nerka*), lake whitefish (*Coregonus clupeaformis*), and rainbow smelt (*Osmerus mordax*) ESTs (Tables 22.1 and 22.2).

Norway's Salmon Genome Project has contributed significantly to the characterization of the Atlantic salmon transcriptome by submitting more than 55,000 ESTs to GenBank (http://www.salmongenome.no/cgi-bin/sgp.cgi). A recent publication from this group involved 6,262 high-quality 5-prime sequences from non-normalized cDNA libraries built using presmolt Atlantic salmon gill and intestine tissues (Hagen-Larsen et al. 2005). These ESTs assembled into 2,974 putative transcripts (779 contigs and 2,195 singletons), approximately half of which had significant BLAST hits (E values less than 10^{-15}) (Hagen-Larsen et al. 2005). A project based at the National University of Ireland surveyed 1,152 Atlantic salmon ESTs from six non-normalized cDNA libraries (Davey et al. 2001), and 733 ESTs from a non-normalized liver cDNA library (Martin et al. 2002).

The rainbow trout EST project of the USDA/Agricultural Research Service (ARS) National Center for Cool and Cold Water Aquaculture generated approximately 47,000 high-quality 5-prime ESTs arising from a normalized mixed-tissue (brain, gill, liver, kidney, and muscle) cDNA library (Rexroad et al. 2003). These ESTs were added to gene sequences in GenBank to create the first TIGR Rainbow Trout Gene

Table 22.1. cGRASP cDNA Library Summary Statistics.

Salmonid species	Library or library group	No. of good sequences	No. of putative transcripts	No. of singletons	No. of contigs	Maximum no. of clones in contig	Average no. of clones in contig	No. of clones on 16K microarray
Salmo salar	Brain—higher molecular weight (MW)	2,565	1,990	554	1,436	27	1.3	90
	Eye	1,259	892	211	681	24	1.4	0
	Esophagus	749	502	102	400	51	1.5	27
	Gill	5,103	3,314	890	2,424	46	1.5	455
	Heart	2,777	1,829	378	1,451	33	1.5	0
	Kidney	5,965	3,915	822	3,093	39	1.5	92
	Head kidney—infected	2,104	1,297	581	716	30	1.6	334
	Head kidney—stimulated	1,155	814	373	441	65	1.4	3
	Head kidney—normalized	1,405	931	192	739	4	1.5	0
	Intestine	2,539	1,771	439	1,332	49	1.4	0
	Liver	2,782	991	169	822	76	2.8	0
	Liver—normalized	1,379	1,006	172	834	16	1.4	244
	Mixed gut	3,753	2,541	465	2,076	28	1.5	676
	Mixed tissue non-normalized	89	76	12	64	3	1.2	8

Mixed tissue—normalized	128,535	50,733	22,943	27,790	89	2.5	8,079
Liver, spleen, head kidney	2,324	1,222	123	1,099	14	1.9	0
Ovary	5,581	4,047	1,096	2,951	32	1.4	527
Pituitary gland	3,114	1,688	429	1,259	142	1.8	314
Pyloric caecum	13,543	5,875	943	4,932	89	2.3	1,157
Red muscle	880	380	36	344	60	2.3	0
Retina	1,237	996	245	751	21	1.2	84
Skeletal muscle—normalized and non-normalized	1,686	1,110	284	826	53	1.5	51
Skin	1,078	682	138	544	34	1.6	0
Spleen	5,954	3,936	938	2,998	28	1.5	267
Swim bladder	2,207	930	155	775	66	2.4	0
Testes	4,806	3,435	892	2,543	16	1.4	101
Thyroid and jaw	17	17	7	10	1	1.0	3
Thyroid—normalized	1,424	973	207	766	6	1.5	0
Whole juvenile—non-normalized	268	230	138	92	6	1.2	0
Whole juvenile—normalized	3,832	2,533	507	2,026	44	1.5	910

(*Continued*)

Table 22.1. (*Continued*)

Salmonid species	Library or library group	No. of good sequences	No. of putative transcripts	No. of singletons	No. of contigs	Maximum no. of clones in contig	Average no. of clones in contig	No. of clones on 16K microarray
Oncorhynchus mykiss	Brain	749	604	254	350	14	1.2	85
	Mixed embryonic	209	184	45	139	7	1.1	40
	Mixed gonad	7,990	6,522	3,160	3,362	58	2.3	514
	Whole juvenile—lower MW	296	251	42	209	5	1.2	57
	Whole juvenile—higher MW	936	769	304	465	16	1.2	316
	Whole juvenile—normalized	22,041	6,546	1,176	5,370	194	3.4	1,574
	Mixed tissue	41,188	21,908	8,461	13,447	108	1.9	0
	Mixed tissue—normalized	44,154	27,505	15,216	12,289	158	1.6	0
	Pituitary gland	3,576	1,584	413	1,171	101	2.3	0
	Neurogenesis	6,731	3,288	984	2,304	69	2.0	1
Osmerus mordax	Liver	1,747	741	441	300	153	2.4	1
	Brain	440	402	286	116	4	1.1	2
	Kidney	339	252	116	136	6	1.3	0
	Spleen	596	303	165	138	45	2.0	0
	Mixed tissue—normalized	4,321	3,099	2,079	1,020	10	1.4	0
Coregonus clupeaformis	Brain	2,212	1,634	1,250	384	31	1.4	2
	Mixed tissue—normalized	3,193	2,379	1,639	740	7	1.3	1

Table 22.2. cGRASP salmonid EST project summary statistics.

	Salmo salar[a]	*Oncorhynchus mykiss*[b]	*Oncorhynchus tshawytscha*[a]	*Oncorhynchus nerka*[a]	*Coregonus clupeaformis*[a]	*Osmerus mordax*[a]
Number of good sequences[c]	251,725	246,704	4,572	1,243	5,405	7,443
Number of contigs[d]	37,281	42,423	791	291	1,013	1,345
Number of singletons	34,820	25,777	2,546	229	2,889	3,087
Number of putative transcripts	72,101	68,200	3,337	520	3,902	4,432
Maximum contig size (No. of ESTs)	395	345	16	21	31	154
Average contig size (No. of ESTs)[e]	5.8	5.2	2.1	4.2	2.3	9.9

[a]Statistics reflect the state of the cGRASP EST database on April 1, 2006.
[b]Statistics reflect the state of the cGRASP EST database on January 1, 2005.
[c]Sequences were considered "good" if their trimmed PHRED20 lengths were at least 100 bases.
[d]A contig (contiguous sequence, or assembled EST) contains two or more ESTs.
[e]Average assembled EST size = (total number of good sequences) / (number of putative transcripts).

Index, which contained 7,956 contigs and 17,669 singletons, totaling 25,625 putative transcripts (Rexroad et al. 2003). Of the 7,646 contigs containing only ESTs, 3,912 (51.2%) had significant BLAST hits (E values less than 10^{-5}); of the 17,479 EST singletons, 4,690 (26.8%) had significant BLAST hits (Rexroad et al. 2003). Since their original publication, the USDA group has submitted additional rainbow trout EST collections to GenBank, including approximately 3,500 sequences from a normalized pituitary gland library (Gahr et al. 2006) and several thousand sequences from cDNA libraries constructed from early life stage trout (Rise, manuscript in preparation). The INRA AGENAE program has sequenced more than 80,000 clones from high-quality rainbow trout cDNA libraries, and has submitted the majority of these sequences to public databases such as the European Molecular Biology Laboratory (EMBL) and GenBank (Aegerter et al. 2004). The rainbow trout EST project of the Institute of Applied Biotechnology, University of Kuopio, Finland, generated approximately 2,000 ESTs from high-complexity cDNA libraries constructed using RNA from tissues of stressed fish (Krasnov et al. 2005b). The characterization of Atlantic salmon and rainbow trout transcriptomes approaches completion, and additional salmonid species are becoming better represented in public sequence data repositories. Following the lead of the National Institutes of Health Mammalian Gene Collection Program (Gerhard et al. 2004), it is now time for salmonid genomic research projects to work together toward generation of functionally annotated, publicly accessible cDNA resources containing complete open reading frames for all salmonid genes.

Large-scale salmonid EST projects have enabled the creation of functional genomic research tools such as cDNA and oligonucleotide microarrays. There are numerous additional uses for salmonid EST collections. For example, salmonid ESTs can be placed on BACs or physical maps (i.e., Thorsen et al. 2005, von Schalburg et al. 2005b), and microsatellite sequences in noncoding regions of salmonid ESTs can be mapped on genetic maps (Rexroad et al. 2005). This will facilitate the integration of salmonid physical and genetic maps, and contribute to the development of genetic maps with the high marker densities needed for fine mapping of quantitative trait loci (QTL). EST-linked microsatellites have been used to detect genetic signatures of divergent selection in eight free-living populations of Atlantic salmon (Vasemägi et al. 2005). In addition, salmonid ESTs have been mined to identify potential functions of organs such as the pyloric caecum (Rise et al. 2004b), to detect morphogenic activities in mature gonads (von Schalburg et al. 2006), and to identify and facilitate molecular phylogenetic analyses of transcribed transposons (Krasnov et al. 2005a).

Construction of High-complexity cDNA Libraries

Normalized cDNA Libraries

Normalization is a well-known method of increasing cDNA library complexity (enriching for rare transcripts and reducing the frequency of abundant transcripts), thereby accelerating gene discovery. Normalized cDNA libraries have been important components of large-scale rainbow trout and Atlantic salmon gene discovery projects (Rexroad et al. 2003, Rise et al. 2004b). To illustrate the utility of cDNA library

Table 22.3. Statistics[1] showing the influence of cDNA library normalization.

Library	% new (species)[2]	% no BLAST hit[3]
Prenormalized lower MW[4] *S. salar* pyloric caecum	11.9	15.5
Normalized lower MW *S. salar* pyloric caecum	36.3	25.5
Prenormalized lower MW *O. mykiss* whole juvenile	37.1	15.4
Normalized lower MW *O. mykiss* whole juvenile	71.6	32.2

[1]Data are from Rise et al. 2004b.
[2]Percent new (species) value is the number of previously unidentified EST clusters (including contigs and singletons) divided by the total number of clusters.
[3]Percentage of EST clusters (including contigs and singletons) with no significant BLASTN or BLASTX hit ($E < 10^{-5}$).
[4]MW = molecular weight.

normalization, statistics indicative of gene discovery rate for a set of prenormalized and normalized salmonid cDNA libraries are shown in Table 22.3. The production of a normalized cDNA library first requires the construction of a "prenormalized" library that is initially evaluated on its average insert size, titre, and complexity. If the newly constructed library possesses the required qualities, it may then proceed on to the normalization process using standard molecular methods (i.e., Soares et al. 1994, Zhulidov et al. 2004).

The following is a sample method for construction of a directionally cloned, prenormalized cDNA library from a high-quality messenger RNA (mRNA) (polyA+ RNA) preparation. Libraries built in this manner are discussed by Rise and others (2004b). Adult male and female Atlantic salmon pyloric caeca are collected and flash frozen in liquid nitrogen. Pyloric caeca mRNA samples are prepared from pulverized tissue samples (ground to a fine powder under liquid nitrogen, with baked mortars and pestles) using the Poly(A)Pure mRNA Isolation Kit (Ambion, Inc., Austin, TX), and mRNA samples are quality checked and quantified by agarose gel electrophoresis and spectrophotometry, respectively. Directionally cloned (5′ *Eco*RI, 3′ *Xho*I) cDNA libraries are prepared from mRNA using reagents and methods such as those included in the pBluescriptII XR cDNA Library Construction Kit (Stratagene, La Jolla, CA). Size fractionation by 1% agarose gel extraction (Qiagen) is performed on *Xho*I digested cDNAs immediately prior to ligation into pBluescript II SK (+). ElectroMAX DH12S cells are transformed by electroporation with the ligation reaction. Twelve clones from each prenormalized cDNA library are selected to estimate insert sizes. Using these methods, mean insert sizes of prenormalized libraries are usually in the range of 1,000–1,300 base pairs (bp) (Rise et al. 2004b). The products mentioned in this chapter reflect the personal experience of the authors, and are not intended to be a list of all possible reagents for a given procedure. For alternative methods, readers are referred to Chapter 20. Web links for these products are provided in the Web Site References section at the end of this chapter.

A detailed method for normalizing cDNA libraries, based on Soares and others (1994), may be found in Gahr and others (2006). This method involves isolation of single-stranded DNA phagemid using helper phage, hybridization of the single-stranded DNA with a "driver" (all inserts of the library polymerase chain reaction (PCR) amplified with universal primers), and posthybridization separation of single-stranded DNA (enriched for rare transcripts) and double-stranded DNA (enriched for abundant transcripts) by hydroxyapatite column chromatography (Soares et al. 1994, Bonaldo et al. 1996). Another method of normalizing cDNA libraries uses a duplex-specific nuclease to degrade reassociated double-stranded fractions containing abundant transcripts, and subsequent PCR amplification of the equalized single-stranded fraction (Zhulidov et al. 2004).

Suppression Subtractive Hybridization (SSH) Libraries

Existing salmonid EST databases and microarrays contain thousands of known and unknown genes. However, because these resources have been derived from cDNA libraries constructed largely from healthy embryonic, larval, and adult tissues (Rise et al. 2004b, Rexroad et al. 2003), they may be missing some genes involved in salmonid responses to various stressors. A complete understanding of the molecular mechanisms involved in pathological processes may require, in addition to the use of existing genomic tools, the development of new resources such as targeted suppression subtractive hybridization (SSH) cDNA libraries and disease-relevant host and pathogen EST databases.

The construction of SSH libraries is relatively straightforward, and may be accomplished using commercially available products such as the Advantage cDNA PCR kit (Clontech). SSH libraries are enriched for genes up-regulated in a "tester" sample relative to a "driver" sample, and reciprocal SSH libraries are often created to identify genes both induced and suppressed by a given stressor. SSH libraries have been used to identify salmonid immune-relevant cell and tissue genes that are potentially responsive to pathogens (i.e., Bayne et al. 2001, O'Farrell et al. 2002, Sangrador-Vegas et al. 2002, Tsoi et al. 2004, Rise et al. 2004b, Ewart et al. 2005), and have also aided in the discovery of salmonid genes involved in normal developmental processes such as ovary development (von Schalburg et al. 2005a) and embryogenesis (Rise et al. 2004b).

An example of steps involved in SSH library construction and characterization follows. MicroPoly(A)Pure kits (Ambion) are used to prepare mRNA from flash-frozen samples (i.e., pathogen exposed spleen versus nonexposed control spleen). Poly(A)+ RNAs are converted into cDNAs and reference (driver) and experimental (tester) cDNAs are subjected to SSH using the PCR-Select cDNA Subtraction kit and the manufacturer's instructions (BD Biosciences—Clontech; see Web Site References for the link to the User Manual). Reciprocal SSH libraries are built, one enriched for genes up-regulated in response to the stressor and the other enriched for genes down-regulated in response to the stressor. SSH library construction yields PCR products that may be cloned into pCR4-TOPO vector and transformed using Top10 electrocompetent cells (Invitrogen). An initial evaluation of insert size and complexity is made by visual comparison of clone restriction fragments with DNA size markers such as *Hin*dIII and 1 kilobase (kb) ladder. One 96-well plate of clones from each SSH library is sequenced, and these sequences are quality checked and assembled into contiguous sequences and singletons.

Characterization of cDNA Libraries

Acquisition of cDNA Library Sequence Data

Obtaining sequence data for each EST clone from a particular cDNA library is a multistep process. The cDNA library is plated and grown at an appropriate density on solid media, and then individual EST clones are picked, subjected to recombinant plasmid DNA isolation and purification, and cycle sequenced to increase the signal of the transcript. The sequencing reaction is then loaded on an automated sequencer. Depending upon scale and the available resources of the project, these processes can be either completely automated by various robotic systems, all performed manually, or a combination of both.

Following the cDNA library construction phase, the efficiency of each library is determined by calculating its titre. The titre represents the number of colony forming units (cfu) contained within a set volume of the library (i.e., 10^5 cfu/ml). A library with a low titre (i.e., due to low efficiency of ligation of inserts into vector) may not be of maximum use to the researcher because it may not allow for enough EST clones to be picked and sequenced to fully capture all of the transcripts found within the portion of the transcriptome represented within the original RNA pool.

The library is initially plated on selective media containing the appropriate antibiotic to select for bacteria carrying the recombinant vector that confers resistance to the antibiotic. Other reagents may be added to the media that will help determine which bacteria are harboring recombinant plasmid. Typically colonies are arrayed into 96-well or 384-well format growth plates and grown up overnight with vigorous shaking (i.e., 300 rpm) to obtain a desired bacterial density. Glycerol stocks (approximately 15% glycerol) of overnight cultures, prepared in the same format corresponding to the "mother" plates, are archived at −80°C.

Isolation of the recombinant plasmid DNA from the bacterium is loosely based on the method by Birnboim and Doly (1979). In general, the overnight bacterial culture is lysed in an alkaline solution, and then neutralized to release the plasmid DNA into the lysate. This lysate is cleared from the bacterial chromosomal DNA and cellular debris by various methods, including high-speed centrifugation, trapping bacterial debris by a filter trap, and selectively binding the plasmid to a column matrix. The plasmid DNA is then precipitated or eluted from the column matrix, and then washed with ethanol. Purified plasmid is resuspended in either water or mildly buffered solution (i.e., TE). All residual RNA is removed by the addition of RNase A either at the beginning or the end of the process. Removal of RNA and bacterial cellular debris is critical because it will interfere with the subsequent downstream sequencing processes. Before sequencing the recombinant plasmid, its quantity and purity are determined by measuring its absorbance at 260 nanometer (nm) and its absorbance ratio at 260 nm/280 nm, respectively.

The medium-throughput plasmid DNA purification method employed by GRASP uses Growth, Clarification and DNA Recovery Plates (Whatman, Clifton, NJ), where the bacterial debris is filter-trapped at the final stages of the extraction process. The high-throughput method employed by GRASP uses 384-well plates (Axygen, Union City, CA) and requires a centrifugation step to pellet the chromosomal DNA, cellular debris, and protein complexes, leaving only cleared lysate containing plasmid DNA.

Sequencing of the isolated plasmid DNA can be performed on a variety of automated platforms (i.e., ABI DNA Analyzer, Amersham MegaBase, Licor, Solexa). For the GRASP project, EST sequencing is performed on both the Applied Biosystems 3700 and 3730 DNA Analyzer (ABI, Foster City, CA). For cycle sequencing reactions (Carothers et al. 1989) to be analyzed on the ABI 3700 DNA Analyzer, a total of 300 nanograms (ng) of purified plasmid DNA, 3.2 picomoles (pmoles) of primer, 1.0 μl of ABI BigDye® Terminator v3.1 Ready Reaction Mix, and 0.5 μl of 5× Big Dye Sequencing Buffer are combined and brought up to a final volume of 5.0 μl. Due to the increased sensitivity of the 3730 DNA Analyzer over the 3700 model, the amount of template and BigDye® Terminator Reaction Mix can be halved when using the ABI 3730.

Cycle sequencing is performed on a thermal cycler (MJ Research, Waltham, MA) with all ramp rates at 1°C/second. The template is initially denatured at 96°C for 1 minute. The reaction is then cycled under the following conditions: 96°C for 10 seconds, 50°C for 5 seconds, 60°C for 3 minutes, and repeated either 40 times or 36 times depending on whether the sequencing product is loaded onto an ABI 3700 or 3730, respectively. The sequencing reaction is then precipitated in 9.3 mM EDTA and 70% ethanol. Salts are removed by washing the pellet with 70% ethanol. The ethanol is removed and the pellet is air-dried for 10 minutes. The sequencing product is rehydrated in 10 μl of nuclease-free water and loaded onto a DNA Analyzer.

Sequence Data Processing and Management

Due to the ability to automate much of the sequencing processes described, large amounts of sequence data can be rapidly generated. It is therefore critical to have an automated pipeline that can deposit the sequence chromatograms into a database, store, archive, and automatically process the data to give the researcher current project statistics.

There are essentially three components involved in the curation of an EST database: data preprocessing, data processing, and data postprocessing. In the preprocessing stage, base-calling from chromatogram traces is performed using PHRED (Ewing et al. 1998, Ewing and Green 1998). Vector, poly-A tails, and low quality regions are trimmed from EST sequences. For the GRASP project, sequences with fewer than 100 good quality bases after trimming are discarded.

These preprocessed ESTs are assembled into contiguous sequences (contigs) using PHRAP (http://www.phrap.org/). In this processing step, contig consensus sequences and singleton sequences are aligned with nonredundant GenBank nucleotide and amino acid sequence databases using BLASTN and BLASTX, respectively (Altschul et al. 1990, 1997) and assigned "identifiers." Each EST sequence is submitted to GenBank, assigned an accession number, and uploaded into the public database. Each new EST sequence that is entered into the database must be "keyed" to correspond with the archived cDNA clone from which the information was originally obtained.

It is important to determine what information is of most value for the user to glean from the stream of sequences being deposited into the database. The data postprocessing component deals with this step. In the case of GRASP, tracking "length of sequence reads," "good versus bad reads," and "length of insert size" is important as a measure of data quality control. Individual library summary statistics are generated

providing information on the percent singletons, number of clusters, largest cluster size, and the average cluster size (Tables 22.1 and 22.2). Also, an assessment of library complexity is of great importance. The complexity of each library is initially assessed by performing BLASTN and BLASTX alignments of ESTs against a database composed of all nonredundant nucleic and amino acid sequences from the appropriate salmonid species in GenBank, plus an in-house collection of nonredundant ESTs from that species at the time of new data entry. The higher "percent new" and "percent no significant BLAST hit" values indicates a higher rate of gene discovery and library complexity (Table 22.3). Statistics (i.e., percent new within species versus time of EST entry into the database) for the most complex high molecular weight and normalized libraries establish a threshold and provide a means to discriminate the level of complexity of other cDNA libraries. This method of evaluating cDNA library complexity has been effectively used by GRASP to maximize the rate of gene discovery, and sequencing efforts have focused on those cDNA libraries comparing favorably to the threshold. Quality controlled sequences in an EST database can be used to select genes for DNA microarray construction, and may be mined to identify paralogs, "unknown" genes, microsatellites, and single nucleotide polymorphisms (SNP).

How the archived sequence data are maintained and organized is largely dependent on the requirements of the user. If there is a need to access large amounts of summary statistics rapidly, then of course a large amount of computing power is required. Considerations regarding backup of stored data, power loss protection, and accessible interfaces are also paramount to the curation of an EST database. Consideration must be given to the security of the data and how it will be shared with internal and external researchers.

Global Gene Expression Profiling

Salmonid DNA Microarrays

At least six different salmonid DNA microarrays have been developed to date (Table 22.4). Each microarray platform will be presented, along with the relevant literature.

1. GRASP ~3500 gene (3.5K) and 16K cDNA microarrays: Construction of the GRASP microarrays was described in Rise and others (2004b) (3.5K) and von Schalburg and others (2005b) (16K). Briefly, clones were stringently selected from an EST database containing approximately 80,000 sequences derived from more than 30 high-complexity salmonid cDNA libraries at the time the 3.5K microarray was designed and built, and more than 300,000 sequences from more than 175 high-complexity salmonid cDNA libraries at the time the 16K microarray was constructed. Inserts were amplified by PCR, cleaned, and printed as double, side-by-side spots on ez-rays™ aminosilane slides (Matrix) in 3× SSC. *Arabidopsis* cDNAs were spotted on each microarray (Rise et al. 2004b, von Schalburg et al. 2005b), for use in thresholding or data normalization.

   The majority of the cDNA features on GRASP microarrays are from Atlantic salmon. The 3.5K GRASP microarray contains 3,119 unique Atlantic salmon and

Table 22.4. Publications involving salmonid DNA microarray platforms.

Platform	Application	Target Species[1]	Reference
GRASP 3.5K[2]	Cross-species hybridizations	AS, RT, LW, RS	Rise et al. 2004b
GRASP 3.5K[2]	Host response to *P. salmonis* infection	AS	Rise et al. 2004a
GRASP 3.5K[2]	Gene expression profiling (GEP) of brain	AS	Aubin-Horth et al. 2005
GRASP 3.5K[2]	GEP of developing precocious ovary	RT	von Schalburg et al. 2005a
GRASP 3.5K[2]	Morphogen expression in developing gonad	RT	von Schalburg et al. 2006
GRASP 3.5K[2]	Liver genes responsive to tumor promoters	RT	Tilton et al. 2006
GRASP 3.5K[2]	Response to live *A. salmonicida* vaccine	AS	Martin et al. 2006
GRASP 3.5K[2]	Comparing transcriptomes of ecotypes	LW	Derome et al. 2006
GRASP 3.5K[2]	GEP of growth hormone transgenic liver	coho	Rise et al. 2006
IMB 4K[3]	Host response to *A. salmonicida* infection	AS	Ewart et al. 2005
IMB 4K[3]	GEP of growth hormone transgenic liver	coho	Rise et al. 2006
Kuopio 1.3K[4]	GEP of ten 14-3-3 paralogs	RT	Koskinen et al. 2004a
Kuopio 1.4K[4]	Responses to sublethal doses of toxicants	RT	Koskinen et al. 2004b
Kuopio 1.3K[4]	Trancription of transposons	RT	Krasnov et al. 2005a
Kuopio 1.3K[4]	Brain, kidney response to handling stress	RT	Krasnov et al. 2005b
Kuopio 1.3K[4]	Responses to sublethal doses of toxicants	RT	Krasnov et al. 2005c
Kuopio 1.4K[4]	Effect of temperature acclimation on heart	RT	Vornanen et al. 2005
Kuopio 1.4K[4]	GEP of fry with early life stage mortality	AS	Vuori et al. 2006
Kuopio 1.4K[4]	GEP of macrophages stimulated with LPS	RT	MacKenzie et al. 2006
GRASP 16K[5]	Cross-species hybridizations, intraspecific variation, organ complexity, and gene content of bacterial artificial chromosomes	AS, RT, LW, RS, chinook	von Schalburg et al. 2005b
GRASP 16K[5]	Gill GEP of fish with amoebic gill disease	AS	Morrison et al. 2006
GRASP 16K[5]	Response to DNA vaccine against ISAV	RT	Purcell et al. 2006

GRASP 16K[5]	GEP of growth hormone transgenic liver	coho	Rise et al. 2006
OSU 1.6K[6]	GEP in AFB_1-induced hepatocarcinogenesis	RT	Tilton et al. 2005
Chum mtDNA[7]	Genetic stock identification	chum	Moriya et al. 2005
Lipid metab.[8]	Influence of diet on liver gene expression	AS	Jordal et al. 2005
~200 gene RT[9]	Hepatic gene expression response to stress	RT	Wiseman et al. 2006

[1]The species affiliations of targets (cDNAs hybridized onto microarray probes): AS = Atlantic salmon (*Salmo salar*), RT = rainbow trout (*Oncorhynchus mykiss*), LW = lake whitefish (*Coregonus clupeaformis*), RS = rainbow smelt (*Osmerus mordax*), coho = coho salmon (*Oncorhynchus kisutch*), chinook = chinook salmon (*Oncorhynchus tshawytscha*), chum = chum salmon (*Oncorhynchus keta*).
[2]The GRASP 3.5K chip includes 3,119 unique AS cDNA probes (features, or spots), and 438 unique RT cDNA probes. GRASP = Genomic Research on Atlantic Salmon Project (Canada). K: approximately 1,000 genes.
[3]The IMB 4K chip is composed of AS cDNA probes. IMB = Institute for Marine Biosciences (Canada).
[4]The Kuopio microarrays contain approximately 1,400 RT cDNAs, and were fabricated by the Institute of Applied Biotechnology, University of Kuopio (Finland) in collaboration with the USDA-ARS (USA).
[5]The GRASP 16K chip includes 13,421 unique AS cDNA probes and 2,576 unique RT cDNA probes.
[6]The OSU 1.6K chip, containing 1,672 oligonucleotide (70-mer) probes representing over 1,400 genes, was developed at Oregon State University (USA). AFB_1: aflatoxin B_1.
[7]The Chum mtDNA microarray contains 38 oligonucleotide (17- to 20-mer) probes designed for use in mitochondrial DNA (mtDNA) haplotype detection.
[8]This microarray, developed by scientists from Norway, Canada, and Scotland, contains 73 primarily Atlantic salmon cDNA features representing genes involved in lipid metabolism, and several control cDNA features.
[9]Created by scientists at the University of Victoria (Canada), this microarray contains approximately 200 cDNA fragments representing genes with known functions.

438 unique rainbow trout cDNAs (Rise et al. 2004b), and the 16K GRASP microarray contains 13,421 Atlantic salmon and 2,576 rainbow trout cDNAs (von Schalburg et al. 2005b). Cross-species hybridizations, also called "heterologous hybridizations," have been used to show that the GRASP cDNA microarrays are effective tools for global gene expression profiling using targets from various salmonid species including Atlantic salmon, rainbow trout, chinook salmon, and lake whitefish (Rise et al. 2004b, von Schalburg et al. 2005b) (Table 22.5). In addition to cross-species hybridization studies, GRASP cDNA microarrays have been used in studies of salmonid brain gene expression (Aubin-Horth et al. 2005) and ovary development (von Schalburg et al. 2005a, 2006), immune-relevant tissue gene expression responses to pathogens (Rise et al. 2004a, Morrison et al. 2006), vaccines (Purcell et al. 2006, Martin et al. 2006), tumor promoters (Tilton et al. 2006), hepatic genes responding to growth hormone transgenesis (Rise et al. 2006), and comparisons of the white muscle transcriptomes of different lake whitefish ecotypes (Derome et al. 2006) (Table 22.4).

2. Institute for Marine Biosciences (IMB) 4K cDNA microarray: Construction of the IMB microarray is described in Ewart and others (2005). Briefly, 4,104 clones were selected from Atlantic salmon liver, hematopoietic kidney, spleen, and macrophage cDNA libraries (not normalized), and SSH libraries from immune tissues or cells of fish infected by the bacterial pathogen *Aeromonas salmonicida.* Clone inserts were amplified by PCR, purified, resuspended in dH_2O, and spotted in duplicate on Gap II glass slides (Corning, NY, USA) in 50% dimethyl sulfoxide (DMSO). The entire array of 8,208 spots was reprinted on the lower half of the slide. Two-fold dilutions of a plasmid encoding chlorophyll synthetase G4 from *Arabidopsis thaliana* were included at the bottom of each subarray to serve as internal controls. The IMB 4K microarray has been used to study salmonid gene expression responses to *A. salmonicida* infection (Ewart et al. 2005), and growth hormone transgenesis (Rise et al. 2006) (Table 22.4).

Table 22.5. Hybridization characteristics of labeled Atlantic salmon (AS), rainbow trout (RT), chinook salmon, and lake whitefish targets to AS and RT probes on the 16K GRASP chip[1].

Target hybridized to chip	Atlantic salmon		Rainbow trout		Chinook salmon		Lake whitefish	
Probes on microarray (number of spots on chip)	AS (13,421)	RT (2,576)	AS (13,421)	RT (2,576)	AS (13,421)	RT (2,576)	AS (13,421)	RT (2,576)
Average % of salmonid probes passing threshold ± standard deviation[2]	52.6± 7.9	59.5± 7.5	60.7± 9.6	74.4± 6.6	48.3± 4.4	62.9± 4.8	48.9± 2.4	57.8± 0.9

[1]Data are from von Schalburg et al. 2005b. All targets were Cy3 or Cy5-labeled and hybridized to slides from the same batch where possible. Number of replicate slides per species: AS = 8, RT = 4, Chinook salmon = 4, Lake whitefish = 2.

[2]Hybridization signal threshold was defined as two standard deviations above mean signal from *Arabidopsis* features.

3. University of Kuopio 1.4K cDNA microarray: The Kuopio 1.4K cDNA microarray was first described by Krasnov and others (2005b). In a collaboration involving the Institute of Applied Biotechnology of the University of Kuopio and the National Center for Cool and Cold Water Aquaculture (USDA), approximately 1,400 primarily rainbow trout cDNA sequences were selected from ESTs arising from SSH and normalized cDNA libraries (Rexroad et al. 2003, Krasnov et al. 2005b). Salmonid cDNAs were amplified by PCR with universal primers, purified, and printed on poly-(L) lysine-coated slides; each clone was printed six times per slide (Krasnov et al. 2005b). The Kuopio 1.4K microarray has been used to study expression of rainbow trout 14-3-3 paralogs (Koskinen et al. 2004a), and salmonid gene expression responses to environmental toxicants (Koskinen et al. 2004b, Krasnov et al. 2005c), and handling stress (Krasnov et al. 2005b). In addition, this microarray has been used to study the transcription of transposons (Krasnov et al. 2005a), the heart gene expression response to temperature acclimation (Vornanen et al. 2005), whole fry genes responsive to M74 (early life stage mortality) (Vuori et al. 2006), and macrophage genes responsive to lipopolysaccharide (Mackenzie et al. 2006) (Table 22.4).
4. Oregon State University 1.6K oligonucleotide (70-mer) microarray: The OSU 1.6K microarray contains 1,672 rainbow trout 70-mer oligonucleotides, representing about 1,400 putative stress-responsive genes (Tilton et al. 2005). The oligonucleotides, designed from unique gene regions using ProbeSelect (Li and Stormo 2001) and synthesized by Operon Technologies (Alameda, CA) and Sigma Genosys (The Woodlands, TX), were resuspended in buffer (3× SSC plus 1.5M betaine) and printed in duplicate onto Corning UntraGap slides (Tilton et al. 2005). The OSU 1.6K microarray contains SpotReport Alien Oligos (Stratagene) and buffer-only control spots, and has been used to assess the impact of aflatoxin B_1-induced carcinogenesis on the rainbow trout liver transcriptome (Tilton et al. 2005) (Table 22.4).
5. Chum salmon mtDNA oligonucleotide (17 to 20-mer) microarray: The chum mtDNA microarray contains 38 short (17 to 20 bases long) oligonucleotides designed to allow microarray-based detection of chum salmon mitochondrial DNA (mtDNA) haplotypes (Moriya et al. 2005) (Table 22.4).
6. Small-scale cDNA microarray containing genes involved in lipid metabolism: This microarray was created by scientists from the National Institute of Nutrition and Seafood Research (Bergen, Norway), the Institute for Marine Biosciences (Halifax, NS, Canada), and the Institute of Aquaculture (Stirling, UK) (Jordal et al. 2005). It contains 73 primarily Atlantic salmon cDNA features, spotted in quadruplicate on CMT-GAPS coated slides (Corning Microarray Technology) (Jordal et al. 2005). This microarray, which also includes control spots (i.e., a concentration gradient of an *Arabidopsis thaliana* cDNA) and several housekeeping genes, has been used to study the effects of different dietary oils on Atlantic salmon liver gene expression (Jordal et al. 2005).
7. Small-scale rainbow trout cDNA microarray: This microarray was created by scientists at the University of Victoria (Victoria, BC, Canada). In its original form, it contained 147 (now more than 200, MM Vijayan, personal communication) unique rainbow trout cDNA fragments whose identity and function are known. Fragment lengths range from 450 to 550 bps and are amplified from the most highly conserved regions of the genes. The products represent a wide range of molecular

functions, including metabolism, signaling, transport, immune, endocrine, and development, as well as oncogenes and chaperones. Each product was brought up in 3× SSC to a final concentration of 50 ng/μl and spotted in triplicate on poly-L lysine coated slides (Telechem, Sunnyvale, CA, USA). The microarray also contains a 500-bp fragment from the Lambda Q bacterial gene as an internal control spot. The array has been used to study the transcriptional response to acute stress in rainbow trout livers (Wiseman et al. 2006).

DNA Microarray Design and Construction

All salmonid DNA microarrays currently represented in the literature are either cDNA or oligonucleotide microarrays (Table 22.4). There are numerous resources for DNA microarray design, construction, and use, including Appendix 10, DNA Array Technology in the Third Edition of *Molecular Cloning: A Laboratory Manual* (Sambrook and Russell 2001) and *DNA Microarrays: A Molecular Cloning Manual* (Editors, Bowtell and Sambrook 2003). These resources provide extensive explanations of the steps and considerations involved in building a DNA microarray (i.e., substrates, buffers, printing robots) and in designing and executing DNA microarray experiments (i.e., target synthesis and labeling, hybridization, image acquisition, and analysis). We refer the reader to these resources for background information, and will present examples of experiments illustrating the utility of currently available salmonid DNA microarrays. The sample microarray experiments have been designed to be performed with GRASP cDNA microarrays (Table 22.4). The fabrication of GRASP microarray involves PCR amplification of transcript sequences using universal primers, purification of PCR products, and robotic printing of salmonid and control cDNAs onto glass slides at the Gene Array Facility at the Prostate Centre, Vancouver General Hospital (von Schalburg et al. 2005b).

Microarray Experimental Design and Data Analysis

All microarray experiments should be MIAME-compliant (Brazma et al. 2001). MIAME (minimum information about a microarray experiment) guidelines were developed to standardize methods of reporting all aspects of a microarray experiment, including platform design, sample origin, target synthesis, hybridization and wash conditions, and methods of data extraction and analysis. The MIAME Checklist (see Web References) assists the microarray user in collection and organization of information required for MIAME compliance.

Currently, salmonid DNA microarrays are being used primarily for global gene expression profiling (Table 22.4). However, the development of new microarray-based procedures, including splice-variant analysis, on-chip chromatin immunoprecipitation (Hoheisel 2006), and nonprotein-coding RNA expression analysis (Hüttenhofer and Vogel 2006), will potentially influence future salmonid microarray-based research. Microarray experiments are often run with groups of pooled samples, and incorporate technical replicates (i.e., two replicates and two dye swaps) to identify microarray features (transcripts or genes) that have reproducible and characteristic expression differences between sample types. RNA sample pooling is thought to decrease the influence of biological variability, allowing the identification of substantive gene expression changes between sample types using relatively few microarrays

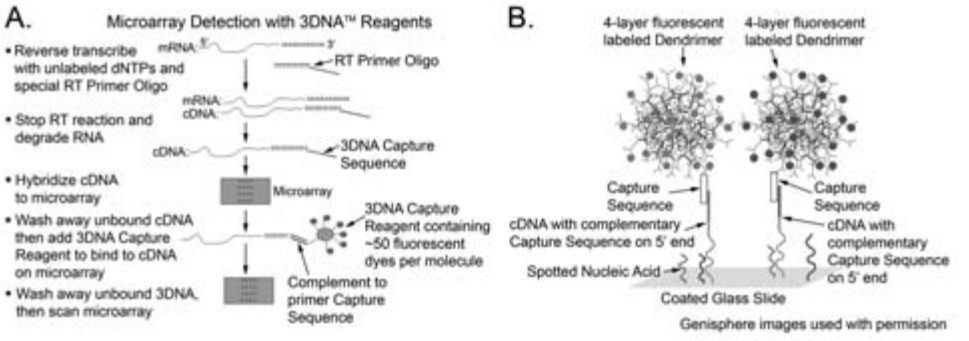


Figure 22.1. A. Steps involved in cDNA target synthesis and microarray hybridizations using Genisphere (Hatfield, PA) Expression Array Detection Kits. B. An example of a dual color nucleic acid microarray hybridization using Genisphere 3DNA reagents and methods. Genisphere figures are used with permission. (Also see color plate.)

(Kendziorski et al. 2005). Although microarray experimental designs may involve individual fish samples and therefore provide information on biological variability (i.e., Aubin-Horth et al. 2005), the issue of biological variability is often addressed using a relatively less expensive technique such as real-time quantitative reverse transcription (RT)—polymerase chain reaction (QPCR). We discuss QPCR validation of microarray results in an upcoming section of this chapter.

Experiments involving custom-built DNA microarrays are often "two-dye" experiments, so named because each chip is hybridized with a mixture of two targets labeled with different fluorescent dyes (i.e., Cy3-labeled infected macrophage target versus Cy5-labeled noninfected macrophage target) (Rise et al. 2004a). DNA microarray experiments may directly compare gene expression in two samples as in Figure 22.1 and 22.2 or they may be of reference design (i.e., von Schalburg et al. 2005a), or loop design (i.e., Aubin-Horth et al. 2005). A discussion of strengths and weaknesses of different microarray experimental designs may be found in Churchill (2002). The inclusion of dye swaps in microarray experimental designs helps to reduce the confounding effects of dye bias (i.e., differences in fluorescence signal between channels based on the dyes used rather than on transcript abundance in the RNA samples being compared) when using nonreference designed array experiments. We will present two examples of direct comparison, two dye experiments using GRASP cDNA microarrays to illustrate some of the methods and products that may be used for microarray-based research.

Sample Microarray Experiment No. 1

The products listed are simply examples from a wide variety of commercially available products related to microarray-based research, and are not necessarily recommended by the authors. The following hypothetical experiment is designed to identify genes differentially expressed in growth-enhanced Atlantic salmon liver tissue and control nonenhanced Atlantic salmon liver tissue, both from adult males. The design of this microarray experiment includes only technical replicates with pooled samples, and therefore will not provide information on biological variability. (QPCR validation of selected informative genes, using cDNA templates from individual growth-enhanced

and nonenhanced fish tissues, may be used to assess biological variability.) Liver tissue is removed from 10 growth enhanced and 10 nonenhanced males, flash-frozen in liquid nitrogen, and archived at −80°C until RNA preparation. Flash frozen tissues are individually ground to a fine powder using baked (220°C, 5 hours) ceramic mortars and pestles under liquid N_2. Total RNA is prepared separately from each tissue using TRIzol reagent (Invitrogen), and cleaned using the RNeasy MinElute kit (Qiagen). Each cleaned RNA sample is quantified by Nanodrop ND-1000 spectrophotometry (Nanodrop Technologies, Wilmington, DE) and quality checked (i.e., by agarose gel or Agilent Bioanalyzer) prior to pooling. The pooled growth-enhanced RNA sample contains equal quantities of high-quality total RNA from each of the ten individuals, and the pooled nonenhanced RNA sample is similarly prepared. Growth-enhanced versus non-enhanced liver cDNA targets are synthesized and hybridized to GRASP 16K microarrays using a commercially available microarray hybridization kit such as Genisphere Array 50 Array Detection Kit for Microarrays. The Array 50 instruction manual (see Web References) provides a detailed explanation of the labeling system chemistry and methods, which involves successive hybridizations of cDNA targets and fluorescent 3DNA reagents to the microarrays (Figures 22.1A, 22.1B). For each cDNA target synthesis, 20 μg total RNA is reverse transcribed using oligo d(T) primers with 3′ or 5′ unique sequence overhangs for the Cy3 or Cy5 labeling reactions, respectively. Microarrays are prepared for hybridization as previously described (Rise et al. 2004a). Labeled targets are hybridized to microarrays in a formamide-based buffer (25% formamide, 4× SSC, 0.5% SDS, 2× Denhardt's solution) for 16 hours at 52°C. The arrays are washed and dried (Rise et al. 2004a), and the Cy3 and Cy5 3-dimensional fluorescent molecules (3DNA™ capture reagent, Genisphere) are hybridized to the bound cDNA on the microarray; the Cy3 and Cy5 3DNA capture reagents bind to their complementary cDNA capture sequences on the Cy3 and Cy5 oligo d(T) primers, respectively (Figures 22.1 A,B). The second hybridization is 3 hours at 52°C, and microarrays are washed and dried (Rise et al. 2004a). Fluorescent images of hybridized arrays are acquired immediately at 10 μm resolution using ScanArray Express (PerkinElmer). The Cy3 and Cy5 cyanine fluors are excited at 543 nm and 633 nm, respectively, and the same laser power (90%) is used for all slides. Photomultiplier tube (PMT) settings are optimized for each slide to give maximum dynamic range. Microarray data extraction and analysis, and evaluation of microarray signal quality, are discussed in relation to sample microarray experiment No. 2. Figures illustrating microarray experimental design and data analysis are also presented for the second sample experiment.

Sample Microarray Experiment No. 2

The following experiment is designed to identify genes differentially expressed during normal rainbow trout embryogenesis. A similar experiment has been performed in collaboration with Dr. R. H. Devlin (Fisheries and Oceans Canada), and is the subject of a manuscript in preparation. In sample experiment No. 2, GRASP 3.5K microarrays are used to compare global gene expression in several embryonic stages to that of a reference sample (50% hatched). One comparison (blastodisc versus 50% hatched) is illustrated in Figure 22.2, along with an example of results from this comparison. MicroPoly(A) Pure kits (Ambion) are used to prepare mRNA from 20 individuals each of 10 developmental stages from blastodisc to 50% hatched. To have adequate

quantities of template for microarray target syntheses, mRNA from each developmental stage is amplified using the MessageAmp aRNA kit (Ambion) and the manufacturer's instructions (see Web References). Amplified RNA (aRNA) samples are visualized on agarose gels to check quality and quantity. Synthesis of fluorescently labeled cDNA targets uses antisense RNA (aRNA) templates, random primers (Roche), dNTPs containing Cy3-dUTP or Cy5-dUTP (Amersham), RNAguard ribonuclease inhibitor (Amersham), and Superscript II RNase H^- reverse transcriptase (Invitrogen), and follows the method in Rise and others (2004a). Labeled targets are treated with RNase, cleaned using the Qiaquick PCR purification kit (Qiagen), and precipitated overnight using standard methods. Each labeled target is recovered by centrifugation (14,000 rpm, 4°C, 1 hour), washed in 70% ethanol, air dried, and resuspended in 60 μl of hybridization buffer: 50% deionized formamide, 5× SSC, 0.1% SDS, 5 μl of 5 μg/μl oligo dT blocker, 5 μl of 2 μg/μl BSA (Pierce), and 2 μl of 10 μg/μl sonicated human placental DNA (Sigma). Targets are incubated at 96°C for 3 minutes, and then at 60°C until applied to microarrays. Microarrays are prepared for hybridization by washing 2 × 5 minutes in 0.1% SDS, washing 5 × 1 minute in MilliQ H_2O, immersing 3 minutes in 95°C MilliQ H_2O, and drying by centrifugation (514 × g, 5 minutes, in 50 ml conical tube). Microarray hybridizations are run in the

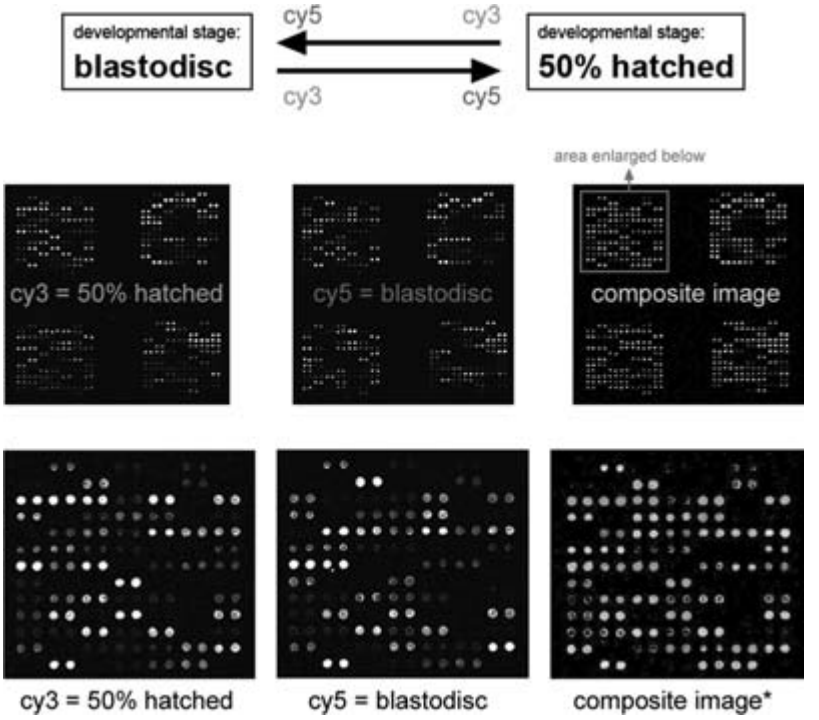


Figure 22.2. Design for a microarray experiment directly comparing global gene expression in two salmonid developmental stages. Arrows symbolize microarrays, with arrow bases showing Cy3-labeled samples and arrow heads showing Cy5-labeled samples. TIFF images of a section of the GRASP 3.5K microarray are shown. (Also see color plate.)

dark under HybriSlips hybridization covers (Grace Biolabs) in slide hybridization chambers (Corning) submerged in a 48°C water bath for 16 hours. Coverslips are floated off, and microarrays are washed and dried as described in Rise and others (2004a).

Fluorescent images of hybridized arrays are acquired immediately at 10 μm resolution using the ScanArray Express (PerkinElmer), or a comparable microarray scanner. Fluorescent intensity data are extracted from Tagged Information File Format (TIFF) images using microarray image processing software such as Imagene 5.5 (BioDiscovery), QuantArray (PerkinElmer), or GenePix Pro (Molecular Devices). Microarray data extraction begins with the creation of a grid representing the two-dimensional arrangement of features on the slide. The TIFF images for both channels from a chip in a two-dye experiment are imported into ImaGene, and the grid is placed over the microarray spots, adjusted, and linked to a gene identification file. ImaGene grids and gene identification files supporting published microarray studies are available on the cGRASP Web site (http://web.uvic.ca/cbr/grasp). Microarray features in areas with spatial effects (i.e., dust particles or high background) may be manually flagged as "marginal," and features with signal values less than a set threshold (such as one standard deviation of mean local background) are automatically flagged as "absent" in ImaGene raw data files.

Microarray data transformation (i.e., background correction, setting background corrected values less than 0.01 to 0.01), normalization (i.e., locally weighted linear regression [Lowess]) (Cleveland 1979), and subsequent analysis (i.e., filtering, clustering) can be performed using proprietary software such as GeneSpring (Agilent Technologies) or GenePix Pro (Molecular Devices), or with the open source software Bioconductor (see Web References). For a review of methods related to microarray data transformation and normalization, see Quackenbush (2002). A scatterplot showing the background-corrected, Lowess normalized (BCLN) data from one microarray (Cy3-labeled 50% hatched versus Cy5-labeled hindbrain swelling developmental stage) in an embryogenesis study is shown in Figure 22.3. Selected transcripts with greater than twofold difference in BCLN signal (expression) between these developmental stages are labeled on the scatterplot (Figure 22.3). Not surprisingly, hatching enzyme-like gene (in capital letters on Figure 22.3) and muscle transcripts such as myosin light chain 2, myosin heavy chain, creatine kinase, and tropomyosin, were greater than twofold up-regulated in the 50% hatched stage fish relative to hindbrain swelling stage fish (Figure 22.3). A hatching enzyme-like gene was selected for QPCR confirmation of microarray results. See the Validation of Microarray Experimental Results section.

The shaded area in Figure 22.3 contains salmonid features with BCLN signal values relatively close to background. As microarray signal data approaches background, variability increases and reproducibility decreases. This higher variability is often illustrated by the increased spread of low-signal data in microarray scatterplots. To evaluate microarray hybridizations, quality statistics may be compiled in Excel from raw fluorescence intensity report files. Data can be sorted by feature type (i.e., salmonid spots, *Arabidopsis* DNA spots, and other control spots for GRASP cDNA microarrays) in Excel. Median signal values from exogenous genome control spots (*Arabidopsis* cDNAs on GRASP chips) may be used to calculate threshold (i.e., mean plus two standard deviations), and mean numbers of salmonid features passing threshold can be determined for Cy3 and Cy5 data separately. The assignment of a

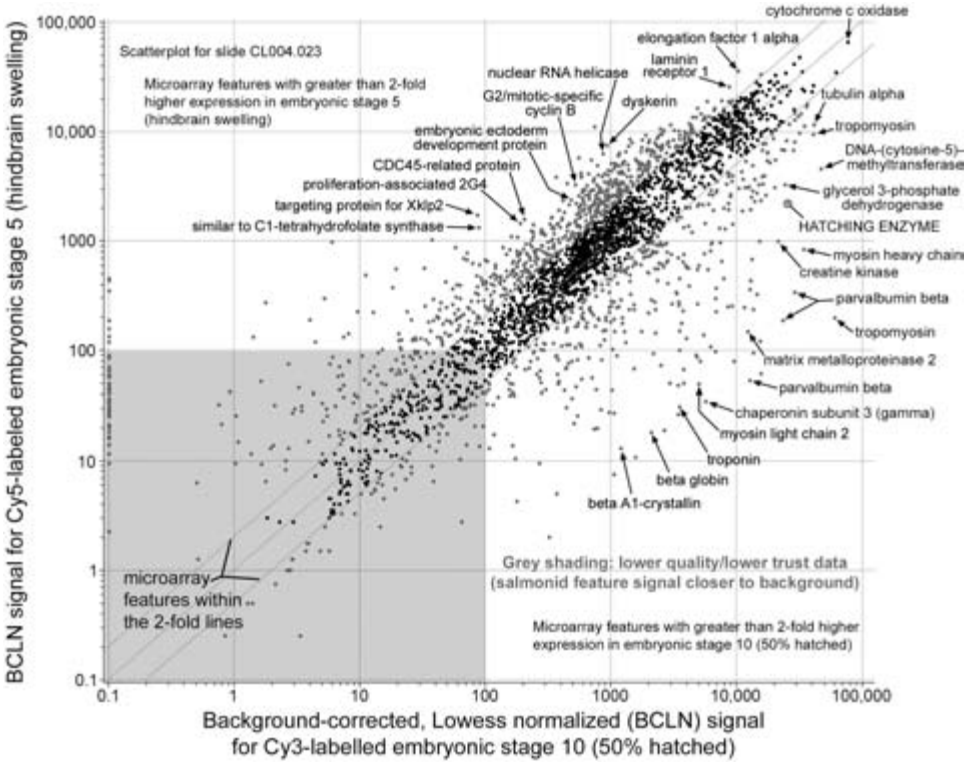


Figure 22.3. Scatterplot showing background corrected, Lowess normalized (BCLN) signal data for one GRASP 3.5K microarray in an embryogenesis study. Select informative genes (differentially expressed in the two developmental stages) are labeled.

threshold based on variability in signal from control features allows the rapid identification of salmonid features with lower trust data.

Fine-tuning Microarray Data Analysis and Interpreting Results

Data and images shown in Figure 22.4 were generated with GeneSpring microarray data analysis tools. Venn diagrams can be used to identify reproducibly informative features on replicate microarrays. For example, using GeneSpring, Venn diagrams identified 98 transcripts that were greater than twofold up-regulated in *Piscirickettsia salmonis*-infected Atlantic salmon macrophages on all three replicate slides of a study (Figure 22.4A) (Rise et al. 2004a). GeneSpring also allows data filtering (i.e., selecting features deemed "present" in ImaGene raw data), and includes clustering methods such as K means, hierarchical, principal components analysis, and QT clustering. In addition, statistical analysis of microarray experimental results may be conducted using GeneSpring. Clustering microarray data can reveal gene expression patterns, providing insight into the molecular mechanisms involved in morphological and physiological processes. With Dr. R. H. Devlin, we used K means (Figure 22.4B) and hierarchical clustering (Figure 22.4C) to analyze data from a GRASP microarray experiment looking at normal global gene expression changes occurring throughout

rainbow trout embryogenesis. One K means cluster from this experiment contains genes that are more highly expressed in early embryos (blastodisc) than in subsequent developmental stages (neural groove through hindbrain swelling) (Figure 22.4B). Hierarchical clustering of filtered data allowed identification of a suite of genes expressed at comparable levels in early (blastodisc) and late (50% hatched) stages, and at lower levels in intermediate stages (i.e., hindbrain swelling, or Stage 5 in Figure 22.4C).

A successful microarray experiment identifies suites of genes that have responded in characteristic ways to an altered experimental parameter (i.e., contact with a pathogen for Figure 22.4A, or time after fertilization for Figures 22.4B and 22.4C).

The next step in a microarray experiment is to use the functional annotations of informative genes to identify altered molecular pathways and biological processes. A salmonid microarray experiment may identify hundreds of reproducibly informative features. Currently, relatively few salmonid nucleotide or amino acid sequences in public data repositories (i.e., GenBank, Swiss-Prot) are functionally annotated. Therefore, the annotations reported for microarray-identified salmonid genes are

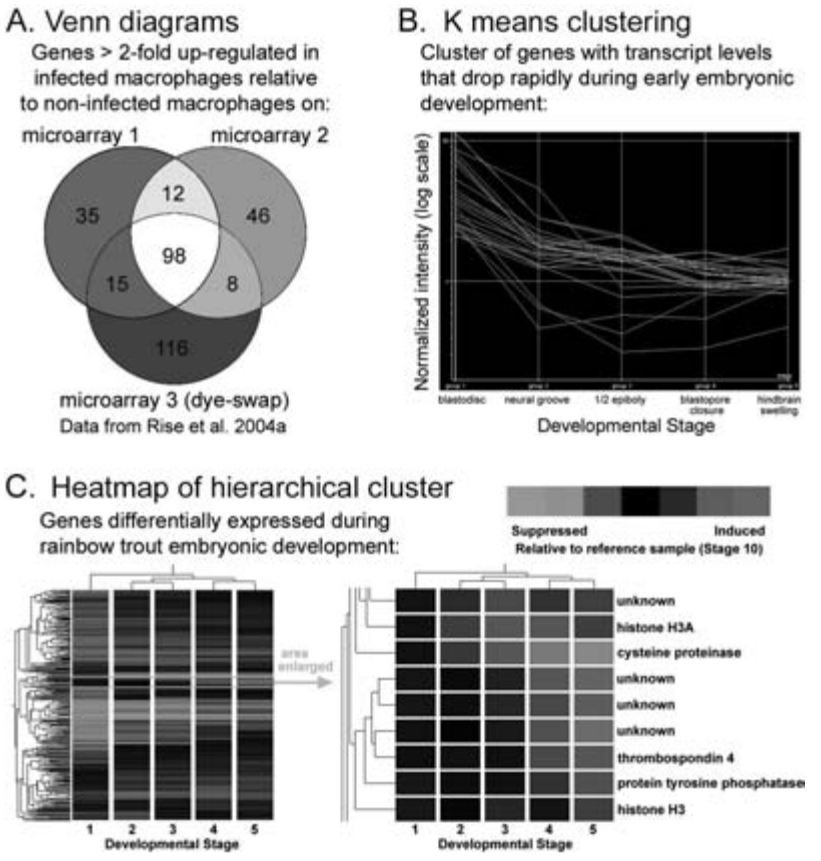


Figure 22.4. Examples of methods for analyzing microarray data. (Also see color plate.)

often those associated with candidate orthologous sequences from well-studied vertebrates such as human, mouse, rat, chicken, and zebrafish (i.e., Tables 1–4 from Rise et al. [2004a]). For each informative chip feature, the top (most negative E value) BLASTN, BLASTX, and tBLASTX hits are analyzed. The extent of aligned sequence, percent identity over the aligned region, and associated E value for a BLAST hit help the researcher to decide if the subject (nonsalmonid) sequence actually represents an ortholog of the query (salmonid) sequence. Functional information on salmonid genes and their putative orthologs may be gathered from public databases such as the Gene Ontology (GO) Database and the UniProt Knowledgebase (see Web References). The GO Consortium, established to develop a standardized language describing gene and protein functions, annotates sequences by molecular function, cellular component, and biological process. Bioinformatics tools such as GenMAPP (Dahlquist et al. 2002), MAPPFinder (Doniger et al. 2003), and PathwayExplorer (Mlecnik et al. 2005) use the functional annotations of differentially expressed genes to identify altered biological pathways.

Prior to submitting reports involving microarray data to peer-reviewed journals, raw (TIFF images) and extracted (i.e., ImaGene files) microarray data should be deposited in a public microarray data repository such as NCBI's Gene Expression Omnibus (see Web References). In addition, quality statistics, TIFF images, a gene identification file, and raw and normalized data files can be made available as online supplemental data supporting submitted manuscripts (for examples, see http://www.uvic.ca/cbr/grasp).

Validation of Microarray Experimental Results

To decrease the influence of biological variation, as well as costs, microarray experiments are often run using pooled RNA samples (i.e., pooled infected spleen RNA from 10 individuals versus pooled noninfected spleen RNA from 10 individuals). In addition, SSH libraries are usually constructed using pooled RNA samples. However, RNA samples from individual fish contributing to the pools should be archived at −80°C for use in experiments designed to validate microarray results. Chauqui and others (2002) present an overview of methods used to validate microarray results, such as quantitative RT-PCR (QPCR), semiquantitative RTPCR, northern blot, and *in situ* hybridization, as well as issues to consider when choosing a validation method. Following a microarray experiment, candidate informative genes are selected for validation. For example, a microarray experiment might identify hundreds of hepatic genes that appear to be reproducibly induced or suppressed in fast-growing fish relative to fish exhibiting normal growth rates. Five to 10 of these candidate informative genes might be selected for validation using QPCR. Both technical replicates (i.e., each reaction run in triplicate) and biological replicates (i.e., 10 different fish for each condition) may be incorporated into QPCR experimental designs. QPCR performed on RNA templates derived from individual fish samples allows assessment of biological variability associated with the expression of genes of interest arising from microarray, SSH, and other studies.

QPCR uses gene-specific primers for informative genes (i.e., genes identified as being differentially expressed between sample types within an experiment) and for a normalizer gene (a gene stably expressed in all samples within the experiment). While there are numerous methods, products, and instruments available for performing

QPCR, they all aim to determine the starting amounts of specific gene sequences within RNA or DNA samples. The simplest and more affordable QPCR methods use fluorescent dyes, such as SYBR Green, that bind double-stranded DNA. QPCR master mixes, such as Brilliant SYBR Green QPCR Master Mix (Stratagene), contain SYBR Green, buffer, $MgCl_2$, dNTPs, and hot-start Taq DNA polymerase. The user supplies gene-specific primers and cDNA templates. We present a sample QPCR method designed to validate microarray-identified Atlantic salmon spleen genes responsive to a viral infection. We also show results of a QPCR experiment (Figure 22.5) designed to confirm an informative gene (hatching enzyme) identified in the embryogenesis microarray experiment depicted in Figures 22.2, 22.3, 22.4B, and 22.4C. The gene encoding HSC71 was shown in microarray experiments to be stably expressed throughout embryonic development, and was therefore deemed to be an appropriate normalizer gene for this QPCR experiment (Figure 22.5). Since the QPCR in Figure 22.5 used pooled RNA templates (n = 20 embryos from each developmental stage), these results do not provide information on biological variability.

Sample QPCR Experiment

Total RNA is prepared individually from flash-frozen infected (n = 10) and noninfected (n = 10) Atlantic salmon spleens from each time point in an experimental infection using TRIzol reagent and methods (Invitrogen). RNA samples are treated with RNase-free DNase (Qiagen) to remove traces of genomic DNA contamination, and cleaned using the RNeasy MinElute kit (Qiagen). For each individual, 2 μg of total RNA are reverse transcribed using 200 ng oligo $d(T)_{16}$ primer. Reverse transcription reactions include 40U RNase inhibitor (Promega), 500 μM dNTPs, 10 mM DTT, and 400U SuperScript II RNase H^- reverse transcriptase (Invitrogen) with the manufacturer's buffer. First-strand cDNAs are diluted 1:10 and used as templates for QPCR analysis. QPCR uses two PCR primers per gene and SYBR Green I dye. Genes of interest (GOI) are selected from the results of microarray experiments. A normalizer gene, with stable expression levels in all infected and noninfected samples at all time points, is selected based on microarray results. Primer pairs for GOI and the normalizer gene are designed from EST FASTA files using Primer3 (see Web References) and the following guidelines: product size 150–250 bp, Tm 60°C ± 1°C, and with at least 3 of the 3′ terminal 6 bases G/C. Twenty-five μl reactions containing 2 μl diluted template, 200 nM each primer, and 1× Brilliant SYBR Green QPCR Master Mix (Stratagene), are run in triplicate using a Mx3000P Real-time PCR System (Stratagene) and the following cycling parameters: 95°C for 9 minutes, then 40 cycles of 95°C for 15 seconds, 52°C for 30 seconds, 72°C for 45 seconds. Controls (no template) and standard curves are run for all primer pairs.

Triplicate threshold cycle values (Ct) for each GOI with each treated (infected) template are first normalized (GOI minus average normalizer gene Ct for the same template). Normalized Ct values from infected and control samples can be compared (converted to fold differences) using the relative quantification method with amplification efficiencies calculated from standard curves (Pfaffl 2001, Pfaffl et al. 2002) or kinetic curves (Liu and Saint 2002). Ct values and calculations are made available as online supplements for associated publications. Melting curves and end-point analysis on agarose gels are run for all GOI and normalizer gene QPCR products to ensure that they are strong, single bands of the expected sizes.

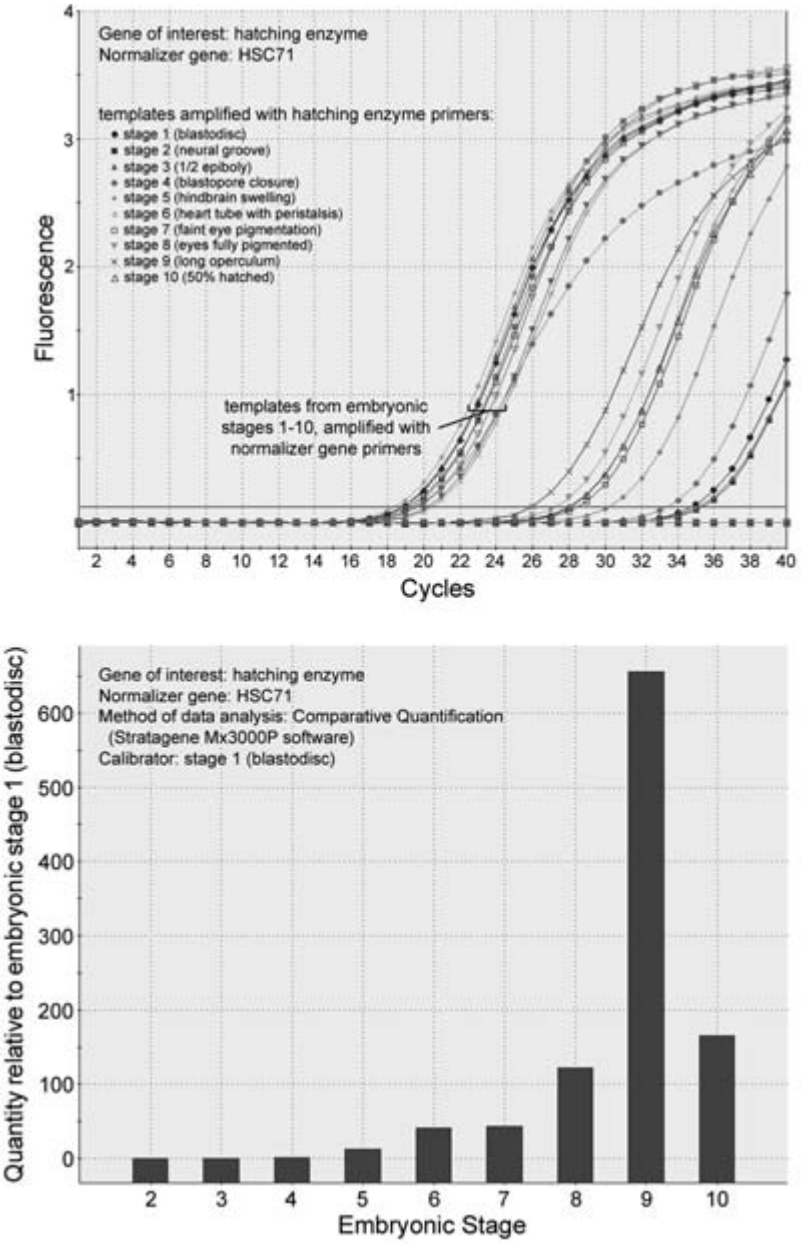


Figure 22.5. Amplification plots and chart showing developmental expression profile of a microarray-identified gene of interest (hatching enzyme) normalized to HSC71. (Also see color plate.)

Other Global Gene Expression Profiling Methods

The two most widely known methods for comprehensive analysis of gene expression in largely uncharacterized genomes are SAGE (serial analysis of gene expression) and MPSS (massively parallel signature sequencing). SAGE has been used for global gene

expression profiling studies with agricultural animals including chicken (Wahl et al. 2004), swine (Blomberg et al. 2005), and cattle (Neill et al. 2006), and MPSS databases have been generated for rice and grapes (Nakano et al. 2006). However, neither of these alternatives to microarray technologies has yet been used in salmonid research. Since this may not always be the case, we present overviews of SAGE and MPSS technologies. SAGE (Velculescu et al. 1995) and MPSS (Brenner et al. 2000a, 2000b) are sequence-based methods that quantify transcripts in a cDNA library. Both techniques generate short (usually 10–14 nucleotide for SAGE, 17–20 nucleotide for MPSS) sequence tags derived from defined regions near the 3′ ends of transcripts (Coughlan et al. 2004).

Serial Analysis of Gene Expression (SAGE)

The following outline of the SAGE method comes from the I-SAGE Kit (Invitrogen) manual (see Web References). This SAGE method involves binding a poly(A)+ RNA sample to oligo(dT) primer-coated magnetic beads, and synthesis of double stranded cDNA on the beads using reverse transcriptase and DNA polymerase. The cDNA molecules are cleaved with an "anchoring enzyme," a restriction enzyme such as *Nla* III with a 4-nucleotide recognition site (therefore cutting approximately every 250 bps). The cDNA samples are divided into two fractions and adapter oligonucleotides containing recognition sequences for a Type IIS restriction endonuclease, and priming sites for amplification by PCR, are ligated onto anchoring sequences. A Type IIS restriction enzyme such as *Bsm*F I, which cuts approximately 13 nucleotides away from the enzyme's binding site, is used to liberate approximately 50 bp (40 bp of adaper sequence plus 10–14 bp cDNA-specific sequence tag) cDNA fragments. The two fractions of cDNA fragments are ligated together, forming approximately 100 bp ditags that are amplified by PCR. The anchoring enzyme is used to liberate 26 bp ditags (each containing 10–14 bp unique sequence tags from two transcripts) that are purified and ligated together into concatenated chains. Concatamers containing 20–50 tags are cloned and sequenced by conventional methods. Transcript abundance is determined by dividing the number of times a transcript-specific tag is identified by the total number of tags sequenced.

Massively Parallel Signature Sequencing (MPSS)

MPSS data sets generally contain over 1 million signature sequences from a single library, providing up to tenfold clone coverage of the transcripts in that library (Stauffer et al. 2004). This depth of transcriptome coverage allows MPSS to provide statistically reliable analysis of genes present at quantities as low as one copy per cell (Jongeneel et al. 2003). MPSS is performed by Lynx Therapeutics (Hayward, CA). Lynx Therapeutics begins the MPSS procedure with at least 100 μg of total RNA for each sample to be analyzed. The processing of samples and generation of MPSS data sets usually take approximately 10 weeks from the date of RNA receipt, and datasets (including signature tag sequences, abundances of signatures in transcripts per million, accession numbers of reference cDNA sequences used to annotate tags, gene annotations, and positions of signatures relative to poly[A] tails of transcripts) are provided by Lynx as tab-delimited files (Jongeneel et al. 2003).

The MPSS method uses Lynx's Megaclone™ technology to construct a cDNA library on the surfaces of microbeads from a poly(A) RNA sample (Brenner et al. 2000a). Poly(A) RNA is used as template for the synthesis of cDNA by reverse transcription with tailed oligo-dT primers. This pool of greater than 1 million original cDNA molecules is digested with *Dpn*II, and the 3-prime *Dpn*II fragments including poly(A) tails are cloned into a plasmid library containing over 16 million different 32-base "address tags" (Tyagi 2000, Constans 2002). Following PCR amplification of the cDNA library (with one primer fluorescently labeled), denatured cDNA-address tag conjugates are hybridized for 72 hours to microbeads. Each microbead is coated with a 32-base "capture oligonucleotide" complementary to an address tag. Approximately 100,000 amplified copies of a particular cDNA fragment hybridize and are ligated to each microbead (Tyagi 2000, Constans 2002). By setting the number of address tag and capture oligonucleotide sequences at greater than 16 million, Lynx ensures that each cDNA molecule in the preamplified library (representing one transcript in the original tissue) is bound to a unique tag and assigned to a unique microbead (Brenner et al. 2000a, Tyagi 2000, Constans 2002). Lynx then separates cDNA-loaded from unloaded microbeads using fluorescence-activated cell sorting (FACS). Approximately 1.5 million fluorescent microbeads are distributed and immobilized in a monolayer at the bottom of a microfluidic flow-cell (Brenner et al. 2000b). Sequencing reagents flow past the stationary microbeads, the fluorescence from which is continuously monitored by imaging. Lynx performs MPSS from each cDNA's *Dpn*II site using proprietary chemistry and instrumentation (Tyagi 2000). On all microbeads in the flow-cell, 20 bp cDNA signature sequences are read by five successive cycles of exposing and sequencing 4-nucleotide segments (Brenner et al. 2000b, Tyagi 2000). Each cycle involves adaptor ligation followed by digestion with a Type II restriction enzyme (e.g., *Bbv* I), characterized by cutting a defined number of bases from the restriction site and leaving 4-nucleotide overhangs (Tyagi 2000). Exposed 4 base-long cDNA segments are sequenced by hybridization to a suite of 1024 "encoded adaptors," followed by successive hybridizations with fluorescent decoder oligonucleotides (Brenner et al. 2000b, Tyagi, 2000). When 20 bp signature tags have been collected on all microbeads in the flow cell, Lynx quantifies tags and translates their abundances to transcripts per million (tpm) values. In addition to the signature sequence and tag abundance dataset for each library, Lynx also provides bioinformatics support including tag identification based on alignments with public and private sequence databases and positioning of tags on cDNAs relative to poly(A) tails.

Applications for Salmonid DNA Microarrays

Salmonids are prominent models for studies involving environmental toxicology, carcinogenesis, comparative immunology, disease ecology, and the molecular genetics and physiology of the stress response, olfaction, vision, osmoregulation, growth, nutrition, and gametogenesis (reviewed in Thorgaard et al. 2002, Rise et al. 2004b). Furthermore, Atlantic salmon and rainbow trout are of particular importance to the global aquaculture industry.

Environmental Toxicant Exposures (Toxicogenomics)

Toxicogenomics is a relatively new field that uses genomic research tools and techniques, such as SSH cDNA library construction and characterization, DNA microarrays, and QPCR, to determine how toxicant exposures impact molecular pathways and biological processes in individuals and populations. Salmonids such as rainbow trout have served as important model organisms for toxicology research (Thorgaard et al. 2002). Salmonids are important in the global aquaculture industry and sports fishery, and therefore represent a way in which bioaccumulated toxicants may impact human populations. Since there are naturally reproducing populations of salmonids throughout the world, this family of fish serves as an ecotoxicogenomic model, bridging the gap between laboratories and the aquatic environment. Existing salmonid microarrays (Table 22.4) contain up to thousands of previously identified salmonid genes, and their use in toxicogenomic studies (i.e., Koskinen et al. 2004b, Krasnov et al. 2005c) provides valuable information on molecular mechanisms of toxicity. However, these genomic resources may be missing some toxicant-responsive salmonid genes. A complete understanding of the molecular mechanisms by which environmental toxicants alter salmonid physiology and development may require, in addition to the use of existing genomic tools, the development of new resources (e.g., targeted SSH libraries and toxicant-responsive EST databases) tailored to toxicogenomic research.

Aquaculture

A key goal of microarray-based research relevant to aquaculture has been to identify genes potentially involved in production-relevant traits such as growth rate (Rise et al. 2006), temperature acclimation (Vornanen et al. 2005), and responses to handling stress (Krasnov et al. 2005b), pathogens (Rise et al. 2004a, Ewart et al. 2005, Morrison et al. 2006, MacKenzie et al. 2006), and vaccines (Martin et al. 2006, Purcell et al. 2006). As an example of this type of work, one might wish to develop a suite of molecular biomarkers for rapid growth potential for use in marker-assisted selection of broodstock with enhanced growth characteristics. This work would potentially focus on tissues involved with growth (i.e., muscle), the regulation of growth (i.e., brain, liver), and blood (nonlethal samples for potential use in selective breeding programs).

The traditional method for identifying candidate genes contributing to quantitative traits involves crosses of inbred strains, and the mapping of QTL by comparing the segregation of the trait to those of multiple genetic markers. Salmonid genetic maps (i.e., Nichols et al. 2003b, Woram et al. 2004, Moen et al. 2004b) have been used to map QTL contributing to production traits such as temperature tolerance (Jackson et al. 1998, Somorjai et al. 2003) and resistance to disease (Nichols et al. 2003a, Rodriguez et al. 2004, Moen et al. 2004a). However, low marker densities on salmonid genetic maps currently prevent the fine mapping needed to identify positional candidate genes for QTL (Rexroad et al. 2005). It is possible that global gene expression profiling methods may augment this approach by identifying heritable gene expression signatures correlated with quantitative traits.

DNA microarrays have been used to identify candidate genes potentially contributing to quantitative traits (Liu et al. 2001, de Koning et al. 2005, Stylianou et al. 2005).

A recent rat microarray study identified coexpressed genes potentially contributing to the quantitative trait hypertension (Hinojos et al. 2005). Genes found to be coexpressed in microarray experiments may have common mechanisms of transcriptional regulation (Thijs et al. 2002). For example, microarray experiments in rice identified suites of genes responsive to cold, drought, high salinity, and abscisic acid treatment, and a set of 15 genes that responded to all four stressors (Rabbani et al. 2003). Analyses of 5-prime sequences of microarray-identified, stress-induced rice genes revealed known and novel *cis*-acting elements relevant to stress response, and may lead to the identification of promoters involved in the agriculturally important quantitative trait, stress tolerance (Rabbani et al. 2003). Microarray studies in *Arabidopsis* have identified coexpressed sets of genes with common *cis*-acting elements in their 5-prime regulatory regions (Oono et al. 2003, Narusaka et al. 2004). There is potential for salmonid DNA microarrays to be used together with QTL mapping studies to identify genes controlling production characteristics such as rapid growth. Computational methods may then be used to scan genomic regions near coexpressed genes for common regulatory motifs (Cora et al. 2004, Yap et al. 2005). Three different salmonid cDNA microarrays were recently used to identify suites of genes that were differentially expressed in growth-enhanced GH transgenic and nontransgenic control coho salmon liver samples (Rise et al. 2006). Hepatic transcripts that were found to be potentially responsive to GH transgenesis had functional annotations related to mitochondrial function, iron homeostasis, metabolism, and cellular proliferation (Rise et al. 2006). Production trait-relevant suites of genes, such as those identified in the GH transgenic coho study, can be placed on the Atlantic salmon physical map using standard molecular techniques (Ng et al. 2005). The integration of genetic and physical maps is a top priority of the international salmonid genomic research community. Once achieved, the map locations of microarray-identified growth-relevant genes may be compared with the results of linkage analyses to identify positional candidate genes for rapid growth QTL. This research will improve understanding of the genetic and physiological causes of complex traits such as growth rate, and potentially lead to the development of molecular methods for selecting individuals with superior production traits. The development of comparative genomic methods such as MONKEY (Moses et al. 2004) and Footer (Corcoran et al. 2005) will also aid this type of research, potentially leading to the identification of conserved *cis*-acting elements controlling traits such as enhanced growth and immune responsiveness.

Health and Disease

Studying the host immune transcriptomic (global gene expression) response to an infection provides information on host-pathogen interactions, and leads to a better understanding of the molecular basis of disease progression. Microarray experiments allow the rapid identification of molecular pathways altered during pathological processes. In agricultural research, DNA microarrays have proven to be valuable tools for studying host immune responses to infection. For example, chicken cDNA microarrays have been used to study the gene expression responses of chicken embryonic fibroblasts to infection by Marek's disease virus (Morgan et al. 2001) and herpesvirus of turkeys (Karaca et al. 2004). Munir and colleagues (2003, 2004) have constructed and characterized SSH cDNA libraries to identify host transcripts

responding to avian metapneumovirus (aMPV) infection, and have used these sequences to build microarrays for continued investigation into the host gene expression response to this viral pathogen. DNA microarrays have been used to identify bovine peripheral blood mononuclear cell transcripts induced or suppressed following infection with the intracellular bacterial pathogen *Mycobacterium avium* (Coussens et al. 2003). The rapidly growing literature involving the use of microarrays to characterize host gene expression responses to infections indicates the widespread acceptance of these reagents and associated methods for studying molecular pathogenesis.

High-density culture in marine net pens is required to maximize productivity and cost-competitiveness in the global salmon market, but also increases the incidence of diseases. Every year, millions of farmed and wild salmonids die as a result of being infected with emerging bacterial and viral pathogens, such as *Piscirickettsia salmonis* and infectious hematopoietic necrosis virus (IHNV), respectively. Salmonid DNA microarrays have been shown to be effective tools for characterizing host immune tissue transcriptomic responses to emerging pathogens. For example, GRASP 3.5K cDNA microarrays were used to identify Atlantic salmon hematopoietic kidney and macrophage genes responding to *Piscirickettsia salmonis* infection (Rise et al. 2004a); Institute for Marine Biosciences (IMB) 4K cDNA microarrays were used to identify Atlantic salmon spleen, hematopoietic kidney, and liver genes responding to *Aeromonas salmonicida* infection (Ewart et al. 2005); and GRASP 16K microarrays have been used to characterize the Atlantic salmon gill transcriptomic response to the parasite causing amoebic gill disease (Morrison et al. 2006). Such studies improve understanding of the host molecular pathways altered during the course of an infection, paving the way for the development of effective diagnostics, vaccines, and therapeutics. GRASP 16K microarrays have also been used to evaluate rainbow trout gill, kidney, and spleen gene expression responses to an IHNV DNA vaccine, shedding light on the molecular mechanisms involved in antiviral immunity in salmonids (Purcell et al. 2006).

It is likely that some salmonid genes involved in responses to various pathogens have not yet been identified. Although microarrays can be used to identify many molecular biomarkers of disease, SSH and other high-complexity cDNA libraries may be useful for identifying novel host genes dysregulated following contact with a pathogen. For example, midway through the GRASP EST project, statistics associated with an Atlantic salmon SSH library enriched for hematopoietic kidney genes induced by *P. salmonis* infection (Rise et al. 2004b) indicated that it was among the best performing GRASP libraries with regard to gene discovery rate. Approximately 66% of the genes identified in this library were not identified in any other library (Rise et al. 2004b), suggesting that this subtracted library effectively identified novel salmonid genes responding to the pathogen *P. salmonis.*

Conservation

Although many government policies are designed to expand aquaculture for the economic and social revitalization of coastal and rural communities, in which opportunities for sustainable development can be elusive, societies are ever mindful of the need

to preserve the environmental integrity and the culture, values, and traditions. Salmonid aquaculture must find ways of minimizing its impact on wild fisheries and the environment. There is a need for scientific information about how farmed salmon interact with their wild counterparts and with the environments in and around their farm sites. The wild salmonid fisheries of North America are currently less than half the size they were 10 years ago, and many wild stocks are considered threatened. Improved conservation measures based on scientific approaches are also badly needed. Genomic studies of salmonids have identified a large number of variable genetic markers (Smith et al. 2005, Hayes et al. 2006). The development and subsequent use of genetic markers for the identification of specific regional stocks are essential for the effective management and sustainability of wild fisheries. The determination of potential harvest location sites and harvest levels in the commercial and sport fisheries are main goals of government agencies and conservation groups. To preserve the long-term health of the species, we need to be able to identify and conserve genetic diversity. Studies to date include looking at parallelism in gene transcription among sympatric lake whitefish ecotypes (Derome et al. 2006), and parallel evolutionary changes of gene transcription profiles of farmed Atlantic salmon in Canada and Norway (Roberge et al. 2006). The ability to identify different populations, local and more global adaptive traits, major biogeographic patterns, and genes and systems that correspond to such patterns, has the ability to revolutionize conservation biology of fish.

Basic Biology

We have previously discussed the use of salmonid microarrays to conduct basic biological research, such as the study of global gene expression changes occurring during the processes of growth and development. Here we provide additional detail on one such application.

Gonadal Development

The development and maturation of the ovary and testis requires precisely coordinated expression of specific gene-classes to produce viable gametes. Endocrine and locally expressed steroids and hormones induce cell growth, differentiation, and maturation of the accessory cells that support the gametes (Erickson and Shimaski 2000, Eppig et al. 2002, Saunders 2003). The assembling support structures, the maturing gametes, and their companion somatic cells undergo cellular remodeling and organization throughout maturation and development. Bidirectional communication between both the gametes and their somatic support cells also mediates reciprocal cell development and function (Erickson and Shimaski 2000, Eppig et al. 2002).

One interesting phenomenon found in a small percentage of juvenile salmon is that they are ready to undergo spawning at least 1 year ahead of their siblings. These precocious males and females undergo dramatic increases in growth and development of their gonads in comparison to their normal (less mature) cohorts. This provides an opportunity to compare and characterize the genes expressed in immature, normal, and precocious reproductive tissues of the same age.

To understand what genes are involved in these dynamic developmental processes, we used 3,557-gene salmonid cDNA microarrays to profile gene expression at three stages of precocious ovary development (June, August, and October) relative to reference (June) normal ovary (von Schalburg et al. 2005a). On average, approximately 240 genes were developmentally regulated during the study period. Some examples of these developmentally regulated genes were CR2 receptor, retinol-binding protein, complement components, immunoglobulins, and various matrix proteinases, elastases, and their inhibitors. Classes of genes maintaining relatively steady-state levels of expression were also identified. Within this group, we found unequivocal evidence for expression of the transcripts that encode the common glycoprotein-α, LHβ, FSHβ, and TSHβ subunits in both the ovary and testis of trout (von Schalburg et al. 2005a). Expression and synthesis of these glycoprotein hormones have been classically considered to be restricted to the pituitary gland.

During this study, the activity of a subset of morphogens in trout gonads was also observed. Morphogens are developmental regulators that modulate different tissue patterning, proliferation, differentiation, or remodeling processes in embryonic and adult tissues, and may evoke specific regulatory programs in stem cells (Lin 2002). In this category, expression of genes that encode anterior gradient-2, BMP-4, epimorphin, flightless, frizzled, GW112, notch, tiarin, twisted gastrulation, and wnt were demonstrated in adult trout gonads (von Schalburg et al. 2006). To understand the potential functional roles of some of these regulators in mature gonadal tissue, we localized their expression to separate mammalian cell-types isolated from ovary (cumulus or mural granulosa cells) and testis (Leydig or Sertoli cells). We observed differences in the expression patterns of some morphogens, which may point to divergent sex-specific (BMP-4, notch-3, and wnt-11) and species-specific (BMP-4, epimorphin) reproductive functional processes (von Schalburg et al. 2006).

Our salmonid-gene specific microarray analyses revealed the changes that occur in the expression of genes involved in tissue remodeling, immunoregulation, cell-cycle progression, apoptosis, and growth during gonadal development. We also found the expression of genes more commonly associated with embryonic development processes, such as anterior gradient-2, epimorphin, flightless, tiarin, and twisted gastrulation for the first time in the adult gonad of any vertebrate. Moreover, this work showed that the salmonid microarrays can serve as useful tools for the detection of unexpected tissue-specific expression of genes.

Future Directions

With respect to providing social and economic benefits, the objectives of genome research on salmon are threefold. First, it harnesses the power of genomics to yield practical benefits for salmonid breeding and aquaculture. Second, it provides a better understanding of how natural populations of salmonids adapt to local conditions, which will benefit agencies that make management decisions concerning stock assessment and harvesting plans. Third, genomic research on salmon helps develop novel methodologies that will enable more sensitive and more accurate environmental monitoring of salmonid populations. Therefore, genomic research on salmon has had

significant economic and social impacts in areas vital for the future of Canada's economy, namely aquaculture, conservation, and the environment.

From the technologicial perspective, as more of the salmonid genome is characterized, resequencing technologies will likely become more important in identifying and characterizing variation. In the very near future we expect that one or two reference genomes will be established. As new resequencing technologies are developing at an extremely rapid pace and the costs are plummeting, it is likely that sample sequencing of RNA may supplant hybridization-based technologies. In addition, since the direct connection between the environment and the genome is at the protein level, proteomics and metabolomics will take an increasingly important role in assessing the interaction between the genome and environment. The rate at which genome science is advancing provides tremendous optimism for solving problems in fish aquaculture, conservation, environmental assessment, and basic physiology and biology. It further provides a new era of collaboration and cooperation as we begin to take on some of the very large fundamental scientific problems in aquatic sciences.

Many of the resources required to carry out genomic studies on salmonids were developed as a result of the Genome Canada-funded GRASP and related projects in Norway, France, the United States, and the United Kingdom. These international efforts resulted in a consortium that will conduct genomic research complementary and relevant to all salmonids. To build on the success of these early efforts, Canada, in collaboration with Norway, France, the United States, and the United Kingdom, has funded the Consortium for Genomic Research on All Salmonids Project (cGRASP), which commenced January 1, 2006.

The goals of cGRASP follow:

1. To expand existing genomic resources for Atlantic and Pacific salmon and trout
2. To develop genomic resources for rainbow smelt, a representative of the common ancestor of salmonids
3. To use the existing and expanded genomic resources as tools to answer fundamental scientific questions, as well as those that are of economic and social importance to aquaculture, conservation, and the environment

We invite participation from the academic, environmental, conservation and business groups in providing a robust genomic resource that can provide the basis for sound decisions in all of these areas.

Acknowledgments

Genomic science is performed by large communities of dedicated scientists who unselfishly provide data and ideas to a greater common goal. We cannot thank everyone but would like to particularly acknowledge Bjorn Høyheim, Stig Omholt, Yann Guiguen, Caird Rexroad, and William Davidson, and the research groups in the Koop, Davidson, and Rise laboratories, whose efforts we have too briefly summarized in this chapter. We are also particularly grateful to our funders, Genome Canada, Genome BC, NSERC, and the province of BC.

References

Aegerter S, D Baron, C Carpentier, F Chauvigné, C Dantec, A Estampes, AS Goupil, A Jumel, I Jutel, D Mazurais, N Melaine, J Montfort, J Bobe, P Chardon, C Chevalet, B Fauconneau, A Fostier, M Govoroun, A Le Cam, F Le Gac, C Klopp, S Panserat, F Piumi, C Ralliére, PY Rescan, and Y Guiguen. 2004. The INRA AGENAE program and the AGENAE EST collections: first results applied to fish physiology research. Abstract in: Functional genomics in fish: from genes to aquaculture. Comp Biochem Physiol A Mol Integr Physiol, 137, p. S137.

Altschul SF, W Gish, W Miller, EW Myers, and DJ Lipman. 1990. Basic local alignment search tool. J Mol Biol, 215, pp. 403–410.

Altschul SF, TL Madden, AA Schaffer, J Zhang, Z Zhang, W Miller, and DJ Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 25, pp. 3389–3402.

Aubin-Horth N, CR Landry, BH Letcher, and HA Hofmann. 2005. Alternative life histories shape brain gene expression profiles in males of the same population. Proc Biol Sci, 272, pp. 1655–1662.

Bayne CJ, L Gerwick, K Fujiki, M Nakao, and T Yano. 2001. Immune-relevant (including acute phase) genes identified in the livers of rainbow trout, *Oncorhynchus mykiss,* by means of suppression subtractive hybridization. Dev Comp Immunol, 25, pp. 205–217.

Birnboim HC and J Doly. 1979. A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucleic Acids Res, 7, pp. 1513–1523.

Blomberg LA, EL Long, TS Sonstegard, CP Van Tassell, JR Dobrinsky, and KA Zuelke. 2005. Serial analysis of gene expression during elongation of the per-implantation porcine trophectoderm (conceptus). Physiol Genomics, 20, pp. 188–194.

Bonaldo MF, G Lennon, and MB Soares. 1996. Normalization and subtraction: two approaches to facilitate gene discovery. Genome Res, 6, pp. 791–806.

Bowtell D and J Sambrook, Eds. 2003. DNA microarrays: a molecular cloning manual. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Brazma A, P Hingamp, J Quackenbush, G Sherlock, P Spellman, C Stoeckert, J Aach, W Ansorge, CA Ball, HC Causton, T Gaasterland, P Glenisson, FC Holstege, IF Kim, V Markowitz, JC Matese, H Parkinson, A Robinson, U Sarkans, S Schulze-Kremer, J Stewart, R Taylor, J Vilo, and M Vingron. 2001. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. Nat Genet, 29, pp. 365–371.

Brenner S, M Johnson, J Bridgham, G Golda, DH Lloyd, D Johnson, S Luo, S McCurdy, M Foy, M Ewan, R Roth, D George, S Eletr, G Albrecht, E Vermaas, SR Williams, K Moon, T Burcham, M Pallas, RB DuBridge, J Kirchner, K Fearon, J Mao, and K Corcoran. 2000b. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol, 18, pp. 630–634.

Brenner S, SR Williams, EH Vermaas, T Storck, K Moon, C McCollum, J Mao, S Luo, JJ Kirchner, S Eletr, RB DuBridge, T Burcham, and G Albrecht. 2000a. *In vitro* cloning of complex mixtures of DNA on microbeads: Physical separation of differentially expressed cDNAs. Proc Natl Acad Sci USA, 97, pp. 1665–1670.

Carothers AM, G Urlaub, J Mucha, D Grunberger, and LA Chasin. 1989. Point mutation analysis in a mammalian gene: rapid preparation of total RNA, PCR amplification of cDNA, and Taq sequencing by a novel method. Biotechniques, 7, pp. 494–496, pp. 498–499.

Chuaqui RF, RF Bonner, CJM Best, JW Gillespie, MJ Flaig, SM Hewitt, JL Phillips, DB Krizman, MA Tangrea, M Ahram, WM Linehan, V Knezevic, and MR Emmert-Buck. 2002. Post-analysis follow-up and validation of microarray experiments. Nature Genet, 32, pp. 509–514.

Churchill GA. 2002. Fundamentals of experimental design for cDNA microarrays. Nature Genet, 32, pp. 490–495.

Cleveland WS. 1979. Robust locally weighted regression and smoothing scatterplots. J Amer Stat Assoc, 74, pp. 829–836.

Constans A. 2002. A new approach to gene expression analysis. The Scientist, 16, p. 44.

Cora D, F Di Cunto, P Provero, L Silengo, and M Caselle. 2004. Computational identification of transcription factor binding sites by functional analysis of sets of genes sharing overrepresented upsream motifs. BMC Bioinformatics, 5, p. 57.

Corcoran DL, E Feingold, J Dominick, M Wright, J Harnaha, M Trucco, N Giannoukakis, and PV Benos. 2005. Footer: A quantitative comparative genomics method for efficient recognition of cis-regulatory elements. Genome Res, 15, pp. 840–847.

Coughlan SJ, V Agrawal, and B Meyers. 2004. A comparison of global gene expression measurement technologies in *Arabidopsis thaliana.* Comp Funct Genom, 5, pp. 245–252.

Coussens PM, CJ Colvin, GJM Rosa, J Perez Laspiur, and MD Elftman. 2003. Evidence for a novel gene expression program in peripheral blood mononuclear cells from *Mycobacterium avium* subsp. *paratuberculosis*-infected cattle. Infect Immun, 71, pp. 6487–6498.

Dahlquist KD, N Salomonis, K Vranizan, SC Lawlor, and BR Conklin. 2002. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nature Genet, 31, pp. 19–20.

Davey GC, NC Caplice, SA Martin, and R Powell. 2001. A survey of genes in the Atlantic salmon (*Salmo salar*) as identified by expressed sequence tags. Gene, 263, pp. 121–130.

de Koning D, O Carlborg, and CS Haley. 2005. The genetic dissection of immune response using gene-expression studies and genome mapping. Vet Immunol Immunopathol, 105, pp. 343–352.

Derome N, P Duchesne, and L Bernatchez. 2006. Parallelism in gene transcription among sympatric lake whitefish (Coregonus clupeaformis Mitchill) ecotypes. Mol Ecol, 15, pp. 1239–1249.

Doniger SW, N Salomonis, KD Dahlquist, K Vranizan, SC Lawlor, and BR Conklin. 2003. MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biol, 4, p. R7.

Eppig JJ, K Wigglesworth, and F Pendola. 2002. The ovarian oocyte orchestrates the rate of ovarian follicular development. Proc Natl Acad Sci USA, 99, pp. 2890–2894.

Erickson GF and S Shimasaki. 2000. The role of the oocyte in folliculogenesis. Trends Endocrinol Metab, 11, pp. 193–198.

Ewart KV, JC Belanger, J Williams, T Karakach, S Penny, SCM Tsoi, RC Richards, and SE Douglas. 2005. Identification of genes differentially expressed in Atlantic salmon (*Salmo salar*) in response to infection by *Aeromonas salmonicida* using cDNA microarray technology. Dev Comp Immunol, 29, pp. 333–347.

Ewing B and P Green. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res, 8, pp. 186–194.

Ewing B, L Hillier, MC Wendl, and P Green. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res, 8, pp. 175–185.

Gahr SA, ML Rise, P Hunt, B Koop, and CE Rexroad III. 2006. Characterization of expressed sequence tags from the pituitary of rainbow trout (*Oncorhynchus mykiss*). Anim Genet, in preparation.

Gerhard DS, L Wagner, EA Feingold, CM Shenmen, LH Grouse, G Schuler, SL Klein, S Old, R Rasooly, P Good, et al. 2004. The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). Genome Res, 14, pp. 2121–2127.

Hagen-Larsen H, JK Laerdahl, F Panitz, A Adzuhubei, and B Høyheim. 2005. An EST-based approach for identifying genes expressed in the intestine and gills of pre-smolt Atlantic salmon (*Salmo salar*). BMC Genomics, 6, p. 171.

Hayes BJ, A Gjuvsland, and S Omholt. 2006. Power of QTL mapping experiments in commercial Atlantic salmon populations, exploiting linkage and linkage disequilibrium and effect of limited recombination in males. Heredity (Epub ahead of print May 10, 2006).

Hinojos CA, E Boerwinkle, M Fornage, and PA Doris. 2005. Combined genealogical, mapping, and expression approaches to identify spontaneously hypertensive rat hypertension candidate genes. Hypertension, 45, pp. 698–704.

Hoheisel JD. 2006. Microarray technology: beyond transcript profiling and genotype analysis. Nat Rev Genet, 7, pp. 200–210.

Hüttenhofer A and J Vogel. 2006. Experimental approaches to identify non-coding RNAs. Nucleic Acids Res, 34, pp. 635–646.

Jackson TR, MM Ferguson, RG Danzmann, AG Fishback, PE Ihssen, M O'Connel, and TJ Crease. 1998. Identification of two QTL influencing upper temperature tolerance in three rainbow trout (*Oncorhynchus mykiss*) half-sib families. Heredity, 80, pp. 143–151.

Jongeneel CV, C Iseli, BJ Stevenson, GJ Riggins, A Lal, A Mackay, RA Harris, MJ O'Hare, AM Neville, AJG Simpson, and RL Strausberg. 2003. Comprehensive sampling of gene expression in human cell lines with massively parallel signature sequencing. Proc Natl Acad Sci USA, 100, pp. 4702–4705.

Jordal AO, BE Torstensen, S Tsoi, DR Tocher, SP Lall, and SE Douglas. 2005. Dietary rapeseed oil affects the expression of genes involved in hepatic lipid metabolism in Atlantic salmon (*Salmo salar* L.). J Nutr, 135, pp. 2355–2361.

Karaca G, J Anobile, D Downs, J Burnside, and CJ Schmidt. 2004. Herpesvirus of turkeys: microarray analysis of host gene responses to infection. Virology, 318, pp. 102–111.

Kendziorski C, RA Irizarry, KS Chen, JD Haag, and MN Gould. 2005. On the utility of pooling biological samples in microarray experiments. Proc Natl Acad Sci USA, 102, pp. 4252–4257.

Koskinen H, A Krasnov, C Rexroad, Y Gorodilov, S Afanasyev, and H Mölsä. 2004a. The 14-3-3 proteins in the teleost fish rainbow trout (*Oncorhynchus mykiss*). J Exp Biol, 207, pp. 3361–3368.

Koskinen H, P Pehkonen, E Vehniäinen, A Krasnov, C Rexroad, S Afanasyev, H Mölsä, and A Oikari. 2004b. Response of rainbow trout transcriptome to model chemical contaminants. Biochem Biophys Res Commun, 320, pp. 745–753.

Krasnov A, H Koskinen, S Afanasyev, and H Mölsä. 2005a. Transcribed Tc1-like transposons in salmonid fish. BMC Genomics, 6, p. 107.

Krasnov A, H Koskinen, P Pehkonen, CE Rexroad, S Afanasyev, and H Mölsä. 2005b. Gene expression in the brain and kidney of rainbow trout in response to handling stress. BMC Genomics, 6, p. 3.

Krasnov A, H Koskinen, C Rexroad, S Afanasyev, H Mölsä, and A Oikari. 2005c. Transcriptome responses to carbon tetrachloride and pyrene in the kidney and liver of juvenile rainbow trout (*Oncorhynchus mykiss*). Aquat Toxicol, 74, pp. 70–81.

Li F and GD Stormo. 2001. Selection of optimal DNA oligos for gene expression arrays. Bioinformatics, 17, pp. 1067–1076.

Lin H. 2002. The stem-cell niche theory: lessons from flies. Nat Rev, 3, pp. 931–940.

Liu HC, HH Cheng, V Tirunagaru, L Sofer, and J Burnside. 2001. A strategy to identify positional candidate genes conferring Marek's disease resistance by integrating DNA microarrays and genetic mapping. Anim Genet, 32, pp. 351–359.

Liu W and DA Saint. 2002. A new quantitative method of real time reverse transcription polymerase chain reaction assay based on simulation of polymerase chain reaction kinetics. Anal Biochem, 302, pp. 52–59.

MacKenzie S, D Iliev, C Liarte, H Koskinen, JV Planas, FW Goetz, H Mölsä, A Krasnov, and L Tort. 2006. Transcriptional analysis of LPS-stimulated activation of trout (*Oncorhynchus mykiss*) monocyte/macrophage cells in primary culture treated with cortisol. Mol Immunol, 43, pp. 1340–1348.

Martin SA, NC Caplice, GC Davey, and R Powell. 2002. EST-based indentification of genes expressed in the liver of adult Atlantic salmon (*Salmo salar*). Biochem Biophys Res Commun, 293, pp. 578–585.

Martin SA, SC Blaney, DF Houlihan, and CJ Secombes. 2006. Transcriptome response following administration of a live bacterial vaccine in Atlantic salmon (*Salmo salar*). Mol Immunol, 43, pp. 1900–1911.

Mlecnik B, M Scheideler, H Hackl, J Hartler, F Sanchez-Cabo, and Z Trajanoski. 2005. PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. Nucleic Acids Res, 33, pp. W633–W637.

Moen T, KT Fjalestad, H Munck, and L Gomez-Raya. 2004a. A multistage testing strategy for detection of quantitative trait loci affecting disease resistance in Atlantic salmon. Genetics, 167, pp. 851–858.

Moen T, B Hoyheim, H Munck, and L Gomez-Raya. 2004b. A linkage map of Atlantic salmon (*Salmo salar*) reveals an uncommonly large difference in recombination rate between the sexes. Anim Genet, 35, pp. 81–92.

Morgan RW, L Sofer, AS Anderson, EL Bernberg, J Cui, and J Burnside. 2001. Induction of host gene expression following infection of chicken embryo fibroblasts with oncogenic Marek's disease virus. J Virol, 75, pp. 533–539.

Moriya S, S Urawa, O Suzuki, A Urano, and S Abe. 2005. DNA microarray for rapid detection of mitochondrial DNA haplotypes of chum salmon. Mar Biotechnol, 6, pp. 430–434.

Morrison RN, GA Cooper, BF Koop, ML Rise, AR Bridle, MB Adams, and BF Nowak. 2006. Transcriptome profiling of the gills of amoebic gill disease (AGD)-affected Atlantic salmon (*Salmo salar* L.): a role for the tumor suppressor protein p52 in AGD-pathogenesis? Physiol Genomics, 26, pp. 15–34.

Moses AM, DY Chiang, DA Pollard, VN Iyer, and MB Eisen. 2004. MONKEY: identifying conserved transcription-factor binding sites in multiple alignments using a binding site-specific evolutionary model. Genome Biol, 5, p. R98.

Munir S and V Kapur. 2003. Regulation of host cell transcriptional physiology by the avian pneumovirus provides key insights into host-pathogen interactions. J Virol, 77, pp. 4899–4910.

Munir S, S Singh, K Kaur, and V Kapur. 2004. Suppression subtractive hybridization coupled with microarray analysis to examine differential expression of genes in virus infected cells. Biol. Proced Online, 6, pp. 94–104.

Nakano M, K Nobuta, K Vemaraju, SS Tej, JW Skogen, and BC Meyers. 2006. Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. Nucleic Acids Res, 34, pp. D731–D735.

Narusaka Y, M Narusaka, M Seki, T Umezawa, J Ishida, M Nakajima, A Enju, and K Shinozaki. 2004. Crosstalk in the responses to abiotic and biotic stresses in *Arabidopsis:* analysis of gene expression in cytochrome P450 gene superfamily by cDNA microarray. Plant Mol Biol, 55, pp. 327–342.

Neill JD, JF Ridpath, and E Liebler-Tenorio. 2006. Global gene expression profiling of bovine immature B cells using serial analysis of gene expression. Anim Biotechnol, 17, pp. 21–31.

Nelson JS. 1994. Fishes of the world, 3rd edition. John Wiley & Sons, Inc., New York.

Ng SH, CG Artieri, IE Bosdet, R Chiu, RG Danzmann, WS Davidson, MM Ferguson, CD Fjell, B Hoyheim, SJ Jones, PJ de Jong, BF Koop, MI Krzywinski, K Lubieniecki, MA Marra, LA Mitchell, C Mathewson, K Osoegawa, SE Parisotto, RB Phillips, ML Rise, KR von Schalburg, JE Schein, H Shin, A Siddiqui, J Thorsen, N Wye, G Yang, and B Zhu. 2005. A physical map of the genome of Atlantic salmon, *Salmo salar.* Genomics, 86, pp. 396–404.

Nichols KM, J Bartholomew, and GH Thorgaard. 2003a. Mapping multiple genetic loci associated with *Ceratomyxa shasta* resistance in *Oncorhynchus mykiss.* Dis Aquat Organ, 56, pp. 145–154.

Nichols KM, WP Young, RG Danzmann, BD Robison, C Rexroad, M Noakes, RB Phillips, P Bentzen, I Spies, K Knudsen, FW Allendorf, BM Cunningham, J Brunelli, H Zhang, S Ristow, R Drew, KH Brown, PA Wheeler, and GH Thorgaard. 2003b. A consolidated linkage map for rainbow trout (*Oncorhynchus mykiss*). Anim Genet, 34, pp. 102–115.

O'Farrell C, N Vaghefi, M Cantonnet, B Buteau, P Boudinot, and A Benmansour. 2002. Survey of transcript expression in rainbow trout leukocytes reveals a major contribution of interferon-responsive genes in the early response to a rhabdovirus infection. J Virol, 76, pp. 8040–8049.

Oono Y, M Seki, T Nanjo, M Narusaka, M Fujita, R Satoh, M Satou, T Sakurai, J Ishida, K Akiyama, K Iida, K Maruyama, S Satoh, K Yamaguchi-Shinozaki, and K Shinozaki. 2003. Monitoring expression profiles of Arabidopsis gene expression during rehydration process after dehydration using ca 7000 full-length cDNA microarray. Plant J, 34, pp. 868–887.

Pfaffl MW. 2001. A new mathematical model for relative quantification in real-time RT-PCR. Nucleic Acids Res, 29, p. e45.

Pfaffl MW, GW Horgan, and L Dempfle. 2002. Relative expression software tool (REST©) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. Nucleic Acids Res, 9, p. e36.

Purcell MK, KM Nichols, JR Winton, G Kurath, GH Thorgaard, P Wheeler, JD Hansen, RP Herwig, and LK Park. 2006. Comprehensive gene expression profiling following DNA vaccination of rainbow trout against infectious hematopoietic necrosis virus. Mol Immunol, 43, pp. 2089–2106.

Quackenbush J. 2002. Microarray data normalization and transformation. Nature Genet, 32, pp. 496–501.

Rabbani MA, K Maruyama, H Abe, M Ayub Khan, K Katsura, Y Ito, K Yoshiwara, M Seki, K Shinozaki, and K Yamaguchi-Shinozaki. 2003. Monitoring expression profiles of rice genes under cold, drought, and high-salinity stresses and abscisic acid application using cDNA microarray and RNA gel-blot analyses. Plant Physiol, 133, pp. 1755–1767.

Rexroad III CE, Y Lee, JW Keele, S Karamycheva, G Brown, B Koop, SA Gahr, Y Palti, and J Quackenbush. 2003. Sequence analysis of a rainbow trout cDNA library and creation of a gene index. Cytogenet Genome Res, 102, pp. 347–354.

Rexroad III CE, MF Rodriguez, I Coulibaly, K Gharbi, RG Danzmann, J DeKoning, R Phillips, and Y Palti. 2005. Comparative mapping of expressed sequence tags containing microsatellites in rainbow trout (*Oncorhynchus mykiss*). BMC Genomics, 6, pp. 54.

Rise ML, SRM Jones, GD Brown, KR von Schalburg, WS Davidson, and BF Koop. 2004a. Microarray analyses identify molecular biomarkers of Atlantic salmon macrophage and hematopoietic kidney response to *Piscirickettsia salmonis* infection. Physiol Genomics, 20, pp. 21–35.

Rise ML, KR von Schalburg, GD Brown, MA Mawer, RH Devlin, N Kuipers, M Busby, M Beetz-Sargent, R Alberto, AR Gibbs, P Hunt, R Shukin, JA Zeznik, C Nelson, SRM Jones, DE Smailus, SJM Jones, JE Schein, MA Marra, YSN Butterfield, JM Stott, SHS Ng, WS Davidson, and BF Koop. 2004b. Development and application of a salmonid EST database and cDNA microarray: data mining and interspecific hybridization characteristics. Genome Res, 14, pp. 478–490.

Rise ML, SE Douglas, D Sakhrani, J Williams, KV Ewart, M Rise, WS Davidson, BF Koop, and RH Devlin. 2006. Multiple microarray platforms utilized for hepatic gene expression profiling of growth hormone transgenic coho salmon with and without ration restriction. J Mol Endocrinol, 37, pp. 259–282.

Roberge C, S Einum, H Guderley, and L Bernatchez. 2006. Rapid parallel evolutionary changes of gene transcription profiles in farmed Atlantic salmon. Mol Ecol, 15, pp. 9–20.

Rodriguez MF, S LaPatra, S Williams, T Famula, and B May. 2004. Genetic markers associated with resistance to infectious hematopoietic necrosis in rainbow and steelhead trout (*Oncorhynchus mykiss*) backcrosses. Aquaculture, 241, pp. 93–115.

Sambrook J and DW Russell. 2001. Molecular cloning: A laboratory manual, third edition. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.

Sangrador-Vegas A, JB Lennington, and TJ Smith. 2002. Molecular cloning of an IL-8-like CXC chemokine and tissue factor in rainbow trout (*Oncorhynchus mykiss*) by use of suppression subtractive hybridization. Cytokine, 17, pp. 66–70.

Saunders PTK. 2003. Germ cell-somatic cell interactions during spermatogenesis. Reprod Suppl, 61, pp. 91–101.

Smith CT, CM Elfstrom, LW Seeb, and JE Seeb. 2005. Use of sequence data from rainbow trout and Atlantic salmon for SNP detection in Pacific salmon. Mol Ecol, 14, pp. 4193–4203.

Soares MB, MF Bonaldo, P Jelene, L Su, L Lawton, and A Efstratiadis. 1994. Construction and characterization of a normalized cDNA library. Proc Natl Acad Sci USA, 91, pp. 9228–32.

Somorjai IM, RG Danzmann, and MM Ferguson. 2003. Distribution of temperature tolerance quantitative trait loci in Arctic charr (*Salvelinus alpinus*) and inferred homologies in rainbow trout (*Oncorhynchus mykiss*). Genetics, 165, pp. 1443–1456.

Stauffer Y, G Theiler, P Sperisen, Y Lebedev, and CV Jongeneel. 2004. Digital expression profiles of human endogenous retroviral families in normal and cancerous tissues. Cancer Immun, 4, p. 2.

Stylianou IM, M Clinton, PD Keightley, C Pritchard, Z Tymowska-Lalanne, L Bunger, and S Horvat. 2005. Microarray gene expression analysis of the Fob3b obesity QTL identifies positional candidate gene Sqle and perturbed cholesterol and glycolysis pathways. Physiol Genomics, 20, pp. 224–232.

Thijs G, K Marchal, M Lescot, S Rombauts, B De Moor, P Rouze, and Y Moreau. 2002. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. J Comput Biol, 9, pp. 447–464.

Thorgaard GH, GS Bailey, D Williams, DR Buhler, SL Kaattari, SS Ristow, JD Hansen, JR Winton, JL Bartholomew, JJ Nagler, PJ Walsh, MM Vijayan, RH Devlin, RW Hardy, KE Overturf, WP Young, BD Robison, C Rexroad, and Y Palti. 2002. Status and opportunities for genomics research with rainbow trout. Comp Biochem Physiol B: Biochem Mol Biol, 133, pp. 609–646.

Thorsen J, B Zhu, E Frengen, K Osoegawa, PJ de Jong, BF Koop, WS Davidson, and B Høyheim. 2005. A highly redundant BAC library of Atlantic salmon (*Salmo salar*): an important tool for salmon projects. BMC Genomics, 6, pp. 50.

Tilton SC, LG Gerwick, JD Hendricks, CS Rosato, G Corley-Smith, SA Givan, GS Bailey, CJ Bayne, and DE Williams. 2005. Use of a rainbow trout oligonucleotide microarray to determine transcriptional patterns in aflatoxin B_1-induced hepatocellular carcinoma compared to adjacent liver. Toxicol Sci, 88, pp. 319–330.

Tilton SC, SA Givan, CB Pereira, GS Bailey, and DE Williams. 2006. Toxicogenomic profiling of the hepatic tumor promoters indole-3-carbinol, 17b-estradiol and b-naphthoflavone in rainbow trout. Toxicol Sci, 90, pp. 61–72.

Tsoi SC, KV Ewart, S Penny, K Melville, RS Liebscher, LL Brown, and SE Douglas. 2004. Identification of immune-relevant genes from Atlantic salmon using suppression subtractive hybridization. Mar Biotechnol, 6, pp. 199–214.

Tyagi S. 2000. Taking a census of mRNA populations with microbeads. Nat Biotechnol, 18, pp. 597–598.

Vasemägi A, J Nilsson, and CR Primmer. 2005. Expressed sequence tag-linked microsatellites as a source of gene-associated polymorphisms for detecting signatures of divergent selection in Atlantic salmon (*Salmo salar* L.). Mol Biol Evol, 22, pp. 1067–1076.

Velculescu VE, L Zhang, B Vogelstein, and KW Kinzler. 1995. Serial analysis of gene expression. Science, 270, pp. 484–487.

von Schalburg KR, SP McCarthy, ML Rise, JC Hutson, WS Davidson, and BF Koop. 2006. Expression of morphogenic genes in mature ovarian and testicular tissues: potential stem-cell niche markers and patterning factors. Mol Reprod Dev, 73, pp. 142–152.

von Schalburg KR, ML Rise, GD Brown, WS Davidson, and BF Koop. 2005a. A comprehensive survey of the genes involved in maturation and development of the rainbow trout ovary. Biol Reprod, 72, pp. 687–699.

von Schalburg KR, ML Rise, GA Cooper, GD Brown, AR Gibbs, CC Nelson, WS Davidson, and BF Koop. 2005b. Fish and chips: Various methodologies demonstrate utility of a 16,006-gene salmonid microarray. BMC Genomics, 15, p. 126.

Vornanen M, M Hassinen, H Koskinen, and A Krasnov. 2005. Steady-state effects of temperature acclimation on the transcriptome of the rainbow trout heart. Am J Physiol Regul Integr Comp Physiol, 289, pp. R1177–R1184.

Vuori KAM, H Koskinen, A Krasnov, P Koivumäki, S Afanasyev, PJ Vuorinen, and M Nikinmaa. 2006. Developmental disturbances in early life stage mortality (M74) of Baltic salmon fry as studied by changes in gene expression. BMC Genomics, 7, p. 56.

Wahl MB, RB Caldwell, AM Kierzek, H Arakawa, E Eyras, N Hubner, C Jung, M Soeldenwagner, M Cervelli, Y Wang, V Liebscher, and J Buerstedde. 2004. Evaluation of the chicken transcriptome by SAGE of B cells and the DT40 cell line. BMC Genomics, 5, p. 98.

Wiseman S, H Osachoff, E Bassett, J Malhotra, J Bruno, G VanAggelen, TP Mommsen, and MM Vijayan. 2006. Transcriptional response to acute stress in rainbow trout livers involves upregulation of genes involved in glucose metabolism. (in review).

Woram RA, C McGowan, JA Stout, K Gharbi, MM Ferguson, B Hoyheim, EA Davidson, WS Davidson, C Rexroad, and RG Danzmann. 2004. A genetic linkage map for Arctic char (*Salvelinus alpinus*): evidence for higher recombination rates and segregation distortion in hybrid versus pure strain mapping parents. Genome, 47, pp. 304–315.

Yap YL, MP Wong, XW Zhang, D Hernandez, R Gras, DK Smith, and A Danchin. 2005. Conserved transcription factor binding sites of cancer markers derived from primary lung adenocarcinoma microarray. Nucleic Acids Res, 33, pp. 409–421.

Zhulidov PA, EA Bogdanova, AS Shcheglov, LL Vagner, GL Khaspekov, VB Kozhemyako, MV Matz, E Meleshkevitch, LL Moroz, SA Lukyanov, and DA Shagin. 2004. Simple cDNA normalization using kamchatka crab duplex-specific nuclease. Nucleic Acids Res, 32, p. e37.

Web Site References

Methods and reagents discussed in the chapter:

http://www.clontech.com/clontech/techinfo/manuals/PDF/PT1117-1.pdf, PCR-Select cDNA Subtraction Kit User Manual (BD Biosciences – Clontech).

http://www.genisphere.com/pdf/array50v2_10_19_04.pdf, Array 50 Kit Manual (Genisphere).

http://www.invitrogen.com/content/sfs/manuals/sage_man.pdf, I-SAGE Kit Manual (Invitrogen).

http://www.lynxgen.com/wt/tert.php3?page_name=mpss, MPSS technology (Lynx Therapeutics).

http://www.ambion.co.jp/techlib/prot/bp_1751.pdf, MessageAmp RNA Amplification for Array Analysis Kit Manual (Ambion).

Salmonid genomic research projects and affiliated facilities discussed in the chapter:

http://www.uvic.ca/cbr/grasp, Genomic Research on Atlantic Salmon Project (GRASP, now called cGRASP for the consortium for Genomic Research on All Salmon Project).

http://www.imb.nrc.gc.ca/programs/microarray/index_e.php, Institute for Marine Biosciences (National Research Council Canada) Microarray Program.

http://www.salmongenome.no/cgi-bin/sgp.cgi, the Norwegian Salmon Genome Project.

http://w3.toulouse.inra.fr/lgc/agena, INRA AGENAE rainbow trout genomic research program.

http://www.sigenae.org/uploads/media/SIGENAE_trout_salmon_assembly.pdf, INRA SIGENAE rainbow trout and salmon EST assembly.

http://www.ars.usda.gov/main/site_main.htm?modecode=19-30-00-00, National Center for Cool and Cold Water Aquaculture, USDA-ARS (location of rainbow trout genome project).
http://www.science.oregonstate.edu/mfbsc/facility/micro.htm, Marine and Freshwater Biomedical Sciences Center, Oregon State University, Microarray Facility Core.
http://prostatelab.org/arraycentre, Gene Array Facility at The Prostate Centre, Vancouver General Hospital.

Genomic data analysis software and other resources discussed in the chapter:

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Taxonomy, NCBI Taxonomy Browser.
http://www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=r_trout, TIGR rainbow trout gene index.
http://www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=salmon, TIGR Atlantic salmon gene index.
http://www.molecularcloning.com, Molecular Cloning, a Laboratory Manual on the Web, Cold Spring Harbor Laboratory Press.
http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi, Whitehead Institute for Biomedical Research.
http://www.ncbi.nlm.nih.gov, National Center for Biotechnology Information (NCBI).
http://www.ncbi.nlm.nih.gov/geo, NCBI GEO (Gene Expression Omnibus).
http://www.mged.org/Workgroups/MIAME/miame_checklist.html, MIAME Checklist.
http://www.phrap.org, PHRAP program for DNA sequence assembly.
http://www.biodiscovery.com/index/imagene, Information on ImaGene microarray image processing and visualization software (BioDiscovery, Inc.).
http://www.silicongenetics.com/cgi/SiG.cgi/Products/GeneSpring/features.smf, Information on GeneSpring microarray data analysis software (Agilent Technologies).
http://www.moleculardevices.com/pages/software/gn_genepix_pro.html, Information on GenePix Pro 6.0 microarray image analysis software (Molecular Devices).
http://www.bioconductor.org, Bioconductor 1.8 (open source software for genomic data analysis).
http://www.geneontology.org, Gene Ontology Consortium home page.
http://ca.expasy.org, Expert Protein Analysis System (ExPASy) proteomics server.
http://www.genmapp.org/introduction.asp, GenMAPP Gene Map Annotator and Pathway Profiler.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 23
Computational Challenges for the Analysis of Large Datasets Related to Aquatic Environmental Genomics

Gregory W. Warr, Jonas S. Almeida, and Robert W. Chapman

Introduction

The condition of the aquatic environment is of interest to humans for a multitude of reasons, ranging from the preservation of threatened species, to the quality of life in communities that receive "ecosystem services" from wetlands, waterways, and oceans, and to the sustainable commercial use of the aquatic environment for fisheries, aquaculture, recreation, and tourism. An excellent discussion of these issues can be found in the Report of the U.S. Commission on Ocean Policy (http://www.oceancommission.gov/documents/full_color_rpt/welcome.html) and in the Millennium Ecosystem Assessment Synthesis Report (http://www.millenniumassessment.org//en/Products. aspx).

There are many different and complementary approaches that can be taken to monitoring the physical, chemical, and biological conditions of the environment. However, when we focus on the implications that environmental conditions may have for human health, the concept of sentinel species is one that has become well established, as reviewed in Golden and Rattner (2003). A sentinel species is generally understood as an organism whose response to changing environmental conditions will provide early warning not only of these changes but also of potential hazards to human health and well being that are emerging in the environment. The monitoring of sentinel species may be carried out to obtain information about the effects of specific contaminants, such as metals (reviewed in Peakall and Burger 2003) or organics that have immunotoxic effects through the aryl hydrocarbon receptor (AHR) pathway (Blanco and Cooper 2004). The use of sentinel species for emerging infectious diseases in the environment is also well established (Reif et al. 2006, Wolfe et al. 1998) and the role of wild waterfowl in the ecology of the H5N1 strain of avian influenza is particularly topical as discussed in two recent publications (Webster et al. 2005, Sturm-Ramirez et al. 2005). The assessment of the health of sentinel species can also be used in more integrated, global approaches to measuring environmental quality (Moore et al. 2004, Tom and Auslander 2005).

Ecogenomics

The development, in the last 2 decades, of the "omic" sciences (genomics, transcriptomics, proteomics, lipidomics, chomics, metabonomics, etc., as discussed in Chapter 1)

has presented the biomedical research community with the opportunity to study organisms at a completely different scale from what was used in the past. Instead of looking at the expression of individual genes, proteins, and pathways, or using cellular assays tailored to investigate a specific phenomenon, the omics technologies present the opportunity to take a global perspective on cellular or organismal responses. In the biomedical field, the use of transcriptomics (the assessment of messenger RNA [mRNA] levels as a measure of relative gene expression) and proteomics (the assessment of differentials in protein expression) has been a goldmine of discovery. In the case of cancer, the application of transcriptomic methods has improved not only the classification of tumors but has also permitted the refinement of diagnostic, prognostic, and therapeutic methods (Sanchez-Carbayo et al. 2006, Espinosa et al. 2006, Chi et al. 2006, Pusztai and Gianni 2004, Gianni et al. 2005). The range of fields in which transcriptomic and proteomic approaches are having an impact is enormous, and in this regard the study of ecosystems, especially in the context of environmental toxicology, is no exception. Traditional toxicological assays are often focused on measuring either the levels of a specific contaminant (or family of contaminants) or the biological effects of a similarly restricted set of contaminants or stressors. The global nature of transcriptomic and proteomic approaches is particularly suited to environmental studies, where a diverse range of multiple stressors can be present simultaneously, and the global assessment of the cellular/organismal response offers the potential to assess not only the individual impacts, but the interactions, of multiple stressors on populations of organisms. The power and potential of the -omic approaches to the environmental sciences is recognized by the substantial resources invested by the National Institute of Environmental Health Sciences (NIEHS) in the field of toxicogenomics, including the establishment of a National Center for Toxicogenomics (http://www.niehs.nih.gov/nct/home.htm), and by the emergence of a broader field termed ecogenomics (Chapman 2001) that seeks to integrate environmental -omic information to develop diagnostic and prognostic models of ecosystem health. Although genomics and proteomics are legitimate research disciplines in their own right, their impact on the environmental sciences stems from their value as techniques that expand enormously both the type and quantity of information accessible to investigators. The omic approaches can be applied at differing scales and levels of complexity (and expense!) to environmental research. The topic of this chapter is aquatic environmental genomics, and the impact of genomic approaches (and their often complementary proteomic approaches) to the environment will be considered in the following sections from several different perspectives.

ESTs and Novel Gene Discovery

Although its relevance to environmental genomics may not be obvious, we should not overlook the impact of the simple process of investigating and annotating the transcriptomes of aquatic species. (Also see Chapter 20.) This effort results in a broad expansion of our knowledge, as exemplified by the substantial collections of Expressed Sequence Tags (EST) for aquatic organisms that are accessible in the public databases, including www.marinegenomics.org. The EST collections provide a huge amount of sequence information that is valuable for (among others) studies of

genetic polymorphism and molecular evolution. EST collection also facilitates the identification of novel genes, and it is worth noting that in this respect it is exposing our enormous ignorance of biological diversity. When EST collections are made in organisms (such as crustaceans or mollusks) that represent major taxonomic clades that have previously been little explored, we find that the proportion of unigenes in the collection that are not only novel but also have no known homologues in the sequence databases can range to as high as 50% or more, as illustrated for shrimp and oysters (Jenny et al. 2002, Gross et al. 2001). Surprisingly, this is the case even when we compare species in the same phylum, for example shrimp (Class Crustacea) and *Drosophila* (Class Insecta) as documented at www.marinegenomics.org. EST collections provide the bioinformatic (and in some cases the physical) basis from which microarrays can be designed, constructed, and deployed as environmental biosensors (Chen et al. 2004). The number of important aquatic organisms for which major EST collections are now available is substantial. By far the largest and most complete sets of these are for the bony fishes, including salmonids, catfish, zebrafish, Fugu, medaka, flounder, and cod (see for example, the TIGR Gene Indices (http://www.tigr.org/tdb /tgi/; the Joint Genome Institute http://genome.jgi-psf.org/Takru4/Takru4. home.html; Codgen http://www.codgen.olsvik.info/; consortium for Genomic Research on All Salmon Project http://web.uvic.ca/cbr/grasp/), but the marine mammals such as the bottlenose dolphin (www.marinegenomics.org) and cartilaginous fishes are also represented (http://www.mdibl.org/research/skategenome.shtml), as are many groups of marine invertebrates, such as the urochordates (*Ciona intestinalis,* http:// genome. jgi-psf.org/Cioin2/Cioin2.home.html) and the crustacean and molluscan species (www. marinegenomics.org; http://www.mdibl.org/%7Edtowle/DNA/DNA Facility.htm).

The functional genomics initiative spearheaded by the Hollings Marine Laboratory at Charleston, South Carolina, includes an integrated bioinformatics component, hosted at http://marinegenomics.org (McKillen et al. 2005). Functional genomics requires a level of computational support that spans almost all tasks in the experimental program. It ranges from processing the raw data, such as the images captured from expression arrays, designing the sensor, such as the selection of complementary DNA (cDNA) probes to be spotted, to the calibration of the signal. Even final synthesis where putative pathways are identified from the data cannot be efficiently pursued without close interaction with experimentation. Consequently, Web-based infrastructures, such as the resource www.marinegenomics.org (MG), need to be assembled as an integral part of any functional genomics program, and the aquatic environment is no exception. MG also serves an important role in dissemination of the data, both by holding it as a clearinghouse and preparing it for final deposition in reference databases at the National Center for Biotechnology Information (NCBI), such as GenBank and Gene Expression Omnibus. Deposition in reference public databases is invariably the ultimate objective of these bioinformatic resources. Rather than competing with that more encompassing deposition of molecular biology data, these bioinformatic resources serve the important role of mediating domain-specific data preprocessing and data analysis. Typically, when functional genomics programs are pursued in the absence of this sort of intermediate, the quality of both the data available in GenBank and the ability to use it effectively within the domain that originated it, suffers considerably. MG currently includes results for more than 30 species, 117,000 EST sequences, and 1,500 expression microarrays.

Expanding Knowledge in a Defined Area

Genomic and proteomic approaches can benefit aquatic environmental sciences through research that is tightly focused on a topic area. Examples in the context of the aquatic environment deal with the combined application of transcriptomic and proteomic approaches to expand our knowledge of the metallothioneins involved in the environmental response of the oyster to heavy metals (Jenny et al. 2002, Jenny et al. 2004), the responses of oysters to immune challenge (Gueguen et al. 2003, Gueguen et al. 2006b), and the nature of responses to infectious disease in marine crustacea (de Lorgeril et al. 2005, Destoumieux et al. 1997, Gueguen et al. 2006a, Patat et al. 2004, Supungul et al. 2004).

Microarrays for Ecogenomics

Genomic approaches to studies of the aquatic environment hold the potential to revolutionize the field, through the deployment of gene microarrays as environmental biosensors (Almeida et al. 2005). This field is still in its infancy, but several groups have realized the potential (Gracey and Cossins 2003, Almeida et al. 2005, Ju et al. 2002, Rise et al. 2004, von Schalburg et al. 2005, Cossins and Crawford 2005, Oleksiak et al. 2002, Oleksiak et al. 2001, Renn et al. 2004). Most studies published to date using microarrays in investigations of the response of organisms to the aquatic environment have used teleost fish (see for example, Ju et al. 2002, Gracey et al. 2004, Gracey et al. 2001, Aubin-Horth et al. 2005a, Aubin-Horth et al. 2005b, Oleksiak et al. 2002, Oleksiak et al. 2001, Cossins and Crawford 2005). In our own studies, the Marine Genomics Group in Charleston has produced microarrays for bottlenose dolphin, the oyster (*Crassostrea gigas* and *C. virginica*), and the Pacific whiteleg shrimp, *Litopenaeus vannamei.* The construction and characteristics of these microarrays and their potential value in studies of the aquatic environment will be described in other sections of this chapter, but here we can note that the application of transcriptomic approaches to environmental and ecological research poses some unique problems.

Importance of Negative Data

Transcriptomic analysis in fields such as toxicology, infectious disease, and cancer is conventionally focused on the genes whose expression shows the greatest or most significant differential regulation. Genes that are unregulated are considered uninformative. However, although this is typically quite appropriate in experimental biomedicine where mechanistic molecular science is preeminent, the situation is different in an environmental context, where a lack of change in the transcriptomic signature will usually be very important information (Chapman et al. 2006).

How Many Genes is Enough?

A focus on the most significantly regulated genes revealed in a microarray experiment does not take advantage of the most powerful property of transcriptomic analyses,

which is the opportunity to evaluate the whole transcriptome (or at least that portion of it that is represented on the microarray). A major problem in the analysis of overall transcriptomic patterns (as opposed to the extraction from them of information about individual genes) has been the necessity to deal with massive amounts of data. This has been approached in various ways, the most tractable of which is just to finesse the issue, by a reduction in size of the dataset that needs to be considered. This can be effective if it is appropriately knowledge based. For example, if we can with confidence define the genes (or networks of genes), such as in Calvano and others (2005), that are involved in the phenomena of interest to investigators, such as the response to specific environmental contaminants, then the analysis of transcriptomic signatures can be greatly simplified. This approach represents one of the major directions for research in humans and the well-studied biomedical models of human disease, where whole genomes and transcriptomes have been defined, and millions of man-years of research have generated a well-defined (if still incomplete) understanding of the biochemistry and physiology of humans and their biomedical surrogates. The refinement of the field can be appreciated from the large number of pathway-specific microarrays available from commercial suppliers such as SuperArray (http://www.superarray.com/ArrayList.php) who offer in excess of 70 arrays custom designed to facilitate research in a much focused area. However, such a focused approach is not readily applicable to studies of aquatic environmental genomics, because the necessary knowledge base is missing. Many of the organisms (especially the invertebrates) of critical interest in the aquatic environment are poorly understood in terms of their cell and molecular physiology. We do not have in hand the full transcriptome of most aquatic organisms of interest, and an additional confounding observation is that, of the genes that we have sequenced, we have no inkling of the function of up to 50% of them (i.e., they have no homologues in the databases). Thus, for the effective application of ecogenomics to the study of aquatic organisms, it makes sense to start by asking questions that can be answered without a deep base of molecular, cellular, and physiological knowledge. If we restrict ourselves to asking questions about how overall patterns of gene expression change, then the precise identity and function of the genes whose differential expression defines such patterns is irrelevant. Although we may wish, retrospectively, to identify the function of genes that have been identified as significant components in the response to changes in the aquatic environment, this is not a required component of an effective and informative project in ecogenomics. There are many ways in which complex patterns such as transcriptional signatures can be analyzed, including fractal geometry (Cazalis et al. 2004, Wang et al. 2005, Chapman et al. 2006) and machine learning approaches such as Support Vector Machines (Vapnik 1998), Artificial Neural Networks (Khan et al. 2001), and Genetic Algorithms (Holland 1975). These will be discussed later in this chapter, with respect to the specific analytical methods the Charleston Marine Genomics Group has implemented in our approach to ecogenomics.

One Man's Poison

As ecogenomics and toxicogenomics emerge as major fields of study in the aquatic environment, it will be necessary to examine the "portability" of conclusions from one

species to another. This is not such a problem when microarrays are used purely as environmental biosensors, but it comes to the fore when the species that are being studied are used as "sentinels," or indicators of potential hazards. How might the impact of a given stressor, as monitored through changes in, say, the transcriptomic signature of oyster gills, translate into potential impacts on humans swimming in the ocean or eating the oysters? On the one hand, this issue is an instance (albeit a complex one) of comparative physiology and toxicology, but it also has the potential to fall within the area of public policy and natural resource management. In principle the problems to be solved are no different from the ones facing environmental toxicologists, but in practice the technical complexity and sheer mass of information that is derived from ecogenomics will pose unique challenges to interpretation.

Analysis and Data Reduction

Regardless of the intent or species or environment, analysis of microarrays confronts three conceptual and three technical issues. The conceptual issues are (1) whether the objective is to identify individual genes whose expression changes in response to a particular stress, (2) whether the objective is to understand global changes in expression, the gestalt if you will, or (3) the objective (and the most ambitious one!) is to understand the network connections or regulatory pathways. The technical issues are (1) normalization and calibration of individual hybridization measurements so as to take into account the effect of experimental noise stemming from individual variation, labeling, and hybridization differences on the assessment of gene expression, (2) extracting signals from massively paralleled data, and (3) reduction of the information to its essentials.

The Identification of Significant Signals

By far the bulk of microarray studies have focused upon the identification of transcripts that are significantly changed by some stressor (e.g., Waters et al. 2003, Rieger et al. 2004, Williams et al. 2003). In some respects, this can be likened to *The Hunt for Red October* where the sonar operator (scientist in this case) is trying to identify faint, long period signals in a cacophony of noise. In the main this is a hunt for biomarkers of particular stressors that can uniquely identify the specific stressors that are perturbing the transcript profiles. This approach has been successful in fields ranging from toxicology to cancer and beyond (e.g., Gracey and Cossins 2003, Gracey et al. 2004, Espinosa et al. 2006, Chi et al. 2006, Aubin-Horth et al. 2005a, Oltvai and Barabasi 2002). In our view, this search is a valuable contribution to (1) the field of biomedicine, (2) our understanding of the specifics of the response to a particular stressor, and (3) the elucidation of some aspects of the transcriptional responses to stress. However, in many cases, the observed elicited response is consistent with what we might have predicted based upon our understanding (albeit limited) of biochemical processes (Williams et al. 2003, Rieger et al. 2004, Gracey and Cossins 2003, Gracey et al. 2004); thus, this approach is making only inefficient use of the information presented to investigators in microarray experiments.

Global Signatures

Living systems, whether they are gene regulatory networks, biochemical or metabolic networks, neural networks, or energy and nutrient flows in ecosystems can be represented as relatively simple networks of connections that describe the flow of information (or flux) between nodes. It has been suggested (Chapman 2001, Oltvai and Barabasi 2002) that we (collectively, as biologists) have spent too much time tearing these connections apart, trying to understand the inner workings of the components, while paying insufficient attention to recognizing the emergent properties of these systems that cannot be deduced from component behaviors. The challenge of the global approach is to understand the transcriptome as a whole and is exactly the opposite of studies aimed at identifying significantly up or down regulated genes. It is aimed at understanding the integrated function of the genome, how it functions as a system, and how it responds to internal and external stimuli and to stressors. It is important to note that we are not the first to recognize this, because this conceptual approach extends at least as far back as Sewall Wright (Wright 1831) where networks and epistatic (nonlinear) interactions of genes are prominent components of his Shifting Balance Theory of Evolution.

Network Connections and Regulatory Pathways

Our understanding of enzymatic pathways is extensive and has been built up from thousands of studies of individual enzymes and their properties (cf. http://www.expasy.org/). Only within the past 2 decades have we begun to unravel the complex issues of the regulation of transcription in higher organisms, and only within the last 5 years or so have we begun to appreciate the importance of RNA, particularly microRNAs in this process. In the main this effort has followed the classical approach of one gene (knockout or suppression) at a time, which has proved such a successful approach to defining the structural and functional units of living systems. Although this approach does much to elucidate the basic structure of regulatory networks, it lacks the dynamical aspects of even simple enzyme kinetic theory. Hence, we cannot pass easily to Biochemical Systems Theory (Voit 2000) to study the behavior of the intact components. Recently however, several mathematical approaches have been employed to assess the topology of networks using static data (Liebovitch et al. 2004, Shehadeh et al. 2006, Chapman et al. 2006). It has been shown (Voit and Almeida 2004) that time series information coupled with appropriate computational techniques can elucidate not only the structure of metabolic networks, but important aspects of their dynamics. Whether this approach can be extended to gene networks, and their response to environmental stressors, remains to be seen.

Normalization

A variety of means to standardize raw intensity signals from microarrays have been developed by considering different degrees of stringency in their parametric

assumptions. The stronger the parametric assumption the higher the power to distinguish apparent differential expression but also the higher the risk that the assumptions are not correct and significance is incorrectly assigned. At the two extremes of this range are linear transformations of the data and fully nonparametric empiric approaches such as the quantile normalization. These techniques suffer two major liabilities:

1. Environmental stressors have been shown to reduce the overall rate of transcription (van de Peppel et al. 2003).
2. They are restricted to the dataset in hand, making comparisons between studies (metaanalyses) problematic.

Normalization methods, by their nature, can submerge any overall suppression of transcription and thus fail to recover what might reasonably be considered important data (Altman 2005). Thus, if a transcriptional signature does not change between experimental and control organisms, but the experimental group shows an overall downshift in transcription rates, how are we to recognize and interpret this information? We have no clear-cut answers to this issue because it pertains to individual genes, but will offer an alternative for analysis of the transcriptome as a whole.

Extracting Signals from Noise

The extraction of information from microarray data entails the detection of signals in massively paralleled datasets. In this task, analysis of variance (ANOVA) and other traditional statistical tools are greatly handicapped. For example, the number of samples needed to generate sufficient degrees of freedom to analyze even the univariate associations is defined as the number of genes represented on the arrays plus one. Taking the analysis one step further, to assess additive interactions between genes is not even imaginable, even for a modest metabolic network, using standard statistical approaches. Nevertheless, it is the extraction of information about the dynamic interactions of genes that can be contained in microarray data that constitutes the attraction (perhaps even the Holy Grail) of the technology. It is almost certain that these interactions between genes will be nonlinear in nature and hence inaccessible by linear statistics. We will suggest alternative approaches below.

Data Reduction

The third obstacle to understanding microarray data can be the reduction in the amount of data to the level that is necessary to recover the essential information. Most –omics approaches generate many thousands of data points, many of which may be redundant and not essential to understanding the process in question. Hence, the data could be pruned to the point where the data are comprehensible, but not to the point where essential information is lost. How can we accomplish a reduction in the information to a point at which it still fully represents the original response but

is, at the same time, sufficiently compressed that we can readily comprehend the information about the system?

In addressing these problems it is useful to consider the usual means by which we assess expression profiles. In the main, this is done by plotting in two dimensions, the expression data (raw data, background subtracted data, or ratios of signals) for controls versus controls and for controls versus challenged organisms. From this graphical analysis we can extract the probabilities that the expression of individual genes lies outside the statistical confidence limits of random associations.

However, as discussed above, although the identification of individual transcripts that respond to a given stimulus or stressor is of value, it also conveys very limited information. If change in expression of individual genes is an alphabet, then their concerted and dynamic interactions are the words and syntax that convey a much fuller meaning. It is only within the context of changes in the overall transcriptional signature that individual changes in gene expression can be appropriately understood. Some information about linkages in gene expression can be inferred from the widely used clustering algorithms, but these methods are based in Euclidean geometry or are otherwise limited to linear associations. If, as many others and we believe, and some have shown, there are nonlinear associations in transcriptional cascades, these linear approaches are at best crude approximations of functional genomic complexity.

Several analytical approaches that do not rely on implicit linear interactions to identify potentially interacting suites of genes have been used. The first are artificial neural networks (ANN) or other machine learning algorithms, which have been widely applied to a broad range of massively paralleled datasets including microarrays (Khan et al. 2001, Tarca et al. 2005), genetic polymorphisms at many loci (Motsinger et al. 2006), and mass spectrometry data (Mian et al. 2005, Levner 2005), among many others. It is not our intent to review this voluminous literature, but rather to focus upon machine learning algorithms that show promise in feature selection. ANNs generate, as part of their output, sensitivity values. These are estimates of the impact of specific input variables (for example changes in expression of specific genes) on the outputs. The exact definition is that sensitivities are the slopes (derivatives) of the inputs relative to the outputs, but they also are analogous to the proportion of total variance attributable to a particular input in traditional statistics. In other words, they indicate how sensitive the output is to changes in a particular input variable. As such they can easily serve as a logical basis for selecting inputs (e.g., suites of genes) for subsequent analysis, without losing critical information. The selected genes would, for example, be used in a second round of model development using bootstrapping and data sequestration approaches to examine various model solutions.

Another means of selecting suites of genes for analysis that employs ANN's coupled to Fuzzy logic has recently been advanced (Chen et al. 2006). In essence this approach uses the weights from a two-layer feed forward network to cluster genes according to their relative weights on the hidden nodes of an ANN. Fuzzy logic is used in an all-or-none decision rule algorithm to select the most important genes. The decision rule in this approach is based upon an impact factor (IR), which is basically the impact that changes in the selected gene have upon the network weights.

A third approach, has recently been developed by Chapman and others (2006) that employs fractal geometry to examine the overall shape of complex datasets and

determine how many features of the data are necessary to reconstruction of the original figure. Although the method does not identify optimal sets, it does select significantly up- or down-regulated genes and genes that are necessary to preserve the geometry of the original set. When combined with genetic algorithms (Jirapech-Umpai and Aitken 2005), optimal gene set selection may be achieved, although the necessary algorithms and computer tools to verify this assertion have yet to be developed, much less tested. This is a field ripe for exploration (Chapman et al. 2006).

The reader has probably deduced, from the above discussion, that there are no hard and fast rules about the "proper" way to analyze microarrays or other massively paralleled datasets. This situation contrasts markedly with the well-understood application of traditional linear statistical methods to datasets with only a handful of variables. This is not unexpected as linear statistics have been developed over the past two centuries (at least since Gauss) and is founded in Euclidean geometry that is more than 2,000 years old. Pattern recognition approaches have only become possible with the development of high-speed computers and the understanding that non-Euclidean geometries (fractals) existed. The development of the silicon chip computers and fractal geometry began nearly at the same time and both are less than 5 decades old. Hence, we should not be surprised that standard methods for global transcriptomic analysis are not universally recognized. We do not know how this will play out in the near future, but the one thing we are sure of is that the statistical tools that served biologists in the twentieth century, will not suffice in the twenty-first century. Our technology will force upon us a more team-oriented approach to exploit the wealth of information. Indeed, it already has. Research teams seeking to apply functional genomics techniques to the study of the environment will need to include investigators skilled in a wide range of disciplines if they are to be successful in developing useful diagnostic and prognostic models that will help us understand the environment.

Acknowledgments

The authors would like to thank the numerous colleagues that helped shape the viewpoint expressed in this paper. Special thanks are due to Drs. Paul Gross, Javier Robalino, Matthew Jenny, Charles Cunningham, Nuala O'Leary, and Annalaura Mancia. This paper is contribution #601 of the South Carolina Marine Resources Center, and #36 of the Marine Biomedicine and Environmental Sciences Center. The work was supported in part by the NOS/NOAA Oceans and Human Health Center of Excellence at the Hollings Marine Laboratory.

References

Almeida, JS, DJ McKillen, YA Chen, PS Gross, RC Chapman, and G Warr. 2005. Design and calibration of microarrays as universal transcriptomic environmental biosensors. Comparative and Functional Genomics, 6, pp. 132–137.

Aubin-Horth N, CR Landry, BH Letcher, and HA Hofmann. 2005a. Alternative life histories shape brain gene expression profiles in males of the same population. Proc Biol Sci, 272, pp. 1655–62.

Aubin-Horth N, BH Letcher, and HA Hofmann. 2005b. Interaction of rearing environment and reproductive tactic on gene expression profiles in Atlantic salmon. J Hered, 96, pp. 261–78.

Blanco GA and EL Cooper. 2004. Immune systems, geographic information systems (GIS), environment and health impacts. J Toxicol Environ Health B Crit Rev, 7, pp. 465–80.

Calvano SE, W Xiao, DR Richards, RM Felciano, HV Baker, RJ Cho, RO Chen, BH Brownstein, JP Cobb, SK Tschoeke, C Miller-Graziano, LL Moldawer, MN Mindrinos, RW Davis, RG Tompkins, and SF Lowry. 2005. A network-based analysis of systemic inflammation in humans. Nature, 437, pp. 1032–1037.

Cazalis D, T Milledge, and G Narasimhan. 2004. Probe Selection Algorithms. SCI Conference, Orlando, FL.

Chapman RW. 2001. EcoGenomics—a consilience for comparative immunology? Dev Comp Immunol, 25, pp. 549–51.

Chapman RW, J Robalino, and H Trent. 2006. EcoGenomics: Analysis of Complex Systems. Integr Comp Biol, doi:10.1093/icb/icj049.

Chen CF, X Feng, and J Szeto. 2006. Identification of critical genes in microarray experiments by a Neuro-Fuzzy approach. Computational Biology and Chemistry, in press.

Chen YA, DJ McKillen, S Wu, MJ Jenny, R Chapman, PS Gross, GW Warr, and JS Almeida. 2004. Optimal cDNA microarray design using expressed sequence tags for organisms with limited genomic information. BMC Bioinformatics 5:191.

Chi JT, Z Wang, DS Nuyten, EH Rodriguez, ME Schaner, A Salim, Y Wang, GB Kristensen, A Helland, AL Borresen-Dale, A Giaccia, MT Longaker, T Hastie, GP Yang, MJ Vijver, and PO Brown. 2006. Gene expression programs in response to hypoxia: cell type specificity and prognostic significance in human cancers. PLoS Medicine, 24, p. e47.

Cossins AR and DL Crawford. 2005. Fish as models for environmental genomics. Nat Rev Genet, 6, pp. 324–33.

de Lorgeril J, D Saulnier, MG Janech, Y Gueguen, and E Bachere. 2005. Identification of genes that are differentially expressed in hemocytes of the Pacific blue shrimp (*Litopenaeus stylirostris*) surviving an infection with Vibrio penaeicida. Physiol Genomics, 21, pp. 174–83.

Destoumieux D, P Bulet, D Loew, A Van Dorsselaer, J Rodriguez, and E Bachere. 1997. Penaeidins, a new family of antimicrobial peptides isolated from the shrimp *Penaeus vannamei* (Decapoda). J Biol Chem, 272, pp. 398–406.

Espinosa E, A Redondo, JA Vara, P Zamora, E Casado, P Cejas, and MG Baron. 2006. High-throughput techniques in breast cancer: A clinical perspective. Eur J Cancer, 42, pp. 598–607.

Gianni L, M Zambetti, K Clark, J Baker, M Cronin, J Wu, G Mariani, J Rodriguez, M Carcangiu, D Watson, P Valagussa, R Rouzier, WF Symmans, JS Ross, GN Hortobagyi, L Pusztai, and S Shak. 2005. Gene expression profiles in paraffin-embedded core biopsy tissue predict response to chemotherapy in women with locally advanced breast cancer. J Clin Oncol, 23, pp. 7265–77.

Golden NH and BA Rattner. 2003. Ranking terrestrial vertebrate species for utility in biomonitoring and vulnerability to environmental contaminants. Rev Environ Contam Toxicol, 176, pp. 67–136.

Gracey AY and AR Cossins. 2003. Application of microarray technology in environmental and comparative physiology. Annu Rev Physiol, 65, pp. 231–59.

Gracey AY, EJ Fraser, W Li, Y Fang, RR Taylor, J Rogers, A Brass, and AR Cossins. 2004. Coping with cold: An integrative, multitissue analysis of the transcriptome of a poikilothermic vertebrate. Proc Natl Acad Sci USA, 101, pp. 16970–5.

Gracey AY, JV Troll, and GN Somero. 2001. Hypoxia-induced gene expression profiling in the euryoxic fish *Gillichthys mirabilis.* Proc Natl Acad Sci USA, 98, pp. 1993–8.

Gross PS, TC Bartlett, CL Browdy, RW Chapman, and GW Warr. 2001. Immune gene discovery by expressed sequence tag analysis of hemocytes and hepatopancreas in the Pacific White Shrimp, *Litopenaeus vannamei,* and the Atlantic White Shrimp, *L setiferus.* Dev Comp Immunol, 25, pp. 565–77.

Gueguen Y, JP Cadoret, D Flament, C Barreau-Roumiguiere, AL Girardot, J Garnier, A Hoareau, E Bachere, and JM Escoubas. 2003. Immune gene discovery by expressed sequence tags generated from hemocytes of the bacteria-challenged oyster, *Crassostrea gigas.* Gene, 303, pp. 139–45.

Gueguen Y, J Garnier, L Robert, MP Lefranc, I Mougenot, J de Lorgeril, M Janech, PS Gross, GW Warr, B Cuthbertson, MA Barracco, P Bulet, A Aumelas, Y Yang, D Bo, J Xiang, A Tassanakajon, D Piquemal, and E Bachere. 2006a. PenBase, the shrimp antimicrobial peptide penaeidin database: sequence-based classification and recommended nomenclature. Dev Comp Immunol, 30, pp. 283–8.

Gueguen Y, A Herpin, A Aumelas, J Garnier, J Fievet, JM Escoubas, P Bulet, M Gonzalez, C Lelong, P Favrel, and E Bachere. 2006b. Characterization of a defensin from the oyster *Crassostrea gigas* Recombinant production, folding, solution structure, antimicrobial activities, and gene expression. J Biol Chem, 281, pp. 313–23.

Holland JH. 1975. Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor.

Jenny MJ, AH Ringwood, ER Lacy, AJ Lewitus, JW Kempton, PS Gross, GW Warr, and RW Chapman. 2002. Potential indicators of stress response identified by expressed sequence tag analysis of hemocytes and embryos from the American oyster, *Crassostrea virginica.* Mar Biotechnol (NY), 4, pp. 81–93.

Jenny MJ, AH Ringwood, K Schey, GW Warr, and RW Chapman. 2004. Diversity of metallothioneins in the American oyster, *Crassostrea virginica,* revealed by transcriptomic and proteomic approaches. Eur J Biochem, 271, pp. 1702–12.

Jirapech-Umpai T and S Aitken. 2005. Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes. BMC Bioinformatics, 6, p. 148.

Ju Z, RA Dunham, and Z Liu. 2002. Differential gene expression in the brain of channel catfish (*Ictalurus punctatus*) in response to cold acclimation. Mol Genet Genomics, 268, pp. 87–95.

Khan J, JS Wei, M Ringner, LH Saal, M Ladanyi, F Westermann, F Berthold, M Schwab, CR Antonescu, C Peterson, and PS Meltzer. 2001. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med, 7, pp. 673–9.

Levner I. 2005. Feature selection and nearest centroid classification for protein mass spectrometry. BMC Bioinformatics, 6, p. 68.

Liebovitch LS, LA Shehadeh, and VK Jirsa. 2004. Patterns of genetic interactions: Analysis of mRNA levels from cDNA microarrays. In: Reeke GN, Poznanski RR, Lindsay KA, Rosenberg JR, Sporns O, Eds. Modeling in the Neurosciences: From Biological Systems to Neuromimetic Robotics. CRC Press, New York, pp. 9–24.

McKillen DJ, YA Chen, C Chen, MJ Jenny, HF Trent, J Robalino, DC McLean, PG Gross, RW Chapman, GW Warr, and JS Almeida. 2005. Marine Genomics: A clearing-house for genomic and transcriptomic data of marine organisms. BMC Genomics, 6:34.

Mian S, S Ugurel, E Parkinson, I Schlenzka, I Dryden, L Lancashire, G Ball, C Creaser, R Rees, and D Schadendorf. 2005. Serum proteomic fingerprinting discriminates between clinical stages and predicts disease progression in melanoma patients. J Clin Oncol, 23, pp. 5088–93.

Moore MN, MH Depledge, J Readman, and DR Paul Leonard. 2004. An integrated biomarker-based strategy for ecotoxicological evaluation of risk in environmental management. Mutat Res, 552, pp. 247–268.

Motsinger AA, SL Lee, G Mellick, and MD Ritchie. 2006. GPNN: power studies and applications of a neural network method for detecting gene-gene interactions in studies of human disease. BMC Bioinformatics, 7, p. 39.

Oleksiak MF, GA Churchill, and DL Crawford. 2002. Variation in gene expression within and among natural populations. Nat Genet, 32, pp. 261–6.

Oleksiak MF, KJ Kolell, and DL Crawford. 2001. Utility of natural populations for microarray analyses: isolation of genes necessary for functional genomic studies. Mar Biotechnol, 3, pp. S203–11.

Oltvai ZN and AL Barabasi. 2002. Systems biology Life's complexity pyramid. Science, 298, pp. 763–4.

Patat SA, RB Carnegie, C Kingsbury, PS Gross, R Chapman, and KL Schey. 2004. Antimicrobial activity of histones from hemocytes of the Pacific white shrimp. Eur J Biochem, 271, pp. 4825–33.

Peakall D and J Burger. 2003. Methodologies for assessing exposure to metals: speciation, bioavailability of metals, and ecological host factors. Ecotoxicol Environ Saf, 56, pp. 110–21.

Pusztai L and L Gianni. 2004. Technology insight: Emerging techniques to predict response to preoperative chemotherapy in breast cancer. Nat Clin Pract Oncol, 1, pp. 44–50.

Reif JS, MS Mazzoil, SD McCulloch, RA Varela, JD Goldstein, PA Fair, and GD Bossart. 2006. Lobomycosis in Atlantic bottlenose dolphins from the Indian River Lagoon, Florida. J Am Vet Med Assoc, 228, pp. 104–8.

Renn SC, N Aubin-Horth, and HA Hofmann. 2004. Biologically meaningful expression profiling across species using heterologous hybridization to a cDNA microarray. BMC Genomics, 5, p. 42.

Rieger KE, WJ Hong, VG Tusher, J Tang, R Tibshirani, and G Chu. 2004. Toxicity from radiation therapy associated with abnormal transcriptional responses to DNA damage. Proc Natl Acad Sci USA, 101, pp. 6635–40.

Rise ML, KR von Schalburg, GD Brown, MA Mawer, RH Devlin, N Kuipers, M Busby, M Beetz-Sargent, R Alberto, AR Gibbs, P Hunt, R Shukin, JA Zeznik, C Nelson, SR Jones, DE Smailus, SJ Jones, JE Schein, MA Marra, YS Butterfield, JM Stott, SH Ng, WS Davidson, and BF Koop. 2004. Development and application of a salmonid EST database and cDNA microarray: data mining and interspecific hybridization characteristics. Genome Res, 14, pp. 478–90.

Sanchez-Carbayo M, ND Socci, J Lozano, F Saint, and C Cordon-Cardo. 2006. Defining Molecular Profiles of Poor Outcome in Patients With Invasive Bladder Cancer Using Oligonucleotide Microarrays. J Clin Oncol, 24, pp. 778–89.

Shehadeh LA, LS Liebovitch, and VK Jirsa. 2006. Relationship between global structures of genetic networks and mRNA levels measured by cDNA microarrays. Physica A, 364, pp. 297–314.

Sturm-Ramirez KM, DJ Hulse-Post, EA Govorkova, J Humberd, P Seiler, P Puthavathana, C Buranathai, TD Nguyen, A Chaisingh, HT Long, TS Naipospos, H Chen, TM Ellis, Y Guan, JS Peiris, and RG Webster. 2005. Are ducks contributing to the endemicity of highly pathogenic H5N1 influenza virus in Asia? J Virol, 79, pp. 11269–79.

Supungul P, S Klinbunga, R Pichyangkura, I Hirono, T Aoki, and A Tassanakajon. 2004. Antimicrobial peptides discovered in the black tiger shrimp Penaeus monodon using the EST approach. Dis Aquat Organ, 61, pp. 123–35.

Tarca AL, JE Cooke, and J Mackay. 2005. A robust neural networks approach for spatial and intensity-dependent normalization of cDNA microarray data. Bioinformatics, 21, pp. 2674–83.

Tom M and M Auslander. 2005. Transcript and protein environmental biomarkers in fish—a review. Chemosphere, 59, pp. 155–62.

van de Peppel J, P Kemmeren, H van Bakel, M Radonjic, D van Leenen, and FC Holstege. 2003. Monitoring global messenger RNA changes in externally controlled microarray experiments. EMBO Rep, 4, pp. 387–93.

Vapnik V. 1998. Statistical Learning Theory. Wiley Interscience, New York.

Voit EO. 2000. Computational analysis of biochemical systems: a practical guide for biochemists and molecular biologists. Cambridge University Press, New York.

Voit EO and J Almeida. 2004. Decoupling dynamical systems for pathway identification from metabolic profiles. Bioinformatics, 20, pp. 1670–81.

von Schalburg KR, ML Rise, GA Cooper, GD Brown, AR Gibbs, CC Nelson, WS Davidson, and BF Koop. 2005. Fish and chips: various methodologies demonstrate utility of a 16,006-gene salmonid microarray. BMC Genomics, 6, p. 126.

Wang W, C Yang, G Mathee, and G Narasimhan. 2005. Clustering Using Self-Ordering Maps (ASOM) and Applications. Lecture Notes in Computer Science Vol. 3515, International Workshop on Bioinformatics Research and Applications (IWBRA) Springer-Verlag.

Waters MD, K Olden, and RW Tennant. 2003. Toxicogenomic approach for assessing toxicant-related disease. Mutat Res, 544, pp. 415–24.

Webster RG, Y Guan, L Poon, S Krauss, R Webby, E Govorkovai, and M Peiris. 2005. The spread of the H5N1 bird flu epidemic in Asia in 2004. Arch Virol Suppl, pp. 117–29.

Williams TD, K Gensberg, SD Minchin, and JK Chipman. 2003. A DNA expression array to detect toxic stress response in European flounder (*Platichthys flesus*). Aquat Toxicol, 65, pp. 141–57.

Wolfe ND, AA Escalante, WB Karesh, A Kilbourn, A Spielman, and AA Lal. 1998. Wild primate populations in emerging infectious disease research: the missing link? Emerg Infect Dis, 4, pp. 149–58.

Wright S. 1831. Evolution in Mendelian Populations. Genetics, 16, pp. 97–159.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 24
Functional Genomics

Perry B. Hackett and Karl J. Clark

Introduction

The twenty-first century opened with the publication of the first sequence of a vertebrate genome, the human genome. Called *The Book of Life*, the sequencing of the human genome was expected to reveal genes that regulate development in humans and, by extension, in other vertebrates. Subsequent genetic revelations could be used for developing new medicines and methods for manipulating genetics to understand and improve human health. The human genome project was slated to be the first of several vertebrate animal genomes that would allow scientists to better understand animal physiology as well as enable the modification of genomes in order to alter phenotypes to improve animal health and increase commercial value. The results of the two major human genome projects, one by Celera Corporation and the other under the auspices of the United States National Institutes of Health (NIH), initially were hard to believe and understand—there appeared to be fewer than 30,000 protein-encoding genes in nearly 2.9×10^9 bp of haploid genome. In contrast, the *Drosophila melanogaster* (fruit fly) and *Caenorhabditis elegans* (nematode/roundworm) genomes have about 13,000–19,000 genes in about $9.7–18 \times 10^7$ base pairs (bp). That is, although the human genome was twenty-fold to thirty-fold larger, it had only about twice the number of genes as invertebrate animals that are smaller than a human fingernail. Within a couple of years, the small number of genes required to specify a vertebrate was confirmed with the sequencing of the mouse and rat genomes—again about 30,000 genes at a density of about 1 gene per 100,000 bp. The number of genes in fish is expected to be about the same.

How can a complex vertebrate with so many cells (approximately 10^{13}–10^{14} in humans) require so few genes? One answer came from the finding that the protein encoding sequences of vertebrate genes were most often divided into short sequences called exons. Following transcription into precursor messenger RNAs (mRNA), a single gene could be responsible for a family of mRNAs that encode several different proteins. Diverse amino acid sequences can be derived from the inclusion, removal, or modification of various polypeptide motifs as a result of alternative splicing. These closely related proteins sometimes have independent or overlapping functions. Moreover, some genes have multiple initiation sites for transcription and translation as well as multiple termination sites. As a result, vertebrate genomes are envisioned to encode up to hundreds of thousands of active proteins. The exact number is unknown because differential splicing, transcriptional initiation, and termination events occur in cells of various organs.

The second surprise that was unanticipated by the fruit fly and nematode genome projects was that about two-thirds of the 30,000 genes could be put into one of a relatively few families of polypeptides. Several families had related enzymatic activities

required for DNA, RNA, protein, and carbohydrate synthesis and metabolism that had been worked out from much earlier studies of bacterial genes (e.g., *Escherichia coli* with a genome of about 4.7×10^6 bp has about 3,200 genes at a gene density 70 times that of vertebrate genomes). Other large families encode a variety of transferases, kinases, receptors, ion channels, and transporter genes whose activities are guided by a plethora of regulatory molecules. With many related genes in each gene family, which gives rise to thousands of related proteins from alternative splicing, a second major question arose: What is the bottom line role of each of the encoded polypeptides? Current understanding suggests that some genes play critical roles in initiating certain developmental events, other genes play roles in maintaining the functions of the cells following differentiation events, other genes play supporting or backup roles to augment or protect cellular function, and some genes play very specialized roles that respond to environmental cues or changes. Sorting genes into these categories is not easy or in most cases not even possible to determine by analyzing simple sequence alone. Other tests are necessary and they form the area of functional genomics.

The goals of functional genomics are to better understand the roles of elements of the genome that directly or indirectly affect the development, growth, metabolism, immunity, behavior, and aging of an organism. This includes associating protein products, noncoding RNAs, and *cis*-acting DNA elements with specific functions in the above processes. The goals of functional genomics vary from project to project and organism to organism. In the broadest sense, we conduct functional genomics studies for the following reasons:

1. to understand the health and physiology of organisms
2. to manipulate organisms for specific purposes
3. to understand the evolution of life

Broad arrays of approaches with sophisticated experimental techniques have been developed to study functional genomics. Because of the vast numbers of genes, the even larger number of polypeptides, the variation in expression from tissue to tissue, and the variety of organisms under study, 'high throughput' methods are required so an investigator can examine and sort many activities at the same time. Clearly high-speed computation is an essential feature of most techniques. In the following sections, we describe some of the approaches and techniques with the understanding that new, more powerful methods are constantly emerging, especially as computational power continues to improve. The improvements include tools and instruments for digital recording of information and newer algorithms for sorting and analyzing information developed as our understanding of gene function increases. Bioinformatics is a rapidly developing area of modern biology that looks to unravel the mysteries of biological processes using high-speed analyses by computers with vast memories and computational rates.

Approaches and Methods for Studying Functional Genomics

General Strategies

Functional genomics comprises examining and cataloging expression of genes in normal tissues at various stages of development or in response to particular environments.

There are three overall strategies for elucidating the functions of genes and their importance in development and health of an organism. The combination of all three strategies generally is needed to understand fully the role(s) a gene plays in the well-being of an animal. The first is to examine the genome and compare the predicted genes with genes in other, better-characterized organisms. This strategy is called comparative genomics and it relies extensively on bioinformatics. The second is gene profiling, which comprises examining and cataloging expression of genes in normal tissues at various stages of development or in response to particular environments. The third is loss-of-function and gain-of-function mutagenesis to examine the effects of altering expression of specific genes, or selected groups, to understand their importance and physiological roles.

Expression of genes can be measured in terms of the mRNA or their protein products. Both methods have advantages and disadvantages. mRNA is relatively easy and inexpensive to detect using a variety of hybridization techniques. The spatial localization of mRNAs in cells of multicellular tissues can be determined by *in situ* hybridization in which tissues or even whole organisms are treated to allow penetration of complementary DNAs (cDNA) that will selectively hybridize to mRNAs present in the cells. Alternatively, for quantitative analysis to establish the sequence and/or quantity of an mRNA species, mRNA can be isolated from tissues and either electrophoresed through an agarose gel for northern blotting to establish size, abundance, and number of splicing variants. In some cases, the mRNA may be prehybridized to an assortment of DNA probes to establish particular features of the mRNA such as splicing sites and transcriptional start sites by methods called primer extension and RNA protection. Manuals describing the details of all of these techniques are in abundance (e.g., Ausubel et al. [1998]). Reflecting the availability of sophisticated instrumentation, extremely low levels of mRNA can be detected using reverse transcriptase-polymerase chain reaction (RT-PCR) in which single-stranded mRNA sequences are copied by RT into an RNA-DNA hybrid molecule that is then amplified by the method of PCR to yield thousands of copies of a double-strand DNA fragment of defined size. With automated thermocyclers equipped with fluorescence-detection capabilities, the RT-PCR can be automated and results obtained without gel analysis in what is called real-time RT-PCR. There sometimes is confusion when RT is used to refer to 'real-time' rather than 'reverse transcriptase.'

Hybridization of nucleic acids is simple, but functional activity in cells is generally due to proteins and peptides. Often, proteins are modified by cleavage to remove amino acids at the amino terminus (e.g., for secreted proteins), glycosylation (addition of complex sugars/carbohydrates), phosphorylation (by kinase molecules that add phosphate residues to certain serine, threonine, and tyrosine residues in a polypeptide), acetylation, and methylation. These modifications can greatly affect the activities of polypeptides and cannot be deduced from mRNA sequences alone. These activities depend on the presence of many different enzymes. Hence, there is great interest in examination of proteins and their processed products. As functional genomic investigations enter a new realm of analysis, new fields are created, often with the suffix 'ome' to describe the level (e.g., genome for DNA in genomes; transcriptome for the total number of transcriptional products in a cell, tissue, or organism; proteome for the total number of polypeptide products in a cell, tissue, or organism; kinome for the total number of phosphorylated proteins in a cell, etc.). Proteins and their modifications are far more difficult to study—the combination of

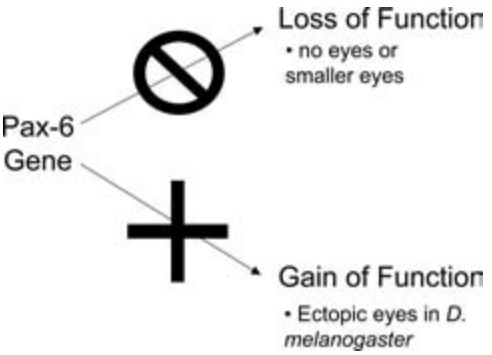


Figure 24.1. Experimental testing for a gene's function can be by inactivating the gene or misexpressing the gene. The consequences of blocking expression and overexpressing the Pax-6 gene, which is a master regulatory gene responsible for eye development, are illustrated.

21 amino acids is far vaster than the combinations of the four bases in DNA or RNA. Whereas hybridization probes are inexpensive and trivial to construct for detection of complementary DNA and RNA sequences, such technology is not available for proteins. Antibodies can be raised to particular motifs in proteins, but this takes months and is expensive. Consequently, most high-throughput analyses have concentrated on detection of mRNA products rather than their encoded proteins.

Merely identifying the players in a cell by either protein or mRNA detection is insufficient for appreciating the roles they play. Their generic functions may be predictable, but the particular importance of the implementation of their function may not be as clear. To use an analogy, one might be able to predict the roles of players on a sports team but not necessarily their relative values; that becomes apparent when an injury occurs and they are removed or replaced. The fact that they are participating in the field does not tell much of their roles in the event. Likewise, to uncover the identities of essential and/or important regulatory genes, one must eliminate or curtail their activities or otherwise alter their expression patterns. This is most often done by either mutating the gene or blocking its expression. Alternatively, extra copies of genes can be expressed in cells, either in a controlled or uncontrolled manner to determine the effects of abnormal expression. When all such data are assembled, it generally is possible to infer the function of a particular gene. Thus, blocking, enhancing, and alternatively expressing genes all provide different clues on the roles a gene plays in the animal (Figure 24.1). In the Experimental Determination of Gene Function section, we discuss how this is actually done.

Comparative Genomics

The evolution of life on earth, which appears to have begun more than 3.5 billion years ago, is recorded in the genomes of current organisms. Comparative genomics is the comparison of genomes from different species. The purpose is to determine the function of genes and noncoding regions of the genome based on the commonality of genetic sequences and observed traits or physical attributes of the organisms from

where the sequences were obtained. Decoding the genomic tree of life has required the examination of chromosomal sequences from many organisms because genes have been transferred in pieces and whole units as well as duplicated and lost over the millions of years that any given species might exist. The earliest examinations of the evolution of genetic sequences were those of ribosomal RNA (rRNA) and protein-encoding genes, whose encoded amino acid sequences were generally more highly conserved than their DNA sequences. This is because the activity of a protein is dependent on its structure and consequently the sequence of its amino acids. Because of the degeneracy of the genetic code, wherein a single amino acid can be encoded by one to six codons, mutations in a DNA sequence do not always lead to a mutation in the amino acid sequence. Indeed, the genetic code has evolved to be modestly robust in accommodating common mutations in the DNA to preserve the protein structure (Modiano et al. 1981). However, only about one-half of the highly conserved DNA actually encodes proteins or ribosomal and transfer RNAs. The other half encodes either transcriptional regulatory motifs, to which specific transcription factors bind to produce tissue-specific mRNAs and proteins, and other regulatory sequences such as small functional RNAs (discussed further in the Reverse Genetics: Knocking Out and Knocking Down Genes section) that contribute to higher levels of gene regulation. For example, a thorough study of the carp (*Cyprinus carpio*) β-actin gene showed that the DNA sequence encoding the protein was 88–91% conserved at the DNA level with the β-actin genes from chickens, rats, and humans whereas the amino acid sequence was 99.0–99.5% conserved relative to the land vertebrates. Even more enlightening was the 100% conservation in transcriptional regulatory motifs for the β-actin gene between the fish and the land vertebrate genes (Liu et al. 1990a). Thus, regulation of a gene is as important as the actual product being regulated. Many of the differences between vertebrate species are not due to their genes, but in the way many common genes are regulated.

Bioinformatics

Many different features of genes are analyzed for comparative genomics, including the lengths, numbers of exons and introns, and locations of genes or their relatives with respect to each other, which is termed synteny. Over the course of evolution, genes, and whole genomes, have undergone many duplications. Some of the duplications are lost since they are of no use but are a 'genetic burden' in terms of energy required for their continued duplication and expression. Often the duplicated genes that are maintained further evolve to assume different but related functions. Genes related by sequence are called homologs. Although the relatedness of genes is often expressed as a percent homology, this term is incorrect; genes are either homologs or not. Relatedness is more correctly referred to as percent conservation of sequence. Because of the amplification of homologs in genomes, it is often difficult to attribute specific functions to specific genes. There are two types of homologous genes:

- Orthologs, genes in different species that evolved from a common ancestral gene by speciation, which are essentially identical in terms of relative genomic position (synteny) and function
- Paralogs, genes related by duplication within a genome, which are related by sequence but not position and/or function

Synteny is a powerful tool because although genomes from vertebrates are scrambled at the chromosomal level, at the gene level, equivalent to say units of about 10^6 bps, recombination is not so frequent and thus neighboring genes are more often conserved than not (Figure 24.2). In comparisons between vertebrate genomes, homologs very rarely lose both neighboring genes. Nevertheless, sometimes it is difficult to determine orthology.

Comparative analysis of genomic sequences involves thousands to millions and even billions of bps from an ever-growing population of organisms. Only computers can digest so much information. The National Center for Biotechnology Information (NCBI), http://www.ncbi.nlm.nih.gov/, a part of the NIH, was established in 1988 as a national resource for molecular biology information. NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information—all for the better understanding of molecular processes affecting human health and disease. The actual storage of genetic information is kept by GenBank, which has the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (Benson et al. 2006). All GenBank sequences have accession numbers that allow them to be directly imputed into many algorithms that have been developed to analyze sequences. There are approximately 6×10^{10} bases in 6×10^7 sequence records in the traditional GenBank divisions as of early 2006. GenBank at NCBI collaborates with the DNA DataBank of Japan and the European Molecular Biology Laboratory, which exchange data daily. Basic Local Alignment Search Tool (BLAST) (http://www.ncbi.nlm.nih.gov/BLAST/) is probably the most used computer program to line up multiple genomes (or all available sequence data) and look for regions of similarity among them. Other sequence-similarity tools are accessible to the public over the Internet. The deposit of sequences into GenBank has been growing exponentially as the cost of sequencing DNA has dropped.

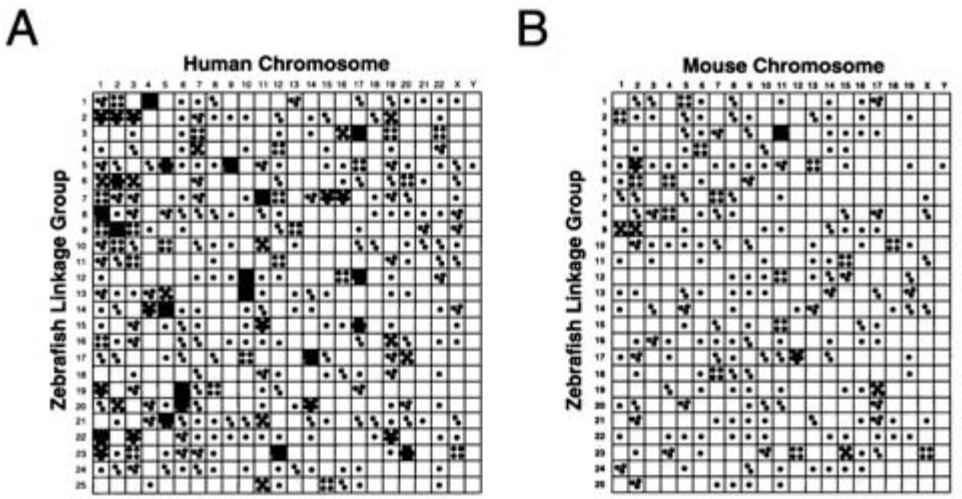


Figure 24.2. Synteny of chromosomal segments from zebrafish with either (A) human or (B) mouse chromosomes. The matrices show that genomes of about 30,000 genes have been broken into several hundred units that have reassorted as blocks of genes in different chromosomes. (Adapted from Woods et al. 2000).

Sequences are obtained in either of two generic ways. The first is targeted sequencing of a particular gene or mRNA copy of a gene. The second is random cleavage of a genome into smaller fragments of 10 kilobase pairs (kbp) to 1 million base pairs (Mbp), which are then sequenced randomly. The random sequences can be assembled into larger units called contigs. A genome is completely sequenced when the number of contigs matches the number of chromosomes, or genetic linkage groups. With current, high-throughput sequencing machines, a simple prokaryotic genome can be sequenced into a complete genome in a few weeks. Some emerging sequencing technologies would allow complete sequencing of a prokaryotic genome in less than 1 day using only a single piece of equipment (see Chapter 25). However, assigning a functional role to each identified genetic sequence takes substantially longer. Vertebrate genomes take much longer. Indeed, this has not yet been achieved for vertebrate genomes due to the extraordinary, in most cases, numbers of repetitive elements and simple sequence repeats that comprise centromeres and other regions of genomes. Consequently, genome sequencing is done in repetitive stages, often called coverage, with each stage extending and combining known contigs as well as correcting mistakes in earlier assemblies. A 1× coverage is equivalent to obtaining sequences that total the genome, in this case nearly 37% of the genome will not be represented whereas another 26% will be represented more than once. As the number of sequences increases, the numbers of contigs increase until there is sufficient overlap to allow contigs to be assembled. By 2× coverage there is more merging of sequences than finding new sequence and the numbers of contigs decreases while the number of bps in contigs increases. Thus, 5× coverage means that the total sequences up to that time are equivalent to 5 times the total genome of the organism—a point at which there is a high confidence that most unique sequences have been found and roughly placed on chromosomal maps. In practice, the presence of repetitive sequences makes final assembly of contigs into chromosomal lengths very difficult. Readers are referred to Chapter 25 and Chapter 26 for sequencing technologies and genome sequencing related issues.

Model Organisms

Conducting experiments in most vertebrates is quite expensive, labor intensive, and in the case of humans, unethical. Hence, because of the conservation of action of most gene products, model organisms have been employed to provide the initial clues about gene function in various organisms. As vertebrates, fish have conserved gene function with mammals such as mice and rats, the premier model organisms used in pharmaceutical and vertebrate developmental biology. Genetic studies in these organisms can be performed faster and less expensively than with most vertebrates. Interestingly, one exception is the zebrafish model system (Dahm and Geisler 2006) that is easy to use, inexpensive, and thereby capable of providing answers about vertebrate gene function at a fraction of the cost of using mice.

Mammalian genomes have approximately the same DNA contents, about 3 billion bps. In contrast, genomes of teleost fish vary more than tenfold, from about 3×10^8 bps (pufferfish, *Chelonodon fluviatilis,* and *Takifugu rubripes* [fugu]) to 4.5×10^9 bps (triploid Atlantic salmon, *Salmo salar*) with a mean of about 10^9 bps (Gregory 2005) (http://www.genomesize.com). Zebrafish (*Danio rerio*), a popular model for developmental biologists, has a genome of about 1.7×10^9 bp. The major difference in the

genomes is not in their numbers of genes, but in their contents of repetitive DNA. More than 90% of the fugu genome is unique in contrast to mammalian and other teleost genomes, which are more than 50% composed of repetitive DNA sequences. Moreover, the introns in fugu genes are few and extremely short, thereby reducing the average size of the genes. Thus, in general the numbers of genes vary only slightly in vertebrates with genome size not being indicative of gene number or genetic complexity.

Gene Families

Ray-finned fish appear to have more copies of many genes than other vertebrates as a result of a genome duplication event that occurred before the teleost radiation. Many of the duplicated genes were lost and many reshuffled, leading to partial duplications. Sequence duplications that are currently present in the genome represent nonfunctional, neofunctional, and subfunctional copies of duplicated genes (Figure 24.3). This has complicated functional assignment to specific genes. For instance, the 39 Hox genes, which are crucial for vertebrate development, are normally organized into four clusters. However, in some ray-finned fish, there are extra clusters that resulted from partial duplications (Mulley et al. 2006, Kurosawa et al. 2006). Expression of the duplicated genes is subtly different during development, with

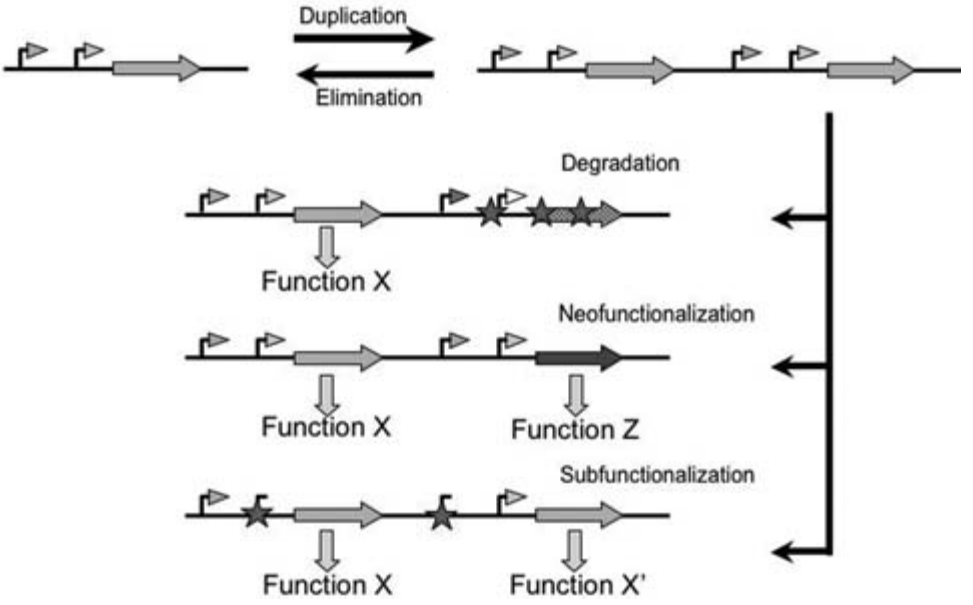


Figure 24.3. Gene duplication and loss. A gene is shown with two transcriptional elements. Duplication of the transcriptional unit and regulators followed by mutation can lead to (1) the original situation if one gene is destructively mutated, (2) new gene function following nondestructive mutations, or (3) splitting the functions of a single gene due to different promoters driving essentially the same gene under different circumstances or in different cell types. (Also see color plate.)

contributions of both genes being important (Tumpel et al. 2006, Kaufman et al. 2000). Duplicated genes, such as the Hox genes, often are tandemly arranged with small variations in their sequences, which can lead to microspecialization of function (Hancock 2005). For instance, olfactory receptors are critical to an animal's interactions with its environment, from recognition of food and enemies to potential mates. There are an estimated 1,500 olfactory receptor genes in mice of which 1,200 are active, whereas there may only be about 400 in humans and fewer in fish, probably because fish receptors are for water-soluble chemical entities whereas vertebrate receptors are for volatile, airborne agents. In Atlantic salmon (*Salmo salar*), the main olfactory and vomeronasal receptors have been characterized and divided into six groups that comprise a number of individual genes that have a common evolutionary history. A database named Wanda (http://www.evolutionsbiologie.uni-konstanz.de/Wanda/) lists groups of orthologous genes from ray-finned fish and arranges them into phylogenetic trees and, where possible, provides genome-map positions and functional information (Van de Peer et al. 2002). Yet to be done is a full characterization of the regulatory sequences that control expression of the various genes in these gene families. In the end, it is the subtle variation in genes and their regulation that leads to the differences in physiological attributes of species and their behaviors. Readers interested in duplicated genes and the study of their functions are referred to Chapter 28 of this book.

Most genes operate in concert with each other. In some cases the cooperation is in the form of multimeric protein complexes (e.g., membrane assemblies required for signal transduction, whose individual members may be from the same or different gene families). Teasing apart the roles of different proteins that combine in various ways is a challenge (Cheng et al. 2003). In addition, most proteins cooperatively operate in networks and signaling cascades, often called systems biology. Thus, the effects of minor differences in the approximate 30,000 genes in fish genomes is multiplied to an enormous, and unknown, number by their interactions with one another. Comparative genomics can be used to identify *cis*-regulatory sequences such as promoters and other sites that influence transcription. Transcriptional regulatory proteins bind to these sequences, and such proteins often are induced as a result of variation in the environment or developmental plan of the animal. Accordingly, genes that are co-regulated are expected to have similar *cis*-regulatory sequences, and by identifying these sequences, networks of gene interactions can be inferred. This has been done in yeast (Tagne et al. 2004), and similar efforts are under way in vertebrates (Davidson and Erwin 2006).

Noncoding RNAs

Regulation by noncoding RNAs (ncRNA) is extremely important in vertebrates (Eddy 2006, Woolfe et al. 2005). ncRNAs play roles in a variety of processes including transcription, DNA replication, RNA processing and modification, translational regulation, and protein degradation and translocation. A particular class of ncRNAs, micro RNAs (miRNA) are 20- to 22-nucleotide RNAs that regulate translation of mRNA by binding to the 3′ untranslated ends of the mRNA to block translation and direct the destruction of the messenger (Humphreys et al. 2005). Other ncRNAs

often are longer and are processed from RNA polymerase II transcripts, the same enzyme responsible for mRNA synthesis. ncRNAs may come from transcripts that lack open reading frames for protein synthesis or from introns of protein-encoding genes. Their expression plus their conservation emphasize their importance. However, their identification presents a challenge to computational biology and functional genomics. In addition to sequence conservation, the activities of many ncRNAs are believed to depend on the structures that they can form by folding into stems and loops via base pairing. Thus, prediction of RNA secondary structure conservation and thermodynamic stability combined with conservation in vertebrates from fish to humans has been used as one means of identification of this important class of functional agents. The number of potential ncRNAs identified in this way is close to the total number of genes (Washiet et al. 2005).

Clearly computer analyses of genetic sequences can identify many regulatory motifs and genes, including those that encode proteins and those that do not, but the bottom-line functions of these genes often is unknown because many are coexpressed in the same cells or are expressed in different tissues at different times. Moreover, genes are expressed at varying levels in different cells, which makes it difficult to associate function of a gene product with its mere presence or, in the case of analyses for mRNA, its implied presence. Thus, experimental tests are crucial to attribute functions to genes.

Profiling Gene Expression

Fish are comprised of a multitude of cell types that compose various organs with specific functions. These functions are mediated by specific sets of proteins, and their respective genes in general are transcribed at rates to provide sufficient levels of these proteins. There are a number of assays that are used to determine the panoply of genes that are expressed in specific tissues or as a consequence of developmental and environmental cues. Several of these assays are discussed in an order that roughly corresponds to their degree of difficulty and informational value (Figure 24.4).

Tissue-Specific cDNA Libraries

Hastie and Bishop (1976) conducted the first semithorough examinations of the diversity of gene expression in differentiated tissues by making cDNAs to bulk isolated mRNAs that were then hybridized to genomic DNA. The results of this work indicated the following:

1. About 10,000–12,000 genes were expressed in most cell types.
2. Genes were expressed at levels from about 10 mRNAs/cell (low abundance: about 98% of the expressed genes) to several hundred copies (middle repetitive abundance: about 2% of the genes expressed) to more than 10,000 mRNAs/cell (abundant: less than 10 genes per cell type).
3. 80–90% of the rarely and middle-repetitively expressed genes were expressed in all cells at about the same rate and only a few genes were differentially expressed.

	Target	Throughput	Comments
Tissue/Temporal analysis			
cDNA Library	mRNA	med	
ESTs	mRNA	high	$, CP
qRT-PCR	mRNA	med	$, QA
Microarray	mRNA	high	$, QA, CP
Cellular/Subcellular domains			
ISH	mRNA	low	
IHC	protein	low	AR
Reporter-Tagged Expression	gene	low	TC

Figure 24.4. Analyses of gene expression with respect to target and rate of analyses (throughput). cDNA libraries, EST databases, quantitative RT-PCR, and microarrays can yield information regarding expression profiles with resolution down to a particular tissue or specific time. *In situ* hybridization (ISH), immunohistochemistry (IHC), and reporter-tagged transgenesis allow localization of expression to individual cell types of a tissue and sometimes to subcellular domains. Notes: $: extensive capital costs; CP: community project; QA: quantitative assay; AR: antibody required; TC: extended time commitment.

Those genes expressed at a common rate in all cells comprise the 'housekeeping' genes that are responsible for basic structures and metabolism in cells. The uniquely or differentially expressed genes are those involved with specific functions of cells. Because the experiments were done with bulk mRNA, they were unable to identify the particular genes whose expression patterns varied in differentiated tissues. Nevertheless, this approach gave approximate answers that stand to this day.

Following the development of automated DNA sequencing, stretches of mRNA isolated from specific tissues were reverse-transcribed into cDNA sequences that provided expressed sequence tags (EST), which could be used to profile gene expression from different tissues (Adams et al. 1991). This became the basis for the first, high-throughput effort for the human genome project. The power of the ESTs was appreciated by members of the fish community who began establishing libraries for genetic linkage mapping from catfish (Liu et al. 1999) and establishment of databases such as that for salmonids (Rise et al. 2004). For more on EST analysis, interested readers are referred to Chapter 20 of this book.

Microarray Analysis

Within the last decade, more advanced methods of measuring gene expression have been developed of which microarrays (see Chapter 21 and Chapter 22) are the most commonly used. The functional landscape of gene expression in the mouse with respect to tissue type and stage of development has been investigated for as many as 40,000 predicted mRNAs (many of which are splicing isoforms derived from the same initial pre-mRNA transcript) in 55 tissues (Weitzman 2004, Holmes and Brown 2004, Zhang et al. 2004). A comprehensive list of developmentally regulated zebrafish

genes and their expression profiles during embryogenesis, including novel information on the temporal expression of several thousand previously uncharacterized genes are accessible (http://giscompute.gis.a-star.edu.sg/~govind/zebrafish/) (Mathavan et al. 2005).

However, the resolution of most microarray experiments generally provides only an overview of gene expression in a sample of tissue or cells; it does not elucidate either the variation of expression from cell to cell nor, in many cases, the variation in expression of genes in various cell types that compose an organ or tissue. For instance, a liver is composed of hepatocytes, Kupfer cells, and sinusoidal epithelial cells as well as cells of the vascular system that provide necessary nutrition to all cells. Each cell type has specific roles that they play in this single organ. Moreover, there is a supposition that knowing the identities of genes that are co-regulated in different tissues or in response to environmental cues indicates gene function. However, we now know that many genes are expressed at varying levels in different tissues, making functional assignments difficult solely on the basis of expression. Thus, further methods are needed to resolve roles of gene expression in complex animals such as fish.

In Situ *Hybridization*

Expression-function correlations on a per-cell basis can be done by hybridization analysis of mRNAs in cells of tissue samples fixed on slides. This is called whole-mount *in situ* hybridization (ISH). ISH is typically performed by targeting a particular mRNA sequence with a cDNA. The cDNA can be labeled with a radioactive nucleotide for detection by autoradiography (old style) or labeled with a chemical moiety that can be detected either enzymatically or by a fluorescently labeled antibody (new style). ISH has the advantage of showing expression levels per individual cell. Controls are needed for accurate analysis. For this, generally sense-DNA probes of the same sequence as the antisense probes are used to detect background binding of the DNAs. Such controls are essential when the transcript levels are low, as they are for most genes. Indeed, quite often it is difficult to detect low expression genes. Detection by reverse transcriptase and *in situ* PCR amplification (RT-PCR followed by hybridization) has been done to detect some rare species of mRNA (Song et al. 2004). Nevertheless, when signals are detected, information only can be inferred about the presence of the encoded protein. Translational rates of mRNA as well as protein stability affect the relative levels of proteins within a cell. Therefore, regardless of the presence of the mRNA (Wills 1999), further testing is necessary to gain information about protein expression, which is often the final gene product. Once again, using the analogy in sports, presence in the field does not itself indicate roles of the players.

Immunohistochemistry

Immunohistochemistry (IHC), similar to *in situ* hybridization, is the localization of proteins in tissue whole-mount or sections using labeled antibodies that bind via antigen-antibody interactions. Fluorescently labeled antibodies are most commonly

used, which allows detection of multiple proteins in a single cell with antibodies that carry various fluorescent tags. IHC has the advantage over ISH because it actually detects the protein of interest, thereby avoiding questions about translational regulation and/or protein stability. The disadvantage of IHC is that it requires a specific antibody that must be raised against the protein of interest; this can take months. In contrast, ISH probes can be obtained within days for far less expense. Consequently, ISH is the method of choice for high throughput screening of gene expression with IHC analysis for those times when certainty of expression of the protein is necessary or location of the protein within the cell is important.

By combining profiles of gene expression with bioinformatics, one can obtain a probable role of a gene (ortholog) in those cases where its function has been determined in another organism. However, when orthology either is questionable, the gene is unique, or the process under investigation is novel, then functional tests are necessary. Tests that directly examine gene function are reviewed in the following section.

Experimental Determination of Gene Function

To this point all of the tests provide inferential information that will suggest, but not really define, the function of a gene product. Moreover, the significance of the roles of proteins may vary in different tissues or at different times of development. Consequently, functional assays are necessary to define the true roles that genes and their products play in the physiological health and development of an organism. There are two main categories of tests:

1. Loss-of-function tests are those that measure the effects of eliminating or significantly reducing expression of a gene.
2. Gain-of-function tests are those that measure the effects of increasing expression of a gene, either in the normal tissue in which it is expressed or in tissues where the gene is normally silent.

There are multiple techniques for conducting each category of testing, and new methods are being developed due to the importance of the questions that can be answered by these approaches. The choice of which techniques to use depends on several parameters, including the organism that one is studying, the tissue/organ of interest, and the period of development that may also be of interest. Loss-of-function assays are the most commonly employed because their genetic basis is straightforward; mutate the responsible gene. Two approaches for loss-of-function studies are discussed in the next two sections followed by a section on gain-of-function tests.

Forward Genetics and Genetic Screens

For more than 100 years, geneticists have looked at mutations that abolish or alter gene expression to identify the presence of genes. Within the last 2 decades, molecular cloning of DNA and sequencing has led to further characterization of genes and their functions. The process of randomly mutating DNA, observing a resulting phenotype that is most probably due to inactivation of a gene, followed by identification of

the mutated gene is a pillar of genetics and is called forward genetics. An alternative approach that is discussed in the Reverse Genetics: Knocking Out and Knocking Down Genes section is reverse genetics wherein the investigator starts with a gene and looks for a phenotype rather than the usual way of finding a phenotype and searching for the responsible gene. Most often today, researchers are less interested in looking at the totality of all possible phenotypes coming from blocking expression of every gene. Rather, they are interested in a certain function, for example, the genetic determinants of the ability of fish such as salmon to be born in fresh water, migrate to live a life stage in salt water, and then return to fresh water to spawn (anadromous) or the genes that allow fish to survive in hypoxic water that has a low oxygen content. There are three steps in finding genes involved with a certain physiological process that involve treating and analyzing large numbers of animals (Figure 24.5):

1. Create populations of mutants.
2. Find individuals in the population of mutants that appear to have a phenotype involved with the process of interest. This step is called phenotypic screening.
3. Identify the gene and the mutation that is responsible for the abnormal phenotype that was caught in the genetic screen.

There is no single or best way to conduct a forward genetic screen and most are difficult because one or more of the three essential steps will be labor intensive, depending on which combination of techniques is used. That is, methods to generate many mutations quickly and inexpensively (Step 1) often make it difficult to identify the mutated genes (Step 3). Alternatively, there are methods that involve tagging a gene with some foreign DNA sequence to allow easy identification, but these techniques generally generate far fewer mutants per treatment. An alternative strategy, examining the inactivation of a selected gene that is suspected of having a particular function, is called gene knockout. This strategy will be discussed after the random mutagenic methods.

Fish and other vertebrates are diploid for most genes and tetraploid for some. Inactivating a single copy of a diploid (or tetraploid) gene rarely has a noticeable effect on an animal; that is, all but a handful of genes are not haplo insufficient. Consequently, even though a gene is mutated, it generally does not create a phenotype that can be identified in a screen. Rather, the fish or other mutant animal must be mated, most often with a wild-type animal, and the progeny of that mating are intermated in order to produce a few offspring (25% as predicted by Mendelian genetics, if there are no complicating issues) that have mutations in both alleles. This is called a three-generation screen and

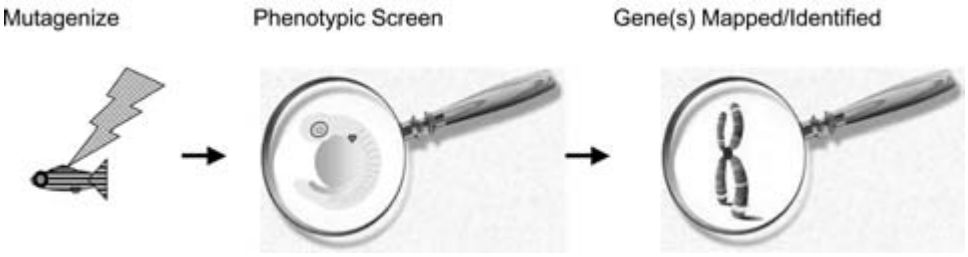


Figure 24.5. Steps in mutagenesis screens to determine gene function.

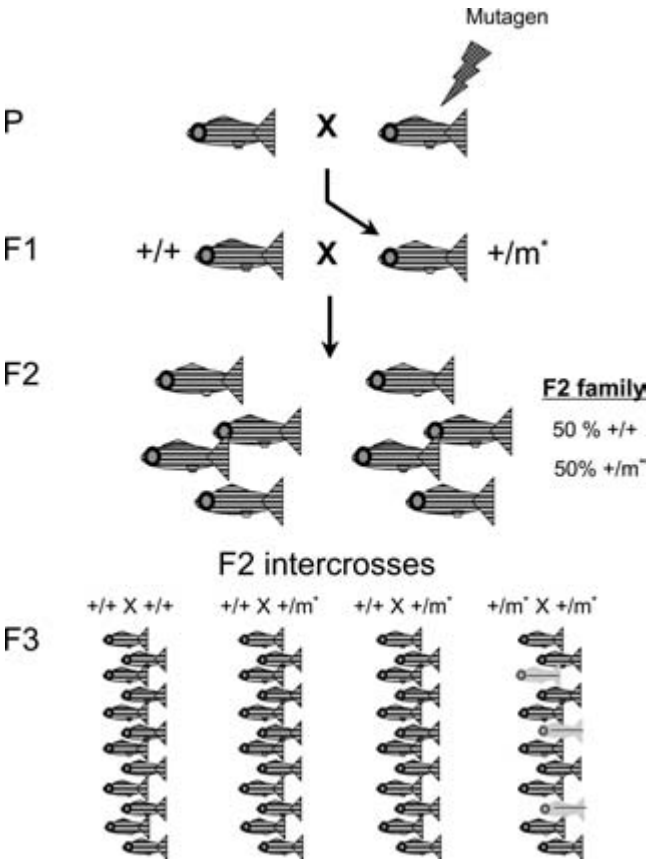


Figure 24.6. Three-generation gene screen.

is illustrated in Figure 24.6. If the mutated gene results in some type of signal that can be detected in the F1 or even the F2 progeny, the screening for mutants is simplified.

Chemical Mutagenesis

There are two principle mutagenesis strategies that have been used by many labs: chemical and insertional. Chemical mutagenesis most often employs the alkylating agent N-ethyl-N-nitrosourea ($C_3H_6N_3O_2$), called ENU. ENU transfers an ethyl group to bases (usually thymine) in DNA. ENU has been used in *Drosophila,* mice, and zebrafish for saturation screens where the goal was to obtain at least one mutation in every gene. Two saturation mutagenesis screens were conducted a decade ago in zebrafish, the first such screens in a vertebrate. Zebrafish were the choice for identifying the key genes in vertebrate development because the fish produce hundreds of offspring and can be mated within 3 months of birth. Moreover, the optical clarity of the embryos allowed visual analysis of all of the organs as they developed in the space of just a few days. In one screen, 49 males were treated with ENU that on average produced 63 families each for a total of 3,075 families of which 2,746 were analyzed by 14,357 matings that produced 3,857 offspring that had an average of 1.1 mutations per

genome (Haffter et al. 1996). A total of 1,163 mutants were characterized by complementation tests to see whether the mutations were in the same gene, and many were, with 894 of the mutations assigned to 372 genes. In the second screen that was conducted simultaneously, 220 genes were identified from 695 mutants that came from mutagenesis of 651 genomes with 1.5 mutations per genome (Driever et al. 1996).

However, of the combined 592 mutated genes between the two studies, less than 10% could be associated with a particular gene within 5 years. The problem was that ENU produced a single mutation in a genome of nearly 1.7×10^9 bp and although the technique of positional cloning could roughly map the sites of the mutations, the positions of the genes were resolved to regions of about 1 million bps. Indeed, within 5 years only four genes, less than 1%, of those identified by ENU were sequenced and identified using positional cloning. Another 23 were identified by the 'candidate gene approach' wherein one guessed what the gene might be based on findings in other vertebrates such as mice (Talbot and Hopkins 2000).

Looking for ENU-induced mutations in a specific gene, to associate with a phenotype, can be accomplished efficiently with a procedure known as TILLING, named after its developer, Till (2003) who called the procedure Targeting Induced Local Lesions IN Genomes. Essentially this procedure, schematized in Figure 24.7, is a high throughput method to find single bp mutations in known sequences. The first step is

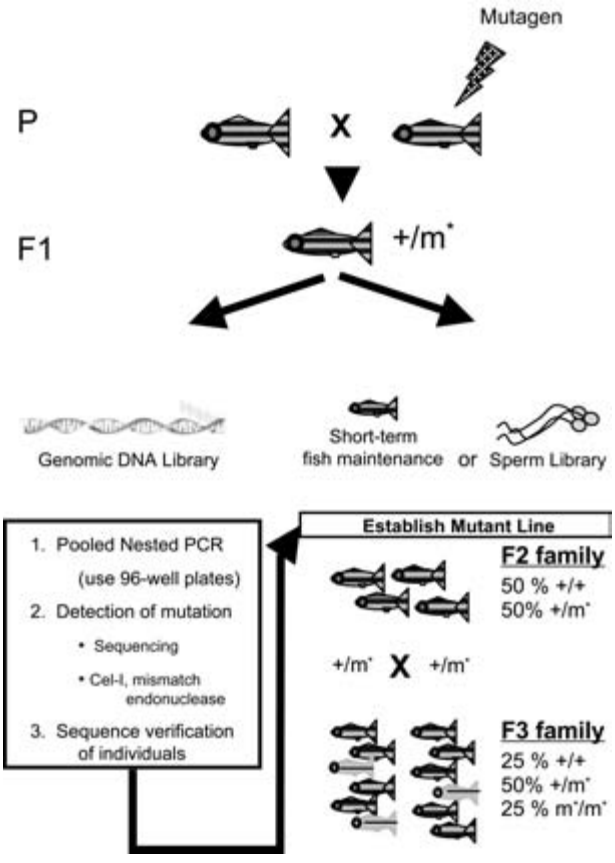


Figure 24.7. TILLING for gene localization. (Adapted from Wienholds et al. 2002.)

to PCR amplify the same locus in wild type and a collection of mutant phenotypes. In fish, this is done by isolating DNA from small sections of fins (fin clips). The second is to denature both sets of double-stranded DNA, mix, and let the mixed strands reanneal to form heteroduplex DNAs wherein a single bp may be mismatched due to the ENU mutation. Cleavage of mismatched DNAs with the plant endonucleases CEL-I produces two bands. In this way hundreds of mutant fish can be screened for mutations in known sequences.

However, when there is not a good candidate gene, finding a gene with a phenotypic mutation requires sequencing millions of bps of both the wild-type region and the mutated region as well as distinguishing background mutations from the mutation of interest. As a result, other methods of mutagenesis were desired that would not only inactivate the gene but also 'tag' it so that it could be easily purified, sequenced, and identified.

Insertional Mutagenesis

The theory behind insertional mutagenesis is that by inserting a known DNA sequence into a genome one both mutates that sequence as well as stamps it with a known identifying tag, as shown in Figure 24.8. The insertional vector, shown as

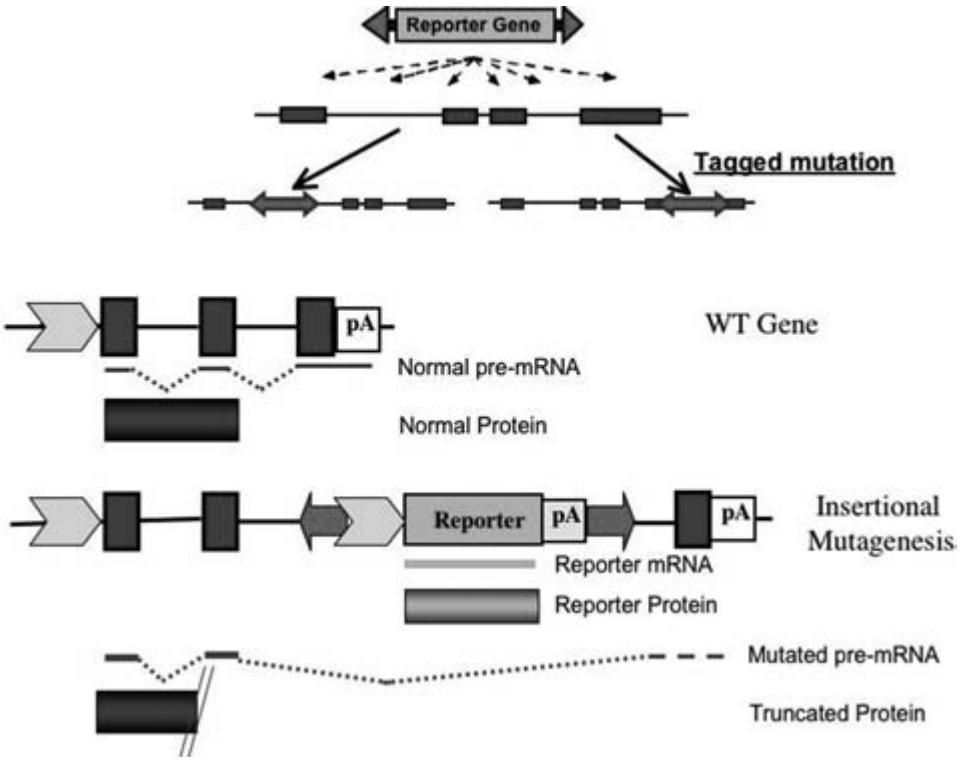


Figure 24.8. Strategies for insertional mutagenesis. The top panel shows that the insertional agent (vector with the reporter gene flanked by inverted arrows) can integrate either between genes (blue boxes) or inside a gene. The bottom panel shows the consequences on transcription and protein synthesis of insertional inactivation of a gene.

inverted arrows, most often contains a reporter gene in order to detect its presence, although this is not necessary and in some cases, for example, in some retroviral vectors described below, the vector does not express a gene.

Nature uses two different approaches for inserting DNA into other DNA sequences. The first is viruses. For insertional mutagenesis, retroviruses, which are associated with cancer, leukemia, and immunodeficiency (e.g., HIV), are used. Retroviruses were first used in mice to infect embryonic stem (ES) cells that then could be raised to produce mutant mice, following the same three-generation crossing scheme shown in Figure 24.6. In one of the largest studies (Zambrowicz et al. 1998), 2,000 genes were disrupted, sequenced, and identified using a mouse retrovirus that had a 3′-gene trap, which is described in more detail in the Trap Vectors section. Retroviruses were quite effective because they are able to integrate their genomes throughout the chromosomes of their host, although preferably close to the transcriptional initiation sites of genes (Wu et al. 2003). A similar procedure was employed with zebrafish except that the fish eggs were not subject to infection due to their chorions. Consequently, a modified murine retrovirus (MLV), related to that used to infect the mouse ES cells, was injected into zebrafish embryos to infect gonadal cells. Because MLV will not infect fish cells, the retroviral envelop glycoprotein gene required for infection of cells was substituted with another glycoprotein gene from the vesicular stomatitis virus to allow infection of fish cells by the xenotropic MLV retrovirus. By using this retrovirus, the Hopkins lab was able to associate more than 50 genes with phenotypes identified by random retroviral mutagenesis or about twice that compared to all the labs combined that worked with the ENU mutants (Talbot and Hopkins 2000). However, there was a problem with this approach. Manufacturing the xenotropic retrovirus is quite difficult, and the virus is able to infect all cells, including those of the human experimenters. Furthermore, it is not clear that retroviruses do not have preferences for integration. Hence, an alternative method was sought.

The second method that nature uses to introduce foreign DNA sequences into chromosomal DNA is transposons. These hopping sequences rarely enter genomes because they have no easy access into cells and nuclei. Consequently, there are few defenses that animals have against transposons, unlike the case with viruses where the immune responses generally quickly suppress viral infections. However, over millions of years, transposons do enter genomes and then become normal residents, often after slowly mutating into inactivity. Transposons come in two types. Retrotransposons translocate via an mRNA intermediate similar to retroviruses. The second type is DNA transposons that move from one DNA sequence to another by a cut-and-paste mechanism (Figure 24.9). DNA transposons have been used for insertional mutagenesis in many nonvertebrate animals such as yeast, fruit flies, nematodes (Zwaal et al. 1993, Rubin et al. 1998, Ross-Macdonald et al. 1999). It was not until an active transposon that could function in vertebrates was artificially created (Ivics et al. 1997) that transposon mutagenesis in vertebrates became possible (Carlson and Largaespada 2007), including in fish (Davidson et al. 2003, Balciunas et al. 2004, Grabher et al. 2002). This transposon is called *Sleeping Beauty* because it was resurrected from an evolutionary sleep that lasted more than 10 million years. Another transposon, *Tol*2, is also used in some labs for insertional mutagenesis in fish (Kawakami et al. 2004). Transposons are simple DNA molecules that can be made and purified in any molecular biology lab and they pose no danger to the investigators using them. Like the xenotropic retroviruses, they need to be injected into embryos

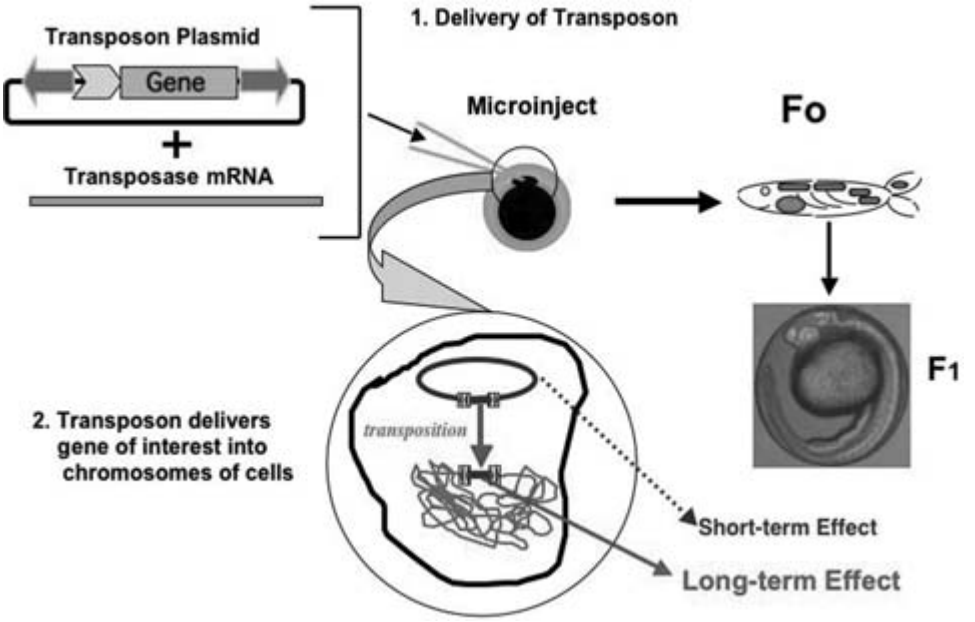


Figure 24.9. Transposon-mediated gene delivery in fish. Injection of *Sleeping Beauty* transposon vectors into a 1-cell fish embryo allows transposition of the transgene in a transposon from the plasmid into a chromosome. The transposon inverted arrows with a promoter-gene inside, is carried on a plasmid. The transposase mRNA is co-injected and expresses the transposase enzyme that cuts the transposon out of the plasmid and integrates it into a chromosome. This leads to long-term expression of the reporter gene, which may interrupt and mutate a gene. Unintegrated plasmids generally express the transgene for only a short period of time. Expression of the transgene is mosaic in Fo animals but they pass on expression of the integrated gene in their offspring. (Also see color plate.)

for insertional mutagenesis. Their activities are comparable to that of retroviruses, but they may have the most random integration patterns of all of the insertional vectors (Yant et al. 2005).

Trap Vectors

A major problem with simple insertional mutagenesis is that the vector will integrate infrequently into exons of genes and more often nearby a gene or into an intron. As a result, most insertions have no effect because the vector sequence is either spliced out of the pre-mRNA transcriptional product (Figure 24.8, when the DNA inserts into an intron) or is ignored if it lands in an untranslated region or in the neighborhood of a transcriptional unit. However, by adding splicing signals and removing either the enhancers or poly(A)-addition sequences, a researcher has a higher assurance that when a vector integrates into a gene, there will be a consequential mutation (Skarnes 2005). Figure 24.10 shows examples of three types of trap vector: (1) gene-trap, (2) 3′-poly(A)-trap, and (3) enhancer-trap. A gene-trap must land in a transcriptional unit for expression of the reporter gene because it lacks its own promoter. By

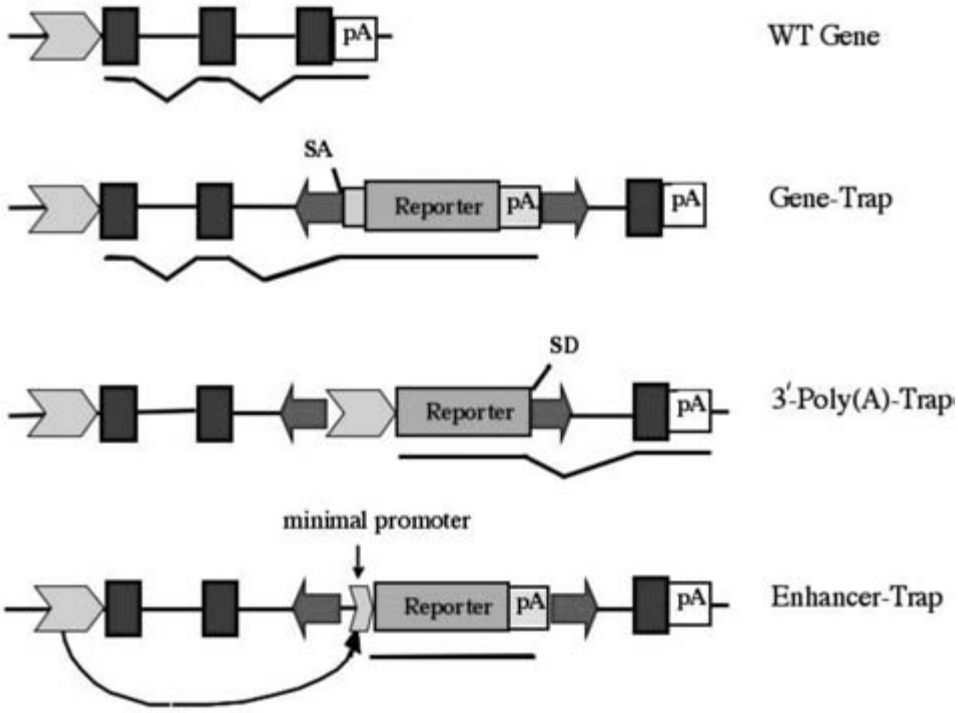


Figure 24.10. Trap vectors are schematized. The consequential expression of the reporter gene is shown for each of the trap vectors. (Adapted from Wadman et al. 2005).

including a splice-acceptor site ahead of the reporter gene, the reporter will be expressed even when the gene-trap lands in an intron. Alternatively, a poly(A) trap vector may have its own promoter, but lack a 3′-poly(A) addition sequence. The pre-mRNA that is made will be unstable and not lead to detectable expression of the reporter unless the 3′-poly(A)-trap lands inside a gene. In this case, a splice donor site is added behind the reporter to ensure that the mRNA from the vector will be expressed when it inserts into an intron. Gene-traps and poly(A)-traps must land in a transcriptional unit for their reporters to be expressed. Expression of the reporter in the gene-trap theoretically will be the same as the gene into which the trap inserted. In contrast, expression of the reporter in the poly(A)-trap will depend on the transcriptional regulators that are carried in the vector; expression of the gene may be constitutive whereas the gene that has been mutated may be highly regulated. An enhancer trap vector has a complete reporter system, but it lacks transcriptional enhancers to boost transcription of the reporter to detectable levels. Reporter genes in enhancer traps can be expressed whether the trap inserts into a gene or in the neighborhood of a gene. Since the local enhancers will activate the reporter gene in the same manner as they would the gene with which they are normally associated, the enhancer-trap theoretically has the same expression profile as the gene in which or next to which it inserted. This type of vector may not mutate the genes into which it inserted, but it can indicate the presence of transcriptional enhancers in the region into which it integrated and thereby identify a transcriptional unit. Further combinations

of a gene trap and 3′-poly(A)-trap can be constructed as well. Trap vectors, generally in transposons, are being employed more extensively over time in model fish species such as zebrafish (Davidson et al. 2003, Grabher et al. 2003, Clark et al. 2004, Parinov et al. 2004, Kawakami et al. 2004, Balciunas et al. 2004, Wadman et al. 2005).

Reverse Genetics: Knocking Out and Knocking Down Genes

Forward genetics using random mutagenesis is one way of finding new genes. However, due to the conservation of function between many, if not most genes, clues of gene function can be gained in one organism for application in another. In this case, investigators often want to unequivocally inactivate a gene or abolish its function for at least a defined period of time. This process is called reverse genetics. There are general methods of reverse genetics: inactivating a specific genetic locus by homologous recombination or inactivating expression of the gene by blocking its mRNA. These strategies are discussed below.

Gene Knockouts

The standard for this procedure has been well established in mice (Carlson and Largaespada 2006). Genes can be knocked out by homologous recombination in embryonic stem cells to mutate a particular genetic sequence that is precise to the bp. Following selection for a stem cell with the desired mutation, a culture is grown for introduction into an embryo to obtain an animal with a specific mutation. Recombination of an integrating DNA sequence into a homologous sequence occurs at a very low rate in mammalian cells, about 10^{-6} to 10^{-7} (Mansour et al. 1988). Hence, there must be a selection for such rare events. As shown in Figure 24.11, by flanking a selectable marker by several kb of DNA sequence on each side, a specific insertion can be obtained in a specific gene. The heterozygous animal that emerges, sometimes, can be outbred and those progeny then crossed with each other (Figure 24.7) to obtain homozygous knockout animals that are deficient in both alleles. The key to using this technology is having embryonic stem cells that can be transformed with the appropriate DNA sequence for further propagation of the mutated gene. In fish, this has been an ongoing quest that has only sporadically been achieved in a few cells in culture (Fan et al. 2006), with or without a recombination aid such as the *E. coli* RecA protein (Cui et al. 2003), or in model fish such as medaka (Hong et al. 2004) and zebrafish (Ma et al. 2001) with pluripotent cells. An alternative procedure that is under current development is transfer of nuclei from somatic cells, a process that has some problems that need solving before it becomes a reliable technique (Rhind et al. 2003).

Gene Knockdowns

As in the sport of boxing, a genetic knockdown is not as permanent as a genetic knockout. Because of the inability to reliably achieve knockouts in vertebrates other than mice, alternative procedures have been developed to achieve essentially the same effects. These approaches have concentrated on blocking expression from mRNA by either enhancing its degradation or blocking its ability to be translated. Two general strategies have been used. The first is to use natural RNA, interfering

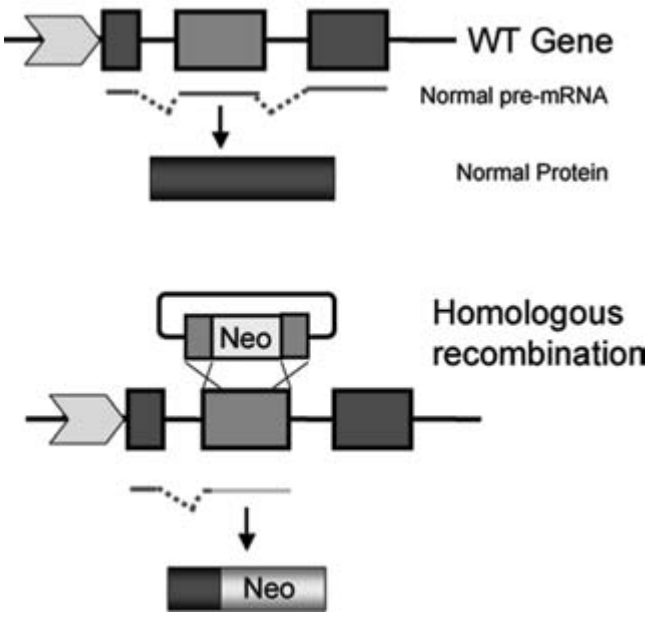


Figure 24.11. Specific inactivation, knockout, of a genetic locus by recombination of a foreign sequence that is flanked by sequences in the target locus. The inserted sequence generally has a reporter gene in order to detect its presence. Since the insertion is the result of pairing of the flanking sequences in the inserted gene, the sequence of the locus with the insert is known exactly. Consequently, the transcriptional regulators of the targeted locus can be used to express the reporter gene. In this example the Neo gene is used to exactly replace exon-2 of the target gene.

RNA (RNAi) as an inhibitor of mRNA function. The second is to introduce artificial oligonucleotides that have essentially the same effect, although they may work through different mechanisms.

RNAi-Mediated Inhibition of Gene Expression

There are several types of RNA that are referred to in several ways that regulate gene expression by inducing the degradation of specific mRNAs. This is a result of the merging of several investigations in plants and animals where processes referred to as post-transcriptional gene silencing and co-suppression were observed. The unifying concept was that short RNAs, commonly referred to as RNAi, could block functional mRNAs. RNAi (interfering RNA), siRNA (silencing RNA), miRNA (microRNAs associated with binding to the 3′ ends of mRNAs), and shRNA (short-hairpin RNA that is a common source of siRNA) all are processed by the dicer/RISC system to degrade mRNA. Figure 24.12 shows the basic elements of the dicer/RISC system. Essentially, animal and even plant cells have a defensive mechanism that recognizes double-stranded RNA, which is often associated with viral infections. In 1998, Fire and his colleagues recognized that the presence of antisense RNA as a part of double-stranded (ds) RNA in cells led to specific inhibition of gene expression in nematodes (Fire et al. 1998). In

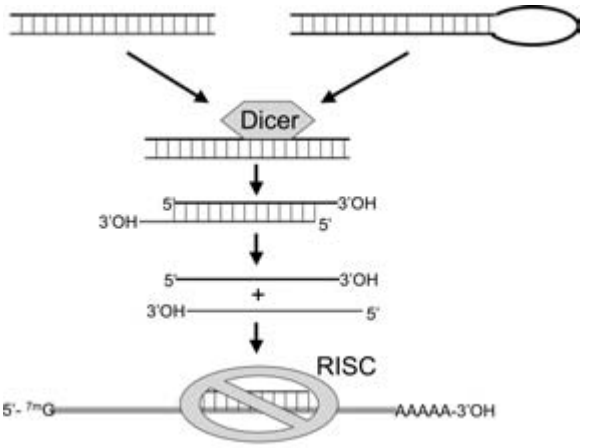


Figure 24.12. Production of RNAi to inhibit mRNA expression. Double-stranded RNA (dsRNA) from same source, top line, is recognized by the Dicer enzyme and cleaved into shorter, dsRNA segments of 21–23 bp with protruding 3′-ends. The strands of the short dsRNA then can denature and reanneal to a complementary sequence on an mRNA (double bar lines), which upon recognition by RISC will lead to mRNA degradation. (Also see color plate.)

the ensuing years, a variety of sources of RNA forming dsRNA were found that led to the production of shorter, duplex RNA molecules composed of 21–23 nucleotides per single strand that had protruding 3′ ends (Figure 24.12). Dicer is the RNaseIII-like enzyme that cleaves dsRNA into the bite-sized, ds 21–23 nucleotide segments. Denaturation of the dsRNA into single strands allows them to bind to complementary sequences on mRNAs to form base paired RNA structures that are recognized by RISC (RNA-induced silencing complex). All of the various short RNAs mentioned above regulate mRNA expression through the common Dicer-RISC pathways.

RNAi-based knockdown strategies have been the subject of thousands of papers. Yet, despite successes in lower animals such as nematodes (Fraser et al. 2000, Gönczy et al. 2000) their track record in vertebrates (e.g., Ngo et al. 2006) is mixed. In particular, RNAi has not been effective in fish for reducing gene expression to the point that gene function is compromised. As shown in Figure 24.13, there is enormous variation in the residual level of gene expression that depends on the particular RNAi sequence. Moreover, both functional genomics (Sumanas and Larson 2002) and gene therapy (Hackett et al. 2004) studies have shown that as little as 1–5% normal activity of many genes is sufficient for nearly normal activities; that is, it is extremely difficult to score an abnormal phenotype unless the expression of a gene is reduced by greater than 98%. This is rarely achieved in vertebrates with RNAi strategies because high expression of RNAi leads to off-target effects. When RNAi levels are sufficient to block more than 95% expression of a targeted gene in vertebrates, often expression of

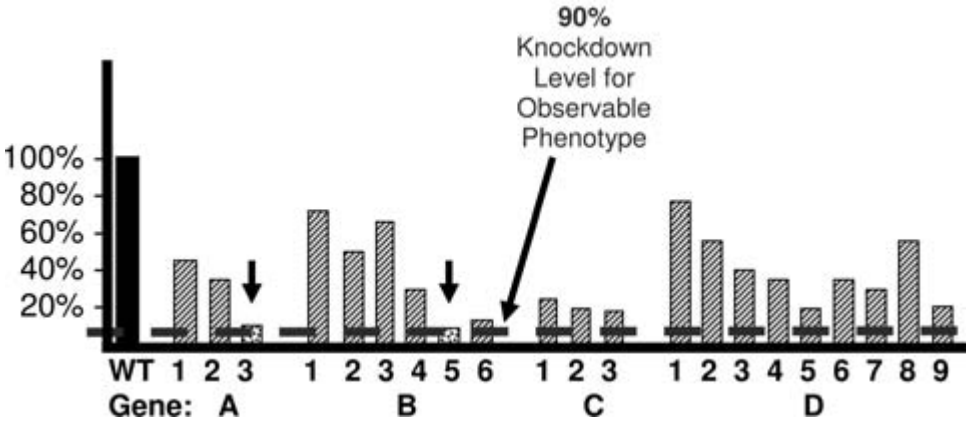


Figure 24.13. Examples of RNAi-mediated reduction in gene expression with respect to different RNAi sequences targeted to the same mRNA. The dashed line indicates the minimal threshold below which gene expression must be reduced for a visible phenotype. Relative expression of four genes, A-D, are shown following treatment with various RNAi sequences (numbers); wild-type (WT) is normalized to 100% for each gene.

other genes is also affected (Jackson et al. 2004, Behlke 2006). Consequently, few gene functions in vertebrate animals, including fish (Zhao et al. 2001), have been discovered using RNAi techniques. As a result, alternative methods of blocking expression from mRNA have been developed that make use of modified oligonucleotides. Nevertheless, there remains intensive effort to find methods that will achieve effective knockdown of gene expression for functional genomics using RNAi-based approaches. This is because RNAi can be synthesized from expression cassettes, notably in the form of precursor shRNAs. This allows a continued presence of the inhibitory RNA over a prolonged period or, alternatively, in selected cell types as a result of appropriate selection of promoters that direct the synthesis of the RNAi.

Oligonucleotide-Mediated Inhibition of Gene Expression

RNA: DNA hybrids rarely exist except transiently during DNA transcription in the nucleus and during infection by viruses in the cytoplasm. As a defense against viral infection, vertebrate cells have an RNaseH activity that degrades the mRNA associated with DNA (Figure 24.14). Although single-stranded cDNAs are effective, they are not stable and will be degraded over time. Consequently, a number of other modifications to oligonucleotide structure were developed that concentrated on their stability in the cell, the stability of the duplex, their abilities to mobilize RNaseH activities, and their ease of synthesis (reviewed in Freier and Altmann 1997). Oligonucleotides with unusual structures such as having a peptide linkage (peptide nucleic acids [PNA]) or altered ring structures so that they were not classical nucleic acids have been employed.

In addition to directing the cleavage of a population of mRNAs, there is an alternative strategy for inhibiting expression that involves blocking either the formation or translation of a targeted mRNA. Morpholino phosphorodiamidate oligonucleotides

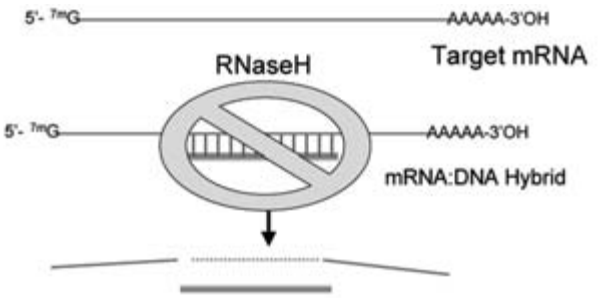


Figure 24.14. RNase H-mediated cleavage of mRNA segments (red lines) base paired to a DNA sequence (blue lines).

(PMO), as shown in Figure 24.15, are particularly useful for this method of gene knockdown. PMOs have a 6-member morpholine ring that replaces the 5-member ribose/deoxyribose rings in normal nucleic acids and the linkage between the bases, which remain the same, is a neutral phosphorodiamidate in which the slight positive charge on the nitrogen balances the slight negative charge on the oxygen in contrast to the phosphodiester backbone of DNA that is negatively charged. The consequence of these changes is that the PMOs can distribute throughout cells (Nasevicius and Ekker 2000). Because of their unusual structure, PMOs do not elicit an RNaseH attack on the mRNA to which they bind. Rather, they act as steric blocks to preinitiating ribosomal subunits in the cytoplasm as well as to the splicing enzymes in the nucleus to block either translation or proper processing of pre-mRNAs (Sumanas and Larson 2002). For this to happen, a 21-base PMO must be properly designed to blanket the initiating AUG codon (Figure 24.16) or equivalently a splicing donor or acceptor site. Most applications of PMOs are to block translation since in some circumstances alternative splicing can generate families of proteins from single genes. In general gene expression using PMOs is reduced more than 98% at doses that are nontoxic, which leads to

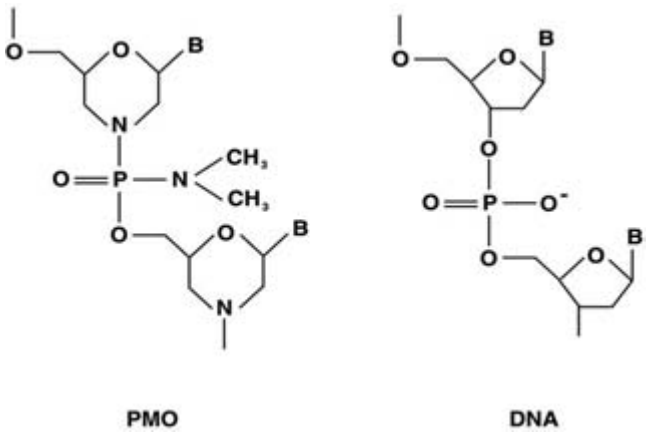


Figure 24.15. Comparison of structures of two bases (B) linked by a phosphodiester linkage that is normally found in DNA (right) and a morpholino phosphorodiamidate (PMO) linkage (left).

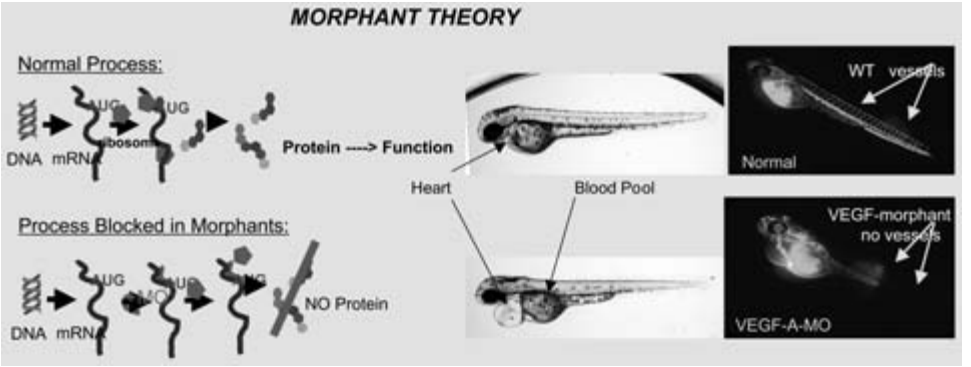


Figure 24.16. Mechanism of action of PMOs and an example of a PMO directed against VEGF in zebrafish. The top row of figures shows normal 48-hour zebrafish embryos with normal vascular that is highlighted by the fluorescence in the blood vessels (right side). The heart is tucked under the head and the blood is distributed throughout the embryo in the vasculature. The bottom row of figures shows 48-hr VEGF Morphant embryos with a grossly enlarged heart and pooled blood (in the left image). The vasculature is not developed. The embryo is alive and can be subjected to further studies. (Also see color plate.)

observable phenotypes at a much higher rate than with RNAi-based strategies. The resulting phenotype from treatment of an embryo or animal with a PMO is called a morphant, in parallel with the term mutant that results from a mutation.

PMO-based gene knockdown has been used extensively in zebrafish, but less so in other fish. This is due to the relatively short effective lifetime of the morpholino, which although it is not degraded, is lost by excretion just as any normal cellular metabolite. PMOs have an effective lifetime of about 4 days, although some functions out to 8 days have been reported. Because zebrafish develop most of their basic organs within this period, many developmental functions can be studied using PMOs. However, adult or later onset functions cannot easily be blocked using PMOs. For those early functions that can be addressed by PMOs, they represent an extremely rapid, cost-effective method for determining the effects of blocking a particular gene's function *in vivo.*

Nevertheless, PMOs and all other artificial oligonucleotide-based inhibitory strategies have the same limitation, an inability to be delivered to every cell of target organs in an older animal. This is a serious drawback and explains the ongoing desire to develop a method that will allow synthesis of an inhibitor that will be able to reduce gene expression by more than 98% without off-target effects.

Dominant-Negative Inhibition of Gene Expression

An alternative forward genetics approach to block expression of a targeted gene is to construct a protein that will interfere with the targeted protein. Dominant negative proteins, first observed as a consequence of antimorphic mutations, act antagonistically to the product of the natural gene (Hudson et al. 2002). For example, modification of a signaling protein that interacts as a dimer with receptors on the cell surface can block dimerization but not binding. This protein will compete with the natural signaling protein and reduce signaling rates. This strategy is useful only when a substantial amount of information is available for the gene. It has not been widely

employed in functional genomics studies, but it is a feasible approach for fine-tuning an understanding of gene interactions.

Over and Ectopic Expression of Genes

The loss-of-function techniques described in the previous sections are relatively difficult. Gene knockouts are extremely difficult and unreliable in all vertebrates except mice, and inhibition by RNAi-based or modified oligonucleotide-based strategies have problems of (1) efficient delivery, (2) enhanced stability, (3) minimization of off-target sites, and (4) identification of sensitive sites in the target RNAs. These problems are acute in their applications to larger fish and other animals. Hence, an alternative approach to investigating gene function has been used, such as gain-of-function or over-expression of genes to determine their effects on physiological and developmental processes. Over-expression strategies have the advantage that they rely on a substantial body of knowledge and can be controlled by the investigator. In general, gain-of-function genetic screens also require delivery of the DNA construct to 1-cell embryos or ES cells in order to obtain animals with the desired phenotypes, although proceeding through a 3-generation screen is not necessary because homozygosity of the additional genetic unit is not necessary (e.g., Pritsker et al. 2006). Their drawback is that determining the functions of genes from over-expression studies is more difficult than from knockout/knockdown studies because of the background effects of the endogenous gene(s). Many times over-expression has no obvious consequence to the phenotype.

There are several strategies for over-expression that can be employed. These include over-expression of a gene product from (1) delivery of the mRNA, a very direct but very transient method that is only good in 1-cell embryos because of the problem of delivery, (2) delivery of expression vectors on extra-chromosomal plasmids, and (3) delivery of integrating vectors such as viruses and transposons, which were discussed earlier in the Insertional Mutagenesis section, that direct integration of the transgenic construct and thereby allow prolonged expression of the transgene. Transgenic fish have been extensively reviewed (e.g., Hackett and Alvarez 2000, Zbikowska 2003). The important conclusions from most of the studies prior to 2000 came from generally transient expression from unintegrated genes or integrated plasmids whose prokaryotic origins were recognized by the host cells, which would methylate and otherwise shut down expression of the transgene. The most important finding was the high conservation of regulatory motifs in fish and other vertebrates, this was found early from the intensive investigations of the β-actin gene in carp compared to the homologous gene in land vertebrates (Liu et al. 1990a, 1990b, 1990c; Liu et al. 1991). This is extremely important because it means that expression cassettes based on mice and other thoroughly studied vertebrates can be used in fish, with the obvious understanding that all such assumptions must be tested because exceptions to the rule occur.

The early transgenic studies were of three sorts:

1. delivery of one of a very few genes, such as those encoding growth hormones and antifreeze proteins, into fish of commercial importance
2. delivery of fluorescent protein reporter genes into model fish such as zebrafish to test transcriptional regulatory motifs or to act as gene traps, as discussed earlier
3. delivery of genes involved with vertebrate development into model fish for basic studies of physiology

Most of these studies involved strong, promiscuous enhancer/promoter expression cassettes that led to ectopic expression of the transgene in most tissues. Use of such promoters rarely allows detailed investigation of gene function because if the level of gene expression is important to the phenotype, the animal is likely to fail over the course of development.

Consequently, ongoing studies and those in the future will employ tissue-specific promoters or genes whose expression can be controlled by regulated deletion of the promoter. Because the chromosomal sites of vector integration cannot be controlled at this time, expression of transgenes can be influenced by enhancers and silencers that are close to the integration site. As a result, an alternative method of regulating gene expression is to use a recombination system such as FLP/FRT or Cre/loxP (O'Gorman et al. 1991, Bunting et al. 1999) that have been enormously successful in mice. The Cre/lox system functions in zebrafish (Pan et al. 2005). Accordingly, a Cre recombinase gene can be put behind an inducible promoter to achieve temporal expression, or behind a tissue-specific promoter to achieve spatial regulation of expression. Cre recombinase will delete DNA sequences that are flanked by 34-bp (two 13-bp inverted repeats separated by 8 bp) lox sequences. Thus, Cre-mediated recombination can remove a sequence that is flanked by lox sites by recombination in order to activate the sequence by removal of a spacer sequence between a gene and its promoter or inactivate a gene by removing its promoter if flanked by lox sequences (Figure 24.17). The deletion of the floxed sequence (sequence flanked by lox sites) can be controlled by placing the Cre gene behind an inducible promoter. For regulation of expression in the whole animal, the tetR system of Gossen and Bujard (1992) can be used wherein a metabolite such as doxycycline is injected or otherwise delivered to the animal or a tissue. Alternatively, the Cre gene can be placed behind a naturally regulated promoter, such as that for prolactin, etc. These tools are under development in fish and are being refined in mammalian model systems such as mice and rats.

Future Directions

At the beginning of this chapter we noted that the goals of functional genomics are to better understand the roles of elements of the genome that directly or indirectly affect the functioning of an organism. We have seen that high throughput methods are available to investigators to examine gene identities and their roles in physiological processes. However, there remains a need to develop methods for inactivating genes in fish (and other vertebrates) if a clear understanding of all genes is to be achieved. In particular, the efforts to obtain embryonic stem cells, or equivalent multifated progenitor cells, are important goals for the future.

For all vertebrates, we have just begun to elucidate the detailed genetic pathways of development in multicellular animals. Such efforts have been initiated in nematodes and other invertebrates that have a few hundred, well-defined cells (Davidson and Erwin 2006, Zhong et al. 2006). In vertebrates where the numbers of genes are significantly larger and the number of differentiated cell types is greater, alternative, computationally driven approaches for determining genetic interactions are being developed. These include the systematic identification of regulatory modules that

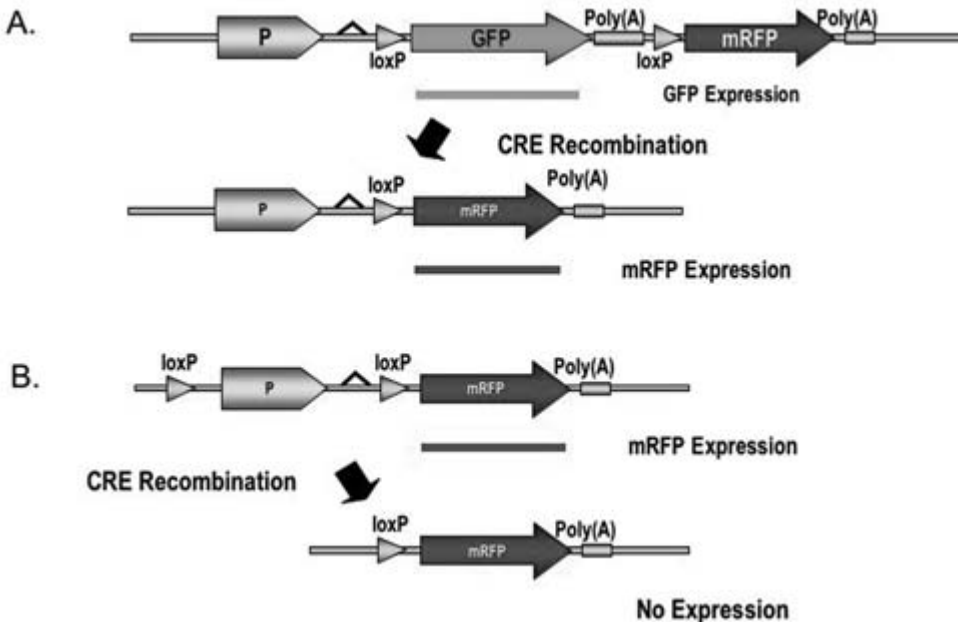


Figure 24.17. Use of the Cre/lox recombination to (A) activate gene expression or (B) inhibit gene expression. Two fluorescent reporter genes are shown. (A) The floxed GFP expression cassette separates the promoter (P) from the mRFP gene, which consequently is not expressed. Cre-mediated recombination leaves only a single *loxP* site between the promoter and mRFP gene, which then can be expressed. (B) The promoter is floxed and upon induction of Cre, the promoter is removed thereby keeping the mRFP gene from being expressed.

distinguish elevated from inhibited expression in the corresponding mRNAs to reveal synergy between *cis*-regulatory modules and explain large-scale tissue-specific differential expression (Smith et al. 2006). For instance, already distinct genetic interactions between multiple vascular endothelial growth factor (VEGF) receptors have been found that are required for development of different blood vessel types in zebrafish (Covassin et al. 2006). As bioinformatics tools are enhanced along with our abilities to deliver a variety of genetic expression cassettes for over-expression as well as blocking expression, our rates of understanding the basic physiological pathways in various species will increase. The relatively short history of functional genomics has shown us that there will be ever more routes, yet to be found, to take to achieve our goals.

Acknowledgments

We thank the Arnold and Mabel Beckman Foundation for support of our work and all members of the Beckman Center for Transposon Research for a long history of contributions of ideas and discussions. The authors were supported by NIH grant 1RO1-DA14546-01 and grants from the USDA (2005-35604-15615) and the Juvenile Diabetes Foundation (2005-1167).

References

Adams MD, JM Kelley, JD Gocayne, M Dubnick, MH Polymeropoulos, H Xiao, CR Merril, A Wu, B Olde, RF Moreno, AR Kerlavage, WR McCombie, and JC Venter. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. Science, 252, pp. 1651–1656.

Amsterdam A and N Hopkins. 2006. Mutagenesis strategies in zebrafish for identifying genes involved in development and disease. Trends Genet, 22, pp. 473–478.

Ausubel FM, R Brent, RE Kingston, DD Moore, JG Seidman, JA Smith, and K Struhl. 1998. *Current Protocols in Molecular Biology.* John Wiley & Sons. New York.

Balciunas D, AD Davidson, S Sivasubbu, Z Welle, and SC Ekker. 2004. Enhancer trapping in zebrafish using the *Sleeping Beauty* transposon. BMC Genomics, 5, 62.

Behlke MA. 2006. Progress towards in vivo use of siRNAs. Mol. Ther., 13, pp. 644–670.

Benson JD, YN Chen, SA Cornell-Kennon, M Dorsch, S Kim, M Leszczyniecka, WR Sellers, and C Lengauer. 2006. Validating cancer drug targets. Nature, 441, pp. 451–456.

Bunting M, KE Bernstein, JM Greer, MR Capecchi, and KR Thomas. 1999. Targeting genes for self-excision in the germ line. Genes Dev, 13, pp. 1524–1528.

Carlson C and DA Largaespada. 2006. Insertional mutagenesis in mice: new perspectives and tools. Nature Rev. Genet., 6, pp. 568–580.

Cheng KC, R Levenson, and JD Robishaw. 2003. Functional genomic dissection of multimeric protein families in zebrafish. Dev. Dyn., 228, pp. 555–567.

Clark KJ, AM Geurts, JB Bell, and PB Hackett. 2004. Transposon vectors for gene-trap insertional mutagenesis in vertebrates. Genesis, 39, pp. 225–233.

Covassin LD, JA Villefranc, MC Kacergis, BM Weinstein, and ND Lawson. 2006. Distinct genetic interactions between multiple Vegf receptors are required for development of different blood vessel types in zebrafish. Proc Natl Acad Sci USA, 103, pp. 6554–6559.

Cui Z, Y Yang, CD Kaufman, D Agalliu, and PB Hackett. 2003. RecA-mediated, targeted mutagenesis in zebrafish. Mar. Biotech., 5, pp. 174–184.

Dahm R and R Geisler. 2006. Learning from small fry: the zebrafish as a genetic model organism for aquaculture fish species. Mar. Biotech., 8, pp. 329–345.

Davidson AE, D Balciunas, D Mohn, J Shaffer, S Hermanson, S Sivasubbu, PB Hackett, and SC Ekker. 2003. Efficient gene delivery and expression in zebrafish using *Sleeping Beauty.* Dev. Biol., 263, pp. 191–202.

Davidson EH and DH Erwin. 2006. Gene regulatory networks and the evolution of animal body plans. Science, 311, pp. 796–800.

Driever W, L Solnica-Krezel, AF Schier, SCF Neuhaus, J Malicki, DL Stemple, DYR Stanier, F Zwartkruis, S Abdelilah, Z Rangini, J Belak, and C Boggs. 1996. A genetic screen for mutations affecting embryogenesis in zebrafish. Development, 123, pp. 37–46.

Dukes JP, R Deaville, D Gottelli, JE Neigel, MW Bruford, and WC Jordan. 2006. Isolation and characterisation of main olfactory and vomeronasal receptor gene families from the Atlantic salmon (*Salmo salar*). Gene, 371, pp. 257–267.

Eddy SR. 2006. Non-coding RNA genes and the modern RNA world. Nature Rev. Genet., 2, pp. 919–925.

Fan L, J Moon, J Crodian, and P Collodi. 2006. Homologous recombination in zebrafish ES cells. Transgenic Res., 15, pp. 21–30.

Fire A, SQ Xu, MK Montgomery, SA Kostas, SE Driver, and CC Mello. 1998. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans.* Nature, 391, pp. 808–811.

Fraser AG, RS Kamath, P Zipperlen, M Martinez-Campos, M Sohrmann, and J Ahringer. 2000. Functional genomic analysis of *C. elegans* chromsome I by systematic RNA interference. Nature, 408, pp. 325–330.

Freier SM and HK Altmann. 1997. The ups and downs of nucleic acid duplex stability: structure-stability studies on chemically-modified DNA-RNA duplexes. Nucl. Acids Res., 25, pp. 4429–4443.

Gönczy P, C Echeverri, K Oegema, A Coulson, SJ Jones, RR Copley, J Duperon, J Oegema, M Brehm, E Cassin, E Hannak, M Kirkham, S Pichler, K Flohrs, A Goessen, S Leidel, AM Alleaume, C Martin, N Ozlu, P Bork, and AA Hyman. 2000. Functional genomic analysis of cell division in *C. elegans* using RNAi of genes on chromsome III. Nature, 408, pp. 331–336.

Gossen M and H Bujard. 1992. Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. Proc Natl Acad Sci USA, 89, pp. 5547–5551.

Grabher C, T Henrich, T Sasado, A Arenz, J Wittbrodt, and M Furutani-Seiki. 2003. Transposon-mediated enhancer trapping in medaka. Gene, 322, pp. 57–66.

Gregory TR. 2005. Genome size evolution in animals. In *The Evolution of the Genome,* T.R. Gregory, Ed. Elsevier, San Diego, pp. 3–87.

Hackett PB and MC Alvarez. 2000. The molecular genetics of transgenic fish. Recent Adv. Mar Biotech, 4, pp. 77–145.

Hackett PB, KJ Clark, SC Ekker, and JG Essner. 2004. Applications of transposable elements in fish for transgenesis and functional genomics. In Chapter 16 *Fish Development and Genetics,* Z Gong and V Korzh, Eds., World Scientific, Singapore, pp. 532–580.

Hackett PB, SC Ekker, DA Largaespada, and RS McIvor. 2004. *Sleeping Beauty* transposon-mediated gene therapy for prolonged expression. In *Non-Viral Vectors for Gene Therapy,* 2nd Edition, L Huang, E Wagner, and M-C Hung, Eds. Adv Genet, 54, pp. 187–229.

Haffter P, M Granato, M Brand, MC Mullins, M Hammerschmidt, DA Kane, J Odenthal, FJ van Eeden, YJ Jiang, CP Heisenberg, RN Kelsh, M Furutani-Seiki, E Vogelsang, D Beuchle, U Schach, C Fabian, and C Nusslein-Volhard. 1996. The identification of genes with unique and essential functions in the development of the zebrafish, *Danio rerio.* Development, 123, pp. 1–36.

Hancock JM. 2005. Gene factories, microfunctionalization and the evolution of gene families. Trends Genet, 21, pp. 591–595.

Hastie ND and JO Bishop. 1976. The expression of three abundance classes of messenger RNA in mouse tissues. Cell, 9, pp. 761–774.

Holmes C and SDM Brown. 2004. All systems GO for understanding mouse gene function. J Biol, 3, p. 20.

Hong Y, S Chen, J Gui, and M Schartl. 2004. Retention of the developmental pluripotency in medaka embryonic stem cells after gene transfer and long-term drug selection for gene targeting in fish. Transgenic Res, 13, pp. 41–50.

Hudson DF, C Morrison, S Ruchaud, and WC Earnshaw. 2002. Reverse genetics of essential genes in tissue-culture cells: 'dead cells talking.' Trends Cell Biol, 12, pp. 281–287.

Humphreys DT, BJ Westman, DIK Martin, and T Preiss. 2005. MicroRNAs control translation initiation by inhibiting eukaryotic initiation factor 4E/cap and poly(A) tail function. Proc Natl Acad Sci USA, 102, pp. 961–966.

Hurles M. 2004. Gene duplication: the genomic trade in spare parts. PLoS Biol, 2, pp. 900–904, e206.

Ivics Z, PB Hackett, RA Plasterk, and Z Izsvak. 1997. Molecular reconstruction of *Sleeping Beauty:* a Tc1-like transposon from fish and its transposition in human cells. Cell, 91, pp. 501–510.

Jackson AL and PS Linsley. 2004. Noise amidst the silence: off-target effects of siRNAs? Trends Genet, 20, pp. 521–524.

Kaufman CD, G Martínez-Rodriguez, and PB Hackett. 2000. Ectopic expression of c-*ski* disrupts gastrulation and neural patterning in zebrafish. Mech Dev, 95, pp. 147–162.

Kawakami K, H Takeda, N Kawakami, M Kobayashi, N Matsuda, and M Mishina. 2004. A transposon-mediated gene trap approach identifies developmentally regulated genes in zebrafish. Dev Cell, 7, pp. 133–144.

Kurosawa G, N Takamatsu, M Takahashi, M Sumitomo, E Sanaka, K Yamada, K Nishii, M Matsuda, S Asakawa, H Ishiguro, K Miura, Y Kurosawa, N Shimizu, Y Kohara, and H Hori. 2006. Organization and structure of *hox* gene loci in medaka genome and comparison with those of pufferfish and zebrafish genomes. Gene, 370, pp. 75–82.

Liu Z, A Karsi, and RA Dunham. 1999. Development of polymorphic EST suitable for genetic linkage mapping of catfish. Mar Biotech, 1, pp. 437–447.

Liu Z, B Moav, A Faras, K Guise, AR Kapuscinski, and PB Hackett. 1990b. Functional analysis of elements affecting expression of the β-actin gene of carp. Mol Cell Biol, 10, pp. 3432–3440.

Liu Z, B Moav, A Faras, K Guise, AR Kapuscinski, and PB Hackett. 1990c. Development of expression vectors for transgenic fish. Bio/Technology, 8, pp. 1268–1272.

Liu Z, B Moav, A Faras, K Guise, AR Kapuscinski, and PB Hackett. 1991. Importance of the CArG box in regulation of β-actin-encoding genes. Gene, 108, pp. 211–217.

Liu Z, Z Zhu, K Roberg, A Faras, K Guise, AR Kapuscinski, and PB Hackett. 1990a. Isolation and characterization of the β-actin gene of carp (*Cyprinus carpio*) DNA Seq, 1, pp. 125–136.

Ma C, L Fan, R Ganassin, N Bols, and P Collodi. 2001. Production of zebrafish germ-line chimeras from embryo cell cultures. Proc Natl Acad Sci USA, 98, pp. 2461–2466.

Mansour SL, KR Thomas, and MR Capecchi. 1988. Disruption of the proto-oncogene int-2 in mouse embryo-derived stem cells: a general strategy for targeting mutations to non-selectable genes. Nature, 336, pp. 348–352.

Mathavan S, SG Lee, A Mak, LD Miller, KR Murthy, KR Govindarajan, Y Tong, YL Wu, SH Lam, H Yang, Y Ruan, V Korzh, Z Gong, ET Liu, and T Lufkin. 2005. Transcriptome analysis of zebrafish embryogenesis using microarrays. PLoS Genetics, 1, p. e29.

Modiano G, G Battistuzzi, and AG Motulsky. 1981. Nonrandom patterns of codon usage and of nucleotide substitutions in human alpha- and beta-globin genes: an evolutionary strategy reducing the rate of mutations with drastic effects? Proc Natl Acad Sci USA, 78, pp. 1110–1114.

Mulley JF, CH Chiu, and PW Holland. 2006. Breakup of a homeobox cluster after genome duplication in teleosts. Proc Natl Acad Sci USA, 103, pp. 10369–10372.

Nasevicius A and SC Ekker. 2000. Effective targeted gene 'knockdown' in zebrafish. Nature Genet, 26, pp. 216–220.

Ngo VN, RE Davis, L Lamy, X Yu, H Zhao, G Lenz, LT Lam, S Dave, L Yang, J Powell, and LM Staudt. 2006. A loss-of-function RNA interference screen for molecular targets in cancer. Nature, 441, pp. 106–110.

O'Gorman S, DT Fox, and GM Wahl. 1991. Recombinase-mediated gene activation and site-specific integration in mammalian cells. Science, 251, pp. 1351–1355.

Pan X, H Wan, W Chia, Y Tong, and Z Gong. 2005. Demonstration of site-directed recombination in transgenic zebrafish using the *Cre*/*loxP* system. Transgenic Res, 14, pp. 217–223.

Parinov S, I Kondrichin, V Korzh, and A Emelyanov. 2004. *Tol2* transposon-mediated enhancer trap to identify developmentally regulated zebrafish genes *in vivo.* Dev Dyn, 231, pp. 449–459.

Pritsker M, NR Ford, HT Jenq, and IR Lemischka. 2006. Genomewide gain-of-function genetic screen identifies functionally active genes in mouse embryonic stem cells. Proc Natl Acad Sci USA, 103, pp. 6946–6951.

Rhind SM, JJ King, LM Harkness, C Bellamy, W Wallace, P DeSousa, and I Wilmut. 2003. Cloned lambs—lessons from pathology. Nature Biotech, 21, pp. 744–745.

Rise ML, KR von Schalburg, GD Brown, MA Mawer, RH Devlin, et al. 2004. Development and application of a salmonid EST database and cDNA microarray: data mining and interspecific hybridization characteristics. Genome Res, 14, pp. 478–490.

Ross-Macdonald P, PSR Coelho, T Roemer, S Agarwal, A Kumar, R Jansen, KH Cheung, A Sheehan, D Symoniatis, L Umansky, M Heidtman, FK Nelson, H Iwasaki, K Hager,

M Gerstein, P Miller, GS Roeder, and M Snyder. 1999. Large-scale analysis of the yeast genome by transposon tagging and gene disruption. Nature, 402, pp. 413–418.

Rubin GM. 1998. The *Drosophila* genome project: A progress report. Trends Genet, 14, pp. 340–345.

Skarnes WC. 2005. Two ways to trap a gene in mice. Proc Natl Acad Sci USA, 102, pp. 13001–13002.

Smith AD, P Sumazin, Z Xuan, and MQ Zhang. 2006. DNA motifs in human and mouse proximal promoters predict tissue-specific expression. Proc Natl Acad Sci USA, 103, pp. 6275–6280.

Song HD, XJ Sun, M Deng, GW Zhang, Y Zhou, XY Wu, Y Sheng, Y Chen, Z Ruan, CL Jiang, HY Fan, LI Zon, JP Kanki, TX Liu, AT Look, and Z Chen. 2004. Hematopoietic gene expression profile in zebrafish kidney marrow. Proc Natl Acad Sci USA, 101, pp. 16240–16245.

Sumanas S and JD Larson. 2002. Morpholino phosphorodiamidate oligonucleotides in zebrafish: a recipe for functional genomics? Brief Func Genomics Proteomics, 1, pp. 239–256.

Tagne JB, DB Reynolds, J Yoo, EG Jennings, J Zeitlinger, DK Pokholok, M Kellis, PA Rolfe, KT Takusagawa, ES Lander, DK Gifford, E Fraenkel, and RA Young. 2004. Transcriptional regulatory code of a eukaryotic genome. Nature, 431, pp. 99–104.

Talbot WS and N Hopkins. 2000. Zebrafish mutations and functional analysis of the vertebrate genome. Genes Dev, 14, pp. 755–762.

Till BJ, SH Reynolds, EA Greene, CA Codomo, LC Enns, JE Johnson, and S Henikoff. 2003. Large-scale discovery of induced point mutations with high-throughput TILLING. Genome Res, 13, pp. 524–530.

Tumpel S, F Cambronero, LM Wiedemann, and R Krumlauf. 2006. Evolution of *cis* elements in the differential expression of two *Hoxa2* coparalogous genes in pufferfish (*Takifugu rubripes*). Proc Natl Acad Sci USA, 103, pp. 5419–5424.

Van de Peer Y, JS Taylor, J Joseph, A Meyer. 2002. Wanda:a database of duplicated fish genes. Nucl. Acids Res., 30, pp. 109–112.

Wadman SA, KJ Clark, and PB Hackett. 2005. Fishing for answers with transposons. Mar Biotech, 7, pp. 135–141.

Washiet S, IL Hofacker, and PF Stadler. 2005. Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci USA, 102, pp. 2454–2459.

Weitzman JB. 2004. Co-regulation of mouse genes predicts function. J Biol, 3, p. 19.

Wienholds E, S Schulte-Merker, B Walderich, and RHA Plasterk. 2002. Target-selected inactivation of the zebrafish rag1 gene. Science, 297, pp. 99–102.

Wills AE. 1999. Translational control of growth factor and proto-oncogene expression. Int J Biochem Cell Biol, 31, pp. 73–86.

Woods IG, PD Kelly, F Chu, P Ngo-Hazelett, YL Yan, H Huang, JH Postlethwait, and WS Talbot. 2000. A comparative map of the zebrafish genome. Genome Res, 10, pp. 1903–1914.

Woolfe A, M Goodson, DK Goode, P Snell, GK McEwen, T Vavouri, SF Smith, P North, H Callaway, K Kelly, K Walter, I Abnizova, W Gilks, YJ Edwards, JE Cooke, and G Elga. 2005. Highly conserved non-coding sequences are associated with vertebrate development. PloS, 3, pp. 116–130, (e7).

Wu X, Y Li, B Crise, and SM Burgess. 2003. Transcription start regions in human genome are favored targets for MLV integration. Science, 300, pp. 1749–1751.

Yant SR, X Wu, Y Huang, B Garrison, SM Burgess, and MA Kay. 2005. High-resolution genome-wide mapping of transposon integration in mammals. Mol Cell Biol, 25, pp. 2085–2094.

Zambrowicz BP, GA Friedrich, EC Buxton, SL Lilleberg, C Person, and AT Sands. 1998. Disruption and sequence identification of 2,000 genes in mouse embryonic stem cells. Nature, 392, pp. 608–611.

Zbikowska HM. 2003. Fish can be first—advances in fish transgenesis for commercial applications. Transgenic Res, 12, pp. 379–389.

Zhang W, QD Morris, R Chang, O Shai, MA Bakowski, N Mitsakakis, N Mohammad, MD Robinson, R Zirngibl, E Somogyi, N Laurin, E Eftekharpour, E Sat, J Grigull, Q Pan, WT Peng, N Krogan, J Greenblatt, M Fehlings, D van der Kooy, J Aubin, BG Bruneau, J Rossant, BJ Blencowe, BJ Frey, and TR Hughes. 2004. The functional landscape of mouse gene expression. J Biol, 3, p. 21.

Zhao Z, Y Cao, M Li, and A Meng. 2001. Double-stranded RNA injection produces nonspecific defects in zebrafish. Dev Biol, 229, pp. 215–223.

Zhong W and PW Sternberg. 2006. Genome-wide prediction of *C. elegans* genetic interactions. Science, 311, pp. 1481–1484.

Zwaal RR, A Broeks, J van Meurs, JTM Groenen, and RHA Plasterk. 1993. Target-selected gene inactivation in *Caenorhabditis elegans* by using a frozen transposon insertion mutant bank. Proc Natl Acad Sci USA, 90, pp. 7431–7435.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Part 4
Preparing for Genome Sequencing

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 25
DNA Sequencing Technologies

Zhanjiang Liu

An understanding of the organization, expression and function, and evolutionary history of the aquaculture genomes requires knowing their primary structure—the linear order of the nucleotide base pairs of the genomes. Currently, demand for low-cost sequencing far outstrips what existing sequencing technology can provide. New sequencing technologies must be developed to dissect genomes of species with relatively small research communities.

DNA sequencing technology fundamentally reshaped the field of biology. It has been one of the largest driving forces in the genomics revolution. In the last 30 years, sequencing capacity has dramatically increased, while the cost of sequencing has been drastically reduced (Figure 25.1). However, the principles behind the various sequencing platforms have remained the same, relying on Sanger's dideoxy chain termination sequencing technology.

Back in the 1970s, two methods were independently developed for DNA sequencing. One method was developed by an American team, and the other, by an English team. The Americans, led by Maxam and Gilbert, used a "chemical cleavage protocol," while the English, led by Sanger, used an enzymatic protocol involving the use of nucleotide analogues for incorporation into DNA, which terminates the growing

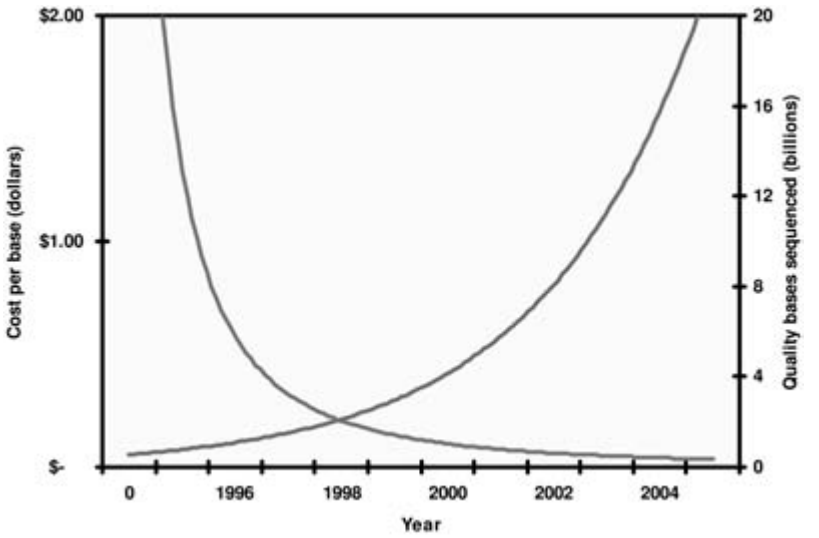


Figure 25.1. The trend of DNA sequencing capacity and cost. The figure is adopted from the Human Genome Project Information Web site (http://www.ornl.gov/sci/techresources/Human_Genome/research/instrumentation.shtml).

chain of DNA synthesis. Even though both teams shared the 1980 Nobel Prize, Sanger's method became the standard because of its practicality, while Maxam-Gilbert's chemical method has become obsolete and is no longer used for large-scale DNA sequencing.

In addition to these two major sequencing strategies, several emerging sequencing technologies are showing great promise. In particular, the 454 sequencing and the Solexa sequencing platforms, both based on the principles of sequencing by synthesis, are gaining momentum currently, and may meet increasing sequencing demand while reducing the cost. In this chapter, we will review the principles of sequencing technologies that brought us where we are today, and also study the potential of new sequencing technologies that may meet the demands of future genome sequencing.

Sanger's dideoxynucleotide chain termination method (the enzymatic method)

Sanger's method, also referred to as the dideoxynucleotide chain termination method, is based on the use of dideoxynucleotides (ddNTP) in addition to the normal nucleotides (dNTP) found in DNA. There are two key elements to Sanger's sequencing method: first, the discovery and application of nucleotide analogs; and second, a balanced proportion of dNTPs and their analogs ddNTPs in such a way that the chain termination occurs randomly at each base position creating a ladder representing each base being synthesized during DNA chain elongation.

dNTPs are building blocks of DNA. The ddNTPs are analogs of the dNTP that are essentially the same as nucleotides except they contain a hydrogen group (H) on the ribose 3′ carbon instead of a hydroxyl group (OH). Because they are the structural analogs of dNTPs, they can be used as substrates by DNA polymerases to be incorporated into DNA. ddNTPs, once incorporated into DNA, prevent the addition of further nucleotides because of their lack of a hydroxyl group at the 3′ of the DNA complex, leading to the termination of the DNA chain. If we suppose that only dNTPs are used for a synthesis reaction, the DNA chain would be synthesized continuously as long as the template and substrate exist. In contrast, if only ddNTPs are used, the DNA chain would be terminated at the first base when a dideoxynucleotide is incorporated. The key to Sanger's method is to use a proper proportion of dNTP and ddNTPs so that a ladder is created representing DNA chains terminated at every base position during DNA synthesis.

Historically, Sanger's method required the preparation of a single-stranded template from the DNA to be sequenced. This requirement was demanded by the DNA polymerase used for sequencing DNA. DNA polymerases all have a couple of important requirements for their polymerase activities: first, DNA polymerases require a template to copy from; second, they require the presence of a primer annealed to the template with the remaining portion of the template single-stranded so the DNA chain can be extended from the primer. When these conditions are present, DNA polymerase can synthesize DNA with a proper buffer system and in the presence of dNTPs.

The preparation of single-stranded templates was a great challenge back in the 1970s and the early 1980s. Use of the single-stranded phage M13 led to the innovation

of technology for the production of single-stranded templates. A group of scientists headed by Dr. Joachim Messing identified a region in the M13 that was required for the packaging of the M13 single-stranded DNA within the intergenic region (Messing et al. 1977). This discovery led to the development of a series of pUC cloning vectors (such as pUC118 and pUC119) aimed at production of single-stranded templates from cloned DNA. The most popular plasmid vectors used today all contain the intergenic region for the production of single-stranded DNA.

The discovery of thermostable DNA polymerases such as *Taq* DNA polymerase, and, thereafter, the invention of polymerase chain reaction (PCR) in 1986 completely eliminated the need for the production of single-stranded DNA. In a typical PCR reaction, the double-stranded DNA templates are denatured using heat. The denatured single-stranded DNA is annealed with the sequencing primer, and the primer is extended by the *Taq* polymerase. Based on the principles of PCR and sequencing, the cycle sequencing procedures were developed to sequence DNA using a double-stranded DNA template.

A typical cycle sequencing reaction contains several steps. Before the DNA can be sequenced, it has to be denatured into single strands using heat. Next, a primer is annealed to one of the template strands. This primer is specifically designed so that its 3′ end is located next to the DNA sequence of interest. Most often, the sequencing primer is designed based on the cloning vector at a region immediately outside of the poly-cloning sites. Typical sequencing primers are the T3 and T7 primers used with pBlueScript cloning vectors, or T7 and Sp6 sequencing primers used with pGEM cloning vectors. To detect the randomly terminated sequencing ladder, the sequencing primer or one of the nucleotides should be radioactively or fluorescently labeled. To illustrate the procedures of sequencing, let us use the classical sequencing reactions as an example. In a typical four-lane sequencing reaction, once the primer is annealed to the DNA, the solution is divided into four tubes labeled "A," "C," "G," "T" and then reagents are added to these samples as follows:

- "A" tube: all four dNTPs, ddATP, and DNA polymerase
- "C" tube: all four dNTPs, ddCTP, and DNA polymerase
- "G" tube: all four dNTPs, ddGTP, and DNA polymerase
- "T" tube: all four dNTPs, ddTTP, and DNA polymerase

As the DNA is synthesized, nucleotides are added on to the growing DNA chain by the polymerase. However, remember that the reaction is using a population of molecules in large numbers. On occasion, a proportional number of molecules incorporate a dideoxynucleotide into the chain in place of a normal nucleotide, resulting in a chain termination. Recall that we had four separate reactions for A, C, G, and T, respectively. In the reaction labeled "A," chain termination products should generate products terminating at positions with A (at the template positions containing T). Very similarly, in the reactions labeled "C," "G," and "T," newly synthesized DNA segments should be terminated at positions with C, G, and T, respectively. The sequencing ladder is resolved by denaturing polyacrylamide electrophoresis on a sequencing gel (Figure 25.2).

Although four-lane sequencing is still used by several platforms such as automated sequencers from LI-COR, the availability of fluorescent labels has allowed the development of single-lane sequencers, which increase sequencing efficiency at least four-fold. The sequencing platforms used by ABI automated sequencers such as ABI

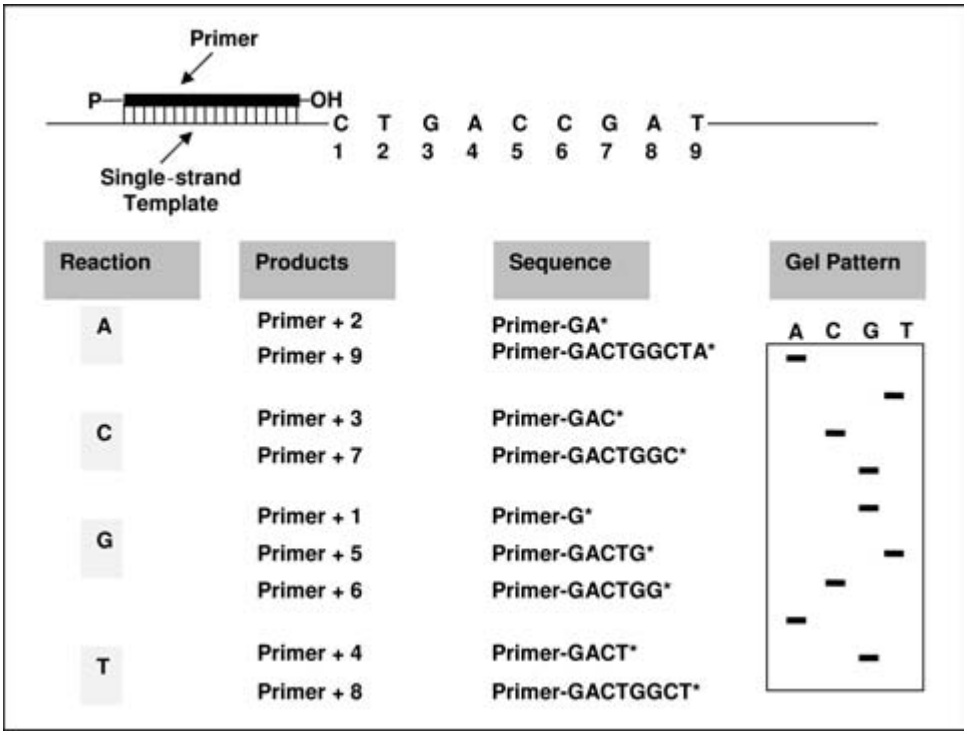


Figure 25.2. Principles of Sanger's dideoxynucleotide chain termination sequencing. Single-stranded template is annealed with a sequencing primer. After the dNTP and ddNTP mix is provided to the reaction containing DNA polymerase with proper buffer, synthesis of the new strand of DNA starts by primer extension. The incorporation of dideoxynucleotide (* in the figure) into DNA terminates the growing DNA chain, creating a sequencing ladder at each base position, depending on the base to be incorporated.

PRISM 3730 XL or 3130 XL, are all single-lane sequencing using BigDye labeling. The sequencing principles of single-lane sequencing are exactly the same as four-lane sequencing except that each dideoxynucleotide is labeled with a fluorescent dye with a different wavelength of excitation and emission so that they can be mixed in the same labeling reaction. The labeled fluorescent products are separated on a sequencing gel and subsequently detected by appropriate laser detectors.

Each sequencing reaction has a limited read length on sequencing gels, usually around 300–500 base pairs (bp). During the early days of sequencing, the ability to sequence long stretches of DNA depended on the development of nested overlapping clones allowing the entire DNA segment to be sequenced because the cost of oligonucleotide synthesis was very high at that time (as high as $30 per base). A number of methodologies for the production of nested clones were then developed (e.g., Anderson 1981, Henikoff 1984, Dale et al. 1985, Liu and Hackett 1993). Because these methodologies are no longer in use for sequencing today, these are only mentioned here as a part of the history of sequencing.

The primer walking approach is routinely used to completely sequence a long segment of DNA. Starting in 1994, primer synthesis costs reached a level of $1 per base from a historical high of $30 per base, and continued technological advancements

since then have allowed further reduction in the cost of primer synthesis. Lower primer costs meant that primer walking for sequencing a long segment of DNA could be cost effective. The primer walking method uses a previously generated sequence for the design of an additional primer for continued sequencing (Kim et al. 2000). The cycle of sequencing primer design followed by sequencing is continued until the desired segment of DNA is completely sequenced. Although cost effective, primer walking is slow when applied to very large pieces of DNA. In these cases, a shotgun sequencing approach is used, and complete sequences are then assembled by overlapping the generated sequences. This approach will be further discussed in the next chapter, Sequencing the Genome.

Maxam-Gilbert sequencing method

The Maxam-Gilbert sequencing method is no longer used for large-scale DNA sequencing, therefore, it is only briefly described here as a part of the historical perspective. This chemical cleavage method involves modification of the bases in DNA followed by base-specific cleavage using chemicals.

The Maxam-Gilbert sequencing method starts with double-stranded DNA. Most often, the DNA to be sequenced is radioactively labeled by attaching a radioactive phosphorus (^{32}P) group to the 5′ end, using a polynucleotide kinase enzyme and γ-^{32}P-ATP. Through exchange reactions of the polynucleotide kinase, the γ-^{32}P is transferred to the DNA molecule. The next step is to separate the strands of DNA. The two strands of the DNA are separated by treatment with dimethyl sulphoxide and heating to 90°C before being purified by electrophoresis. The third step is to conduct chemical modifications. The single-stranded sample is split into four separate samples, and each is treated with one of the cleavage reagents. This step involves alteration of bases. The four reactions follow: (1) G only, treated with dimethylsulphate (DMS); (2) A + G, treated with DMS plus formic acid; (3) C only, treated with hydrazine in 1.5M NaCl; and (4) C + T, treated with hydrazine. After base modifications, all modified bases are cleaved with piperdine. The cleavage products are electrophoresed on a polyacrylamide denaturing gel in four separate lanes with the smallest fragments run fastest. The sequence is read from the bottom (5′) to top (3′) of the gel.

When first developed, the Maxam-Gilbert sequencing method was thought to have the advantage of being able to sequence stretches of DNA that could not be sequenced by the enzymatic method due to high guanine-cytosine (GC) content or strong secondary structures. However, the toxic chemicals involved and the difficulties of adapting such a method for automation doomed this historically Nobel Prize winning technique to eventual obscurity.

Pyrosequencing and the 454 sequencing platform

Sequence determination is most commonly performed using the Sanger method. Decades of technological development and refining of the sequencing platforms made the dideoxy chain termination method the gold standard in the DNA sequencing business. All of the genomes sequenced to date including those of bacteria, archeal,

and eukaryotes (http://www.ncbi.nlm.nih.gov, and http://www.tigr.org) were sequenced using the chain termination method. However, this technique faces limitations in both throughput and cost for most future applications. Research needs have changed from sequencing a single human genome, as done in the human genome project, to sequencing thousands of genomes from groups of people with certain diseases to provide insight into genetic determinants of for example, cancer, heart disease, and high blood pressure. Many research groups around the world have put great effort into the development of alternative principles for DNA sequencing (Ronaghi 2001). Three methods that hold promise are sequencing by hybridization (Bains and Smith 1988, Drmanac et al. 1989, Khrapko et al. 1989), parallel signature sequencing based on ligation and cleavage (Brenner et al. 2000), and pyrosequencing (Ronaghi et al. 1998). Of these, pyrosequencing has emerged as the new sequencing methodology closest to being ready for widespread, practical applications. Pyrosequencing has the potential advantages of accuracy, flexibility, parallel processing, and can be easily automated. Furthermore, the technique dispenses with the need for labeled primers, labeled nucleotides, and gel electrophoresis. Because the newly developed 454 sequencing platform is based on pyrosequencing, here let us first introduce the principles of pyrosequencing.

Pyrosequencing is based on the detection of released pyrophosphate (PP_i) during DNA synthesis. Its principles, procedures, strengths, and weaknesses, as well as potential applications are discussed in a recent review by Ronaghi (2001). Pyrosequencing involves measurement of light generated through a cascade of enzymatic reactions after each nucleotide base incorporation. Technically, pyrosequencing can be divided into several steps for the sake of understanding. Step 1, a sequencing primer is hybridized to a single stranded, PCR amplified, DNA template, and incubated with the enzymes, DNA polymerase, adenosine 5′-triphosphate (ATP) sulfurylase, luciferase and apyrase, and the substrates, adenosine 5′-phosphosulfate (APS), and luciferin. Step 2, the synthesis of the first base by incorporating the needed nucleotide generates PP_i. Pyrophosphate is the natural product of DNA polymerization (Figure 25.3). Step 3, PP_i is converted by sulfurylase into ATP that reacts with luciferin to generate light in the presence of luciferase. The light is recorded by a charge-coupled device (CCD) camera and quantified as pyrograms. Step 4, the last step of the cascade after the recording of light, is to clear up the system, allowing the start of sequencing for the second base. This requires the degradation of all existing nucleotides and ATP by injecting apyrase. The sequential injection of one nucleotide at a time, coupled to the generation of light when the base is incorporated into the growing chain of DNA, allows the determination of sequences. The overall reaction from polymerization to light detection takes place within 3–4 seconds at room temperature. One pmol of DNA in a pyrosequencing reaction yields 6×10^{11} ATP molecules which, in turn, generate more than 6×10^9 photons at a wavelength of 560 nanometers (Ronaghi 2001). When mononucleotide repeats are encountered in the sequence, the pyrosequencing reaction continuously incorporates the repeated nucleotide until it reaches a different nucleotide. The light signal produced is proportional to the number of mononucleotides incorporated up to 8 bases. Mononucleotide repeats greater than 8 bp cannot be accurately sequenced by pyrosequencing (see below).

The theory behind pyrosequencing is sequencing-by-synthesis (Melamede 1985). It has not been used previously for sequencing because of the technical challenges of

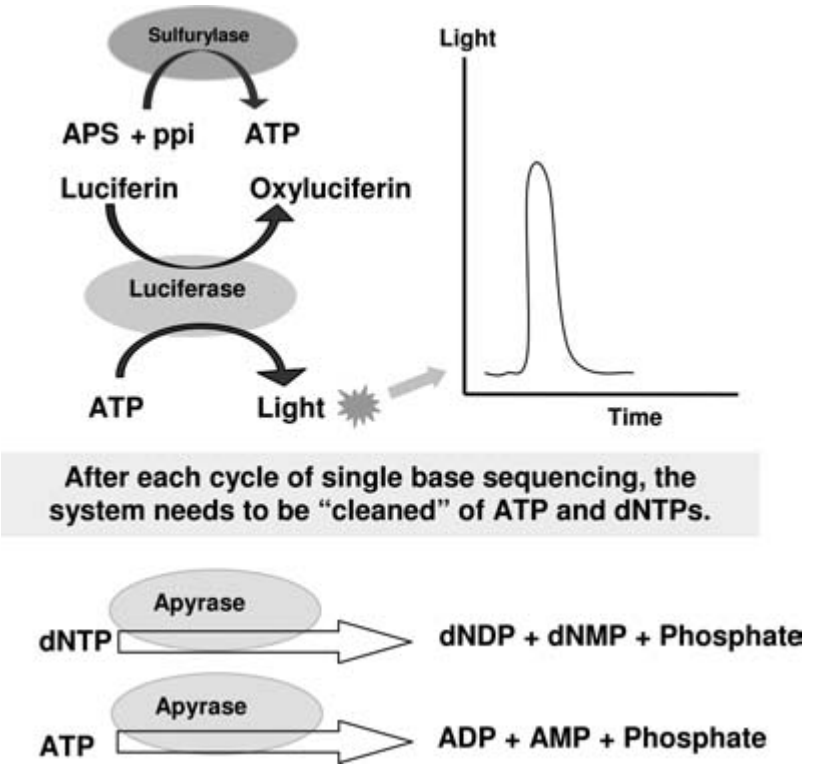


Figure 25.3. Schematic presentation of the principles of pyrosequencing.

synchronization of all the enzymatic reactions. Recent progress in luciferase chemistry and the availability of apyrase and sulfurylase have contributed to the technological advances toward the application of pyrosequencing as a sequencing strategy. However, pyrosequencing still faces great challenges in several areas (Ronaghi 2001). An inherent problem with the described method is de novo sequencing of polymorphic regions in heterozygous DNA material. In most cases, it will be possible to detect the polymorphism. If the polymorphism is a base substitution, it will be possible to obtain a synchronized extension after the substituted nucleotide. If the polymorphism is a deletion or insertion of the same nucleotide (e.g., AAAAAAA versus AA) as the adjacent nucleotide on the DNA template, the sequence after the polymorphism will be synchronized. However, if the polymorphism is a deletion or insertion of another type of nucleotide, the sequencing reaction can become out of phase, making the interpretation of the subsequent sequence difficult. If the polymorphism is known, it is always possible to use programmed nucleotide delivery to keep the extension of different alleles synchronized after the polymorphic region. Another problem is the difficulty in determining the number of incorporated nucleotides in homopolymeric regions, due to the nonlinear light response following incorporation of more than 5–8 identical nucleotides. Finally, but most importantly, pyrosequencing must overcome the problem of short read lengths. Currently, between 100–200 bp can be generated using pyrosequencing-derived sequencing technologies. For genome

sequencing and genome sequence assembly, long reads are definitely desirable. In spite of these inherent problems, pyrosequencing technology has recently been developed into a new sequencing platform called "454 sequencing." The platform has the potential to revolutionize genome sequencing, providing higher throughput and significantly lower costs than current standards.

The 454 sequencing platform

454 Life Sciences has developed a revolutionary technology, producing tens of millions of raw bases per hour on a single instrument. Many biologically meaningful and complex regions of genomes can be analyzed with this system without the time or cost constraints of current DNA sequencing methods. Through this technology, 454 Life Sciences provides an enabling solution for ultra-high-throughput DNA sequencing (http://www.454.com/). As only limited information is available at this time from a publication in *Nature* (Margulies et al. 2005), I will briefly describe the 454 sequencing principles and processes here based on the available information.

The 454 sequencing platform is based on the principles of pyrosequencing and uses microfabricated high-density picolitre reactors (Margulies et al. 2005, Figure 25.4). No cloning is necessary for 454 sequencing, and thus certain terms such as DNA library, etc., have specific meanings here for the description of the 454 sequencing system. The clonal DNA used for sequencing is obtained by clonal PCR amplification of a single molecule in emulsified water-in-oil microreactors. Preparation of the DNA library consists of a few simple steps. Genomic DNA is fractionated into smaller fragments (300–500 bps) that are subsequently filled in to polished ends (blunted), allowing ligation of adaptors to the genomic DNA for PCR amplification. To prevent intermolecular ligation of the genomic DNA fragments, the DNA fragments are dephosphorylated. Short adaptors (A and B) are then ligated onto the ends of the fragments. After ligation, the gap needs to be repaired, presumably using a DNA ligase. The adaptors provide priming sequences for both amplification and sequencing of the sample-library fragments. The two adaptors are different. Adaptor B contains a 5′-biotin tag that enables immobilization of one strand of the library onto streptavidin-coated beads. The nonbiotinylated strand is released and used as a single-stranded template DNA library.

The single-stranded template DNA library is immobilized onto beads carrying short primers complementary to the Adaptor A sequences by base pairing. The key element here is attaining the correct proportion of beads to DNA molecules so that only one molecule is captured by each bead. The beads containing a single molecule of the single-stranded template are emulsified with the amplification reagents in a water-in-oil mixture. Each bead is captured within its own microreactor where PCR amplification occurs. This results in bead-immobilized, clonally amplified DNA fragments.

The single-strand template DNA library beads are added to the DNA Bead Incubation Mix (containing DNA polymerase) and are layered with enzyme beads (containing sulfurylase and luciferase) onto the PicoTiterPlate device. The device is centrifuged to deposit the beads into the wells. The layer of enzyme beads ensures that the DNA beads remain positioned in the wells during the sequencing reaction. Due to the size of the wells in relation to the beads, only one bead containing a

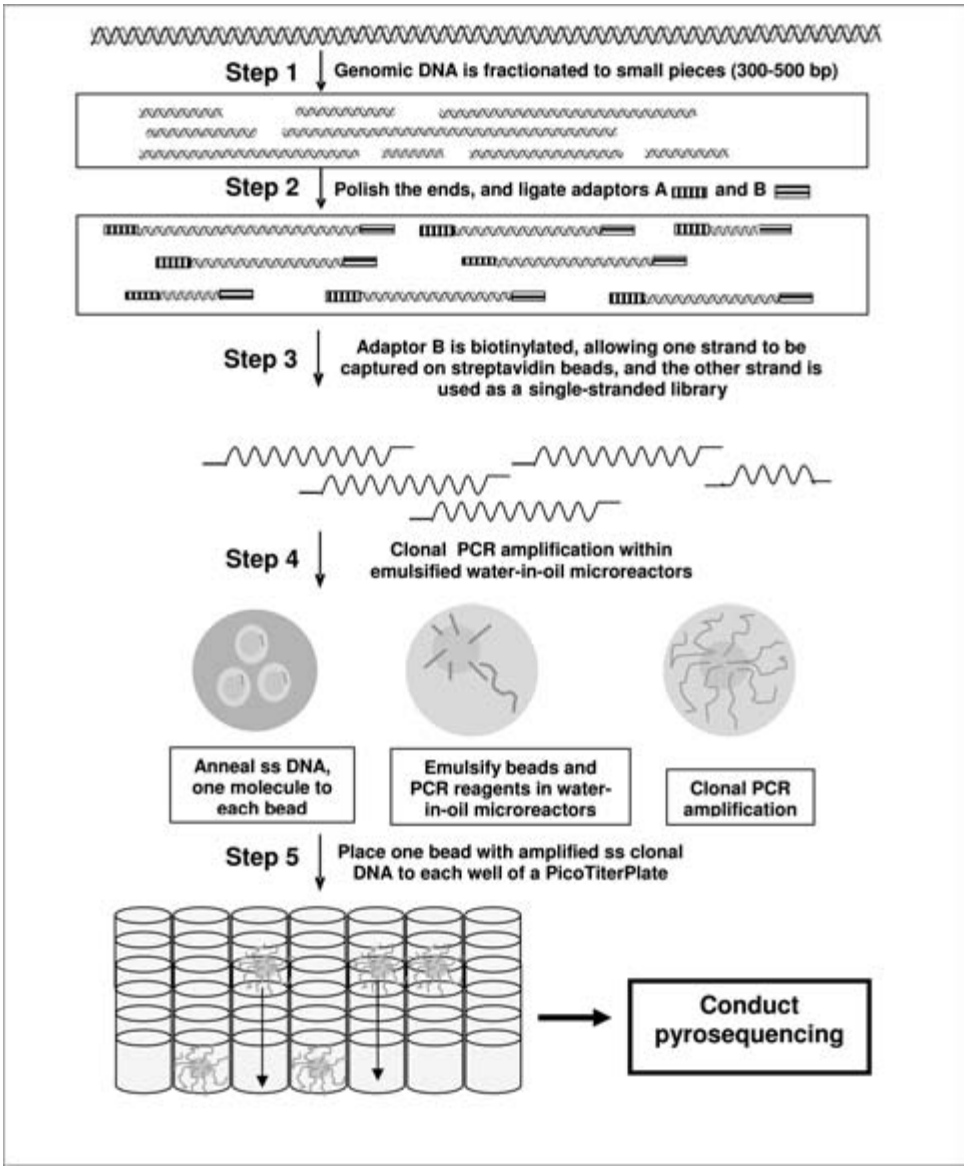


Figure 25.4. Schematic presentation of the principles of the 454 sequencing platform. (Also see color plate.)

specific clonally amplified genomic DNA segment should be placed into each well of the PicoTiterPlate device. The loaded PicoTiterPlate device is placed into the "454 sequencer," the Genome Sequencer 20 Instrument, which performs "pyrosequencing-like" reactions. See above under Pyrosequencing. Unlike a traditional pyrosequencing reaction, hundreds of thousands of beads, each with millions of copies of clonally amplified DNA, are sequenced in parallel. Each well of the PicoTiterPlate device is a separate pyrosequencing reaction. If a nucleotide complementary to the template strand is flowed into a well, the polymerase extends the existing DNA strand

by adding nucleotide(s). Addition of one (or more) nucleotide(s) results in a reaction that generates a light signal that is recorded by the CCD camera in the instrument. The signal strength is proportional to the number of nucleotides incorporated in a single nucleotide flow. Typically, 200,000 reactions are conducted in parallel in a single run. Assuming generation of 100 bp by a single reaction, each run should generate 20 million bp or more of sequence in several hours using a single instrument. Imagine, this is approximately a 10× coverage of a bacterial genome!

From the above introduction, it is clear that the 454 sequencing platform holds much potential. The current major problems prohibiting its application for the sequencing of complex genomes are its relatively short sequencing reads, and difficulties in accurate determination of homopolymeric run in the DNA. The short reads complicate genome sequence assembly, while the inability to determine the number of bases within a long homopolymeric run prohibits accurate sequencing of genomes. These problems are more significant for complex genomes with high levels of repeat structure. However, the technologies' high throughput and low costs are very attractive, especially for aquaculture species. As the technology is perfected to minimize these drawbacks, the 454 sequencing platform will show even greater promise.

The Solexa sequencing platform

Solexa's core technology, the Clonal Single Molecule Array™ technology, allows simultaneous analysis of hundreds of millions of individual molecules. With its unparalleled data density, Solexa's platform will dramatically improve the speed and reduce the cost of a range of genetic analysis applications, including whole-genome resequencing and expression profiling. Although Solexa is developing and plans on commercializing a new platform based on Sequencing-By-Synthesis (SBS), the company currently offers a range of expression profiling and small RNA analysis based on its Massively Parallel Signature Sequencing (MPSS) technology. Both technologies (MPSS and SBS) leverage massively parallel sequencing of short DNA and complementary DNA (cDNA) fragments to generate data from millions of fragments simultaneously. However, by leveraging its Clonal Single Molecule Array™ technology and reversible terminator chemistry, Solexa's SBS approach is anticipated to generate up to 1 billion bases of data per run at costs far more economical than that of MPSS. Given the data density and economical advantages of SBS over MPSS, Solexa plans on transitioning its MPSS-based service offerings to SBS. As limited information is available in published sources, here I introduce some basic concepts of the Solexa sequencing technology using Web-based information (http://www.solexa.com/).

Solexa's SBS uses four proprietary fluorescently labeled modified nucleotides. These specially created nucleotides, which also possess a reversible termination property, allow each cycle of the sequencing reaction to occur simultaneously in the presence of all four nucleotides (A, C, T, G). In the presence of all four nucleotides, the polymerase is able to select the correct base to incorporate, with the natural competition between all four alternatives leading to higher accuracy than methods where only one nucleotide is present in the reaction mix at a time, which require the enzyme to reject an incorrect nucleotide. Sequences where a particular base is repeated one after another ("homopolymer repeats") are dealt with as for any other

sequence and with high accuracy; this avoids the problems of measuring intensity and deducing how many bases were present in the repeat that are the cause of uncertainty seen with "one base per reaction" methods, as described above with the 454 sequencing platform.

Conclusion

Sequencing technology has been applied in research for 30 years. Although multiple sequencing strategies have been invented during this time, Sanger's sequencing strategy has been used for sequencing all the genomes to date. Decades of optimization, refining, and streamlined operations have allowed Sanger's method to be applied on an industrial scale at a reasonable cost for worldwide mutual efforts such as the Human Genome Project. However, the genome projects of most eukaryotic and prokaryotic species have no worldwide interest. In addition, interest in the biomedical community has shifted from obtaining a single human genome sequence to comparing thousands of human genomes in the context of susceptibility to cancer, heart disease, diabetes, high blood pressure, and so on. The current sequencing technologies cannot meet these tremendous sequencing demands with current levels of research funding. Even though the cost of sequencing has been reduced by at least 100 times in the last decade, sequencing a genome of a comparable size to the mammalian genomes still takes $10 million or more using current sequencing technology. This price tag is too expensive for individual human genome sequencing and far too expensive for the underfunded aquaculture community. Current sequencing costs need to be reduced another 100-fold before genome sequencing can become routine. Emerging sequencing technologies such as 454 sequencing and Solexa sequencing have the potential to meet these demands. Aquaculture genome sequencing requires high efficiency and low cost, because we have a small research community. Hopefully the emerging sequencing technologies will become mature soon to provide low-cost and high-quality sequencing. With the current state of technology, 454 sequencing and Solexa sequencing would allow the entire genomes of many fish species to be sequenced for less than $1 million. Although genome sequence assembly may be difficult with current read lengths, it may be possible to anchor "thousands of sequence islands" created by 454 sequencing or Solexa sequencing onto well-developed physical and genetic maps, which may prove to be sufficient for agricultural research objectives.

References

Anderson S. 1981. Shotgun DNA sequencing using cloned DNase I-generated fragments. Nucl Acids Res, 9, pp. 3015–3027.

Bains W and GC Smith. 1988. A novel method for nucleic acid sequence determination. J Theoret Biol, 135, pp. 303–307.

Brenner S, SR Williams, EH Vermaas, T Storck, K Moon, C McCollum, JI Mao, S Luo, JJ Kirchner, and S Eletr. 2000. In vitro cloning of complex mixtures of DNA on microbeads: Physical separation of differentially expressed cDNAs. Proc Natl Acad Sci, 97, pp. 1665–1670.

Dale RM, BA McClure, and JP Houchins. 1985. A rapid single-stranded cloning strategy for producing a sequential series of overlapping clones for use in DNA sequencing: application to sequencing the corn mitochondrial 18 S rDNA. Plasmid, 13, pp. 31–40.

Drmanac R, I Labat, I Brukner, and R Crkvenjakov. 1989. Sequencing of megabase plus DNA by hybridization: Theory of the method. Genomics, 4, pp. 114–128.

Henikoff S. 1984. Unidirectional digestion with exonuclease III creates targeted breakpoints for DNA sequencing. Gene, 28, pp. 351–359.

Khrapko KR, YP Lysov, AA Khorlyn, VV Shick, VL Florentiev, and AD Mirzabekov. 1989. An oligonucleotide hybridization approach to DNA sequencing. FEBS Lett, 256, pp. 118–122.

Kim S, A Karsi, RA Dunham, and Z Liu. 2000. The skeletal muscle alpha-actin gene of channel catfish (*Ictalurus punctatus*) and its association with piscine specific SINE elements. Gene, 252, pp. 173–181.

Liu Z and PB Hackett. 1993. Reverse cloning procedure for generation of subclones for DNA sequencing. Methods Enzymol, 218, pp. 47–58.

Margulies M, M Egholm, WE Altman, S Attiya, JS Bader, LA Bemben, J Berka, MS Braverman, YJ Chen, Z Chen, SB Dewell, L Du, JM Fierro, XV Gomes, BC Godwin, W He, S Helgesen, CH Ho, GP Irzyk, SC Jando, ML Alenquer, TP Jarvie, KB Jirage, JB Kim, JR Knight, JR Lanza, JH Leamon, SM Lefkowitz, M Lei, J Li, KL Lohman, H Lu, VB Makhijani, KE McDade, MP McKenna, EW Myers, E Nickerson, JR Nobile, R Plant, BP Puc, MT Ronan, GT Roth, GJ Sarkis, JF Simons, JW Simpson, M Srinivasan, KR Tartaro, A Tomasz, KA Vogt, GA Volkmer, SH Wang, Y Wang, MP Weiner, P Yu, RF Begley, and JM Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. Nature, 437, pp. 326–327.

Maxam A and W Gilbert. 1977. A new method of sequencing DNA. Proc Natl Acad Sci USA, 74, pp. 560–564.

Melamede RJ. 1985. Automatable process for sequencing nucleotide. US Patent No. US4863849.

Messing J, B Gronenborn, B Muller-Hill, and PH Hofschneider. 1977. Filamentous coliphage M13 as a cloning vehicle: insertion of a *Hin*d II Fragment of the *lac* regulatory region in M13 replicative form in vitro. Proc Natl Acad Sci USA PNAS, 74, pp. 3642–3646.

Ronaghi M, M Uhlæn, and P Nyræn. 1998. A sequencing method based on real-time pyrophosphate. Science, 281, pp. 363–365.

Ronaghi M. 2001. Pyrosequencing Sheds Light on DNA Sequencing. Genome Research, 11, pp. 3–11.

Sanger F, S Nicklen, and AR Coulson. 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA, 74, pp. 5463–5467.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 26
Sequencing the Genome

Zhanjiang Liu

Genome sequencing literally means completely sequencing the entire genome so that all genome bases and their locations are known. It is the process of determining the exact order of the nucleotide bases making up the complete set of chromosomes of the genome. Because DNA sequencing is routine, genome sequencing appears intuitively very straightforward, as if 1,000 base pairs (bp) can be completely sequenced in a single run in a single lane, a genome of 1 billion bps would take 1 million lanes, and because the current automated sequencers can load 96 samples per run, to sequence the entire genome would appear to require only 10,416 runs. Although these intuitive numbers are correct for a 1× genome coverage, completely sequencing a genome is more complex than it appears to be. First, we know that a sequencing run can determine sequences from a short segment, usually approximately 1,000 bp, with shorter quality sequences. As we discussed in the sequencing technology chapter, sequencing of longer segments of DNA requires generation of nested overlapping clones, or primer walking. It is not practical to generate nested clones covering the entire genome, nor is it practical to sequence the entire genome using primer walking. Many genomic regions would be highly repetitive prohibiting direct sequencing. Second, genomic DNA cannot be handled in its entire length with its *in situ* state of chromosomes because sequencing requires a large number of molecules synchronized at the primer binding site. Genomic DNA is typically broken into tens of thousands of random segments of approximately 100–200 kilobase (kb) during DNA isolation from cells due to shearing. Third, from the prospect of project management, sequencing a whole genome also poses special challenges that demand special approaches and strategies.

Strategies for Whole Genome Sequencing

Three strategies have been used for sequencing genomes with large sizes such as those of the mammals: the whole genome shotgun (WGS), the hierarchical shotgun (the bacterial artificial chromosome [BAC] clone-by-clone approach), and the hybrid approach of the two. These strategies, however, were all developed based on the current sequencing technology using the Sanger's dideoxy chain termination method. The adoption of emerging sequencing technologies could lead to the development of new strategies for genome sequencing. We will first introduce the traditional strategies for genome sequencing, and then follow up with strategies on the horizon.

Whole Genome Shotgun (WGS)

Initially, two general strategies for sequencing a complete genome were used: the shotgun sequencing and the clone-by-clone sequencing. WGS sequencing shears

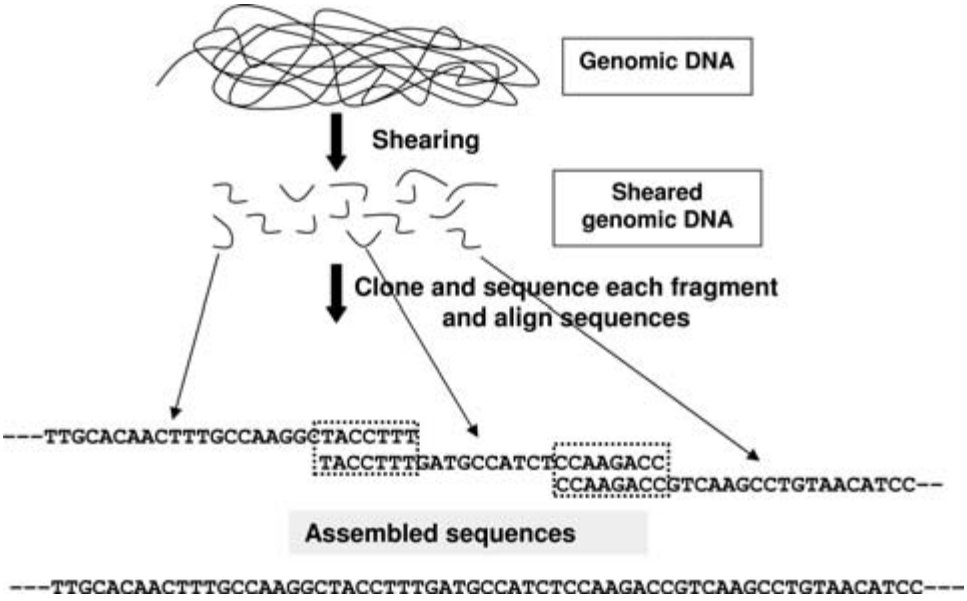


Figure 26.1. Schematic presentation of the whole genome shotgun sequencing strategy. Genomic DNA is sheared into small segments and sequenced. After sequencing, sequences are assembled into continuous sequence elements (contigs) based on overlapping. Note that usually a 6–10× genome coverage is required for the assembly of the genome, depending on the level of repetitive elements of the genome. This strategy was used by Celera for sequencing of the human genome.

genomic DNA into randomly small pieces that are cloned into plasmids and sequenced on both strands. When the sequences are obtained, they are aligned and assembled into draft sequences using bioinformatic tools (Figure 26.1).

The WGS sequencing is generally conducted by construction of shotgun sequencing libraries with different insert sizes. Most typically, the WGS is facilitated with BAC libraries, cosmid libraries, and plasmid libraries. BAC libraries typically have large inserts of 100–200 kb. Cosmid libraries hold intermediate insert sizes of approximately 40 kb. Plasmid libraries used for WGS vary between 2 and 8 kb, but most often are around 4 kb. Sequences generated from the large insert BAC libraries serve as large anchoring islands for sequence assembly and clone orientation; sequences generated from the intermediate insert cosmid libraries serve as smaller anchoring islands for sequence assembly; and the small insert plasmid libraries provide high genome coverage allowing complete sequencing of the genome. The use of large and intermediate insert libraries is very useful for genome assembly, especially for regions with highly repetitive sequences. That is because the BAC clones used for the generation of sequences are also located on the physical map constructed by BAC-based contig construction using restriction fingerprinting. The physical map information would serve as anchoring points of where the sequences are physically located, avoiding mistaken assembly based just on sequence overlaps within the repetitive region.

The key element of whole genome shotgun sequencing is to provide more random libraries with increased coverage, representing the entire genome. In order to be sure of entire genome representation, various tricks can be applied including the above-mentioned libraries with various sizes, the use of various fragmentation methodologies such as shearing, partial restriction digestion, and cloning into high and/or low copy number vectors.

The shotgun method is faster and less expensive, as compared with the hierarchical shotgun sequencing. However, it poses greater challenges for genome sequence assembly since there is no guiding "island" of sequences as there is with the clone-by-clone approach. It is more prone to errors due to erroneous assembly of the entire genome sequence. This is particularly true in genomic regions where repetitive elements are rich, often leading to the over-condensation of the genome sequence with repeats underrepresented due to mistaken assembly. WGS strategy is therefore more adaptable to sequencing of the relatively small genomes, and those with simple repeat structures. However, due to its speed and relatively lower costs, it allows a greater level of genome coverage with the same financial resources.

Hierarchical Shotgun Sequencing (the Clone-by-Clone Approach)

The hierarchical shotgun sequencing strategy is also referred to as the clone-by-clone strategy of genome sequencing (Figure 26.2). In this approach, genomic DNA is first cleaved into segments of about 100–200 kb and inserted into BAC vectors. The BAC clones containing a high level of genome coverage of the genomic DNA are collectively called the BAC library. See Chapter 13 for its construction. BAC libraries are the basis of physical mapping, as described in Chapter 14, by construction of contigs using restriction fingerprinting. Once the contigs are constructed, the use of fingerprints, and most often also BAC-end sequences (see Chapter 15), allows the identification of a set of BAC clones containing the entire genome, but with minimal overlapping. Such a set of BAC clones is called the minimal tiling path (MTP), which is defined as a minimally overlapping set of all clones in the physical map. Each BAC DNA in the MTP is fragmented randomly into smaller pieces and each piece is cloned into a plasmid and sequenced on both strands. These sequences are aligned so that identical sequences are overlapping. These contiguous pieces are then assembled into the finished sequence of the BAC, and the physical map information and BAC-end sequences are then used to assemble the entire genome sequence. The assembly of the genome sequence usually requires 5–10× genome coverage.

The advantage to the hierarchical shotgun approach is its lower level of mistakes when assembling the shotgun sequences into contigs for the generation of genome sequence. The reason is that the chromosomal location for each BAC is known from the physical map, and there are also fewer random pieces to assemble within each BAC. Also, the complexity of repetitive elements within a single BAC should be much lower than that for the whole genome. The weakness of this method is its low speed of sequence generation and its high costs. Once produced, the genome sequences generated using the hierarchical shotgun approach are regarded highly in their qualities.

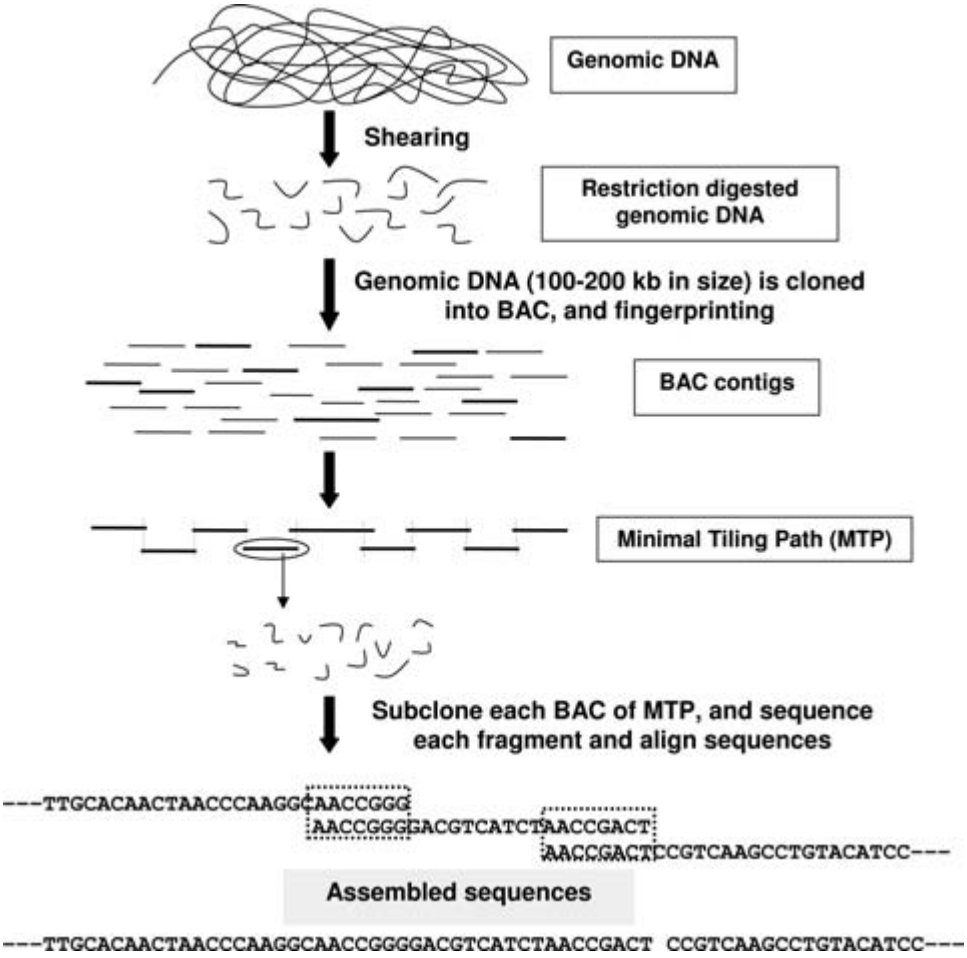


Figure 26.2. Schematic presentation of the hierarchical shotgun sequencing strategy. Genomic DNA is partially digested with restriction enzyme to produce segments of 100–200 kb that are cloned into the BAC vector for contig construction using fingerprinting. Based on the fingerprints (and often also the BAC-end sequences), a minimal tiling path (MTP) can be identified for clone-by-clone sequencing of the genome. Each selected BAC in MTP is further subcloned and shotgun sequenced for the assembly of the sequence of the BAC. The whole genome sequence is assembled from the complete BAC sequences of the MTP BAC clones. This strategy was used by the publicly funded Human Genome Project.

The Hybrid Approach

Because the whole genome shotgun and the hierarchical shotgun approaches are complementary to each other in many aspects, using one approach would alleviate the problems of the other. This gives great motivation for the use of both approaches for the production of a whole genome sequence. The relative proportion of the two approaches used in a genome project depends on the genome size, repeat structure of

the genome, the physical map quality, the availability of BAC-end sequences, and the availability of the MTP set of BAC clones, as well as financial considerations.

Selection of Minimal Tiling Path Clones

Selecting an MTP is the task of picking a set of minimally overlapping clones that span an entire contig. Many of the approaches used for picking MTP clones were developed by the Arizona Genomics Computational Laboratory (http://www.agcol.arizona.edu/), along with the fingerprinting and construction of BAC contigs software package FingerPrint Contig (FPC) (Soderlund et al. 1997, 2000). Their Web page provides a rich source for physical mapping and MTP selection information. Readers interested in the details are referred to their published papers, as well as to their Web site. I will briefly introduce the principles for the selection of MTP BAC clones.

Because of inexact coordinates in the Consensus Bands (CB) map, MTP clones cannot be selected based solely on their positions on the map. Three methods are currently used for picking MTP clones:

1. The fingerprint method, in which overlaps are determined by examination of the clone fingerprints and their map position.This method has been used for the selection of MTP clones for the *Caenorhabditis elegans* project (Coulson et al. 1986), and it was demonstrated to be effective. However, the overlapping regions among the BAC clones were large (e.g., approximately 47.5 kb overlap for the minimal tiling path picked by the International Human Genome Consortium using this approach). This is because the fingerprints were produced using 6-bp cutters of restriction enzyme that cut DNA at a rate of once every 4,096 bp. It requires a large number of overlapping fragments (usually 10–12) to be statistically significant. Recently, four-colored fluorescent fingerprinting has been adopted for physical mapping (Luo et al. 2003), which should greatly reduce the overlapping sizes in the MTP clones while maintaining statistical significance. This is because in this fingerprinting approach, four sets of fingerprints are used for contig construction.
2. The BAC-end sequence method. This approach was initially proposed in 1996 (Venter et al. 1996) using BAC-end sequences as tagged sequence connectors. In this method, BAC-end sequences are first produced (see Chapter 15) as a genome resource for the genome project. When one BAC clone is completely sequenced, its entire sequence is aligned to all BAC-end sequences to identify the clones with minimal sequence overlaps. This approach was found to be very effective in reducing the overlapping regions of MTP clones. However, false positive clones can be identified within repetitive element regions. This is because BAC-end sequences are often short (usually 400–600 bp). When coupled to the relatively low quality of BAC-end sequences in many species, the likelihood of mistaken identification of overlapping sequences is increased, especially within repetitive genomic environs.
3. The hybrid approach of the two methods. It is clear from the above discussion that the two approaches of MTP selection use different resources. The first method involves analyzing the fingerprints of a pair of overlapping clones for shared restriction fragments (bands), and verifying the integrity of the fingerprints of the potentially overlapping pair by matching bands with a spanning and two flanking clones.

In the second method, draft sequence of a BAC is used to determine which additional BAC clones have the minimal overlapping with the completely sequenced BAC clone. The first method uses information that is already present in the physical map, but the overlaps can be inexact. The second method requires sequence information, but gives very exact overlaps (http://www.agcol.arizona.edu/). When both the fingerprints and the BES resources are used in the hybrid approach, the precision should be increased while the overlapping regions are reduced, leading to a reduced sequencing load for genome projects. Software such as MTP is available from Arizona Genomics Computational Laboratory using both sets of information for efficient selection of MTP (http://www.agcol.arizona.edu/software/fpc/userGuide/mtpdemo/#intro).

Emerging Sequence Technologies and Platforms

The above discussion is based on current sequence technologies. However, several promising sequencing strategies are poised to make technical advances to allow them to be effectively used for genome sequencing. A couple of these technologies, such as the 454 sequencing and the Solexa sequencing platforms, as discussed in the previous chapter on sequencing technologies, strive to demonstrate its high efficiency, high throughput, and low cost. Their application could fundamentally change how genome sequencing is conducted, not only at the level of technicalities, but also at the level of psychology and decision making. Their application for genome sequencing could have a revolutionary impact on life sciences, and bring a real opportunity to sequencing aquaculture genomes.

Sequencing the Genome

After the strategy is determined, sequencing of a genome itself is straightforward. The important aspect of large-scale sequencing is the streamlined operation with accurate tracking and recording, as well as the bioinformatics pipeline. In almost all cases, genome sequencing projects have been conducted at large genome centers or by biotechnology industries, both of which not only have the state-of-the-art sequencing equipment but also expertise gained in genome sequencing for management of large sample sets and data sets. Importantly, they are well equipped with bioinformatics capabilities.

Genome Sequence Assembly

Current sequencing technologies only allow an average read length of 1,000–1,500 bp per run, with the first approximately 500–800 bp as high-quality sequences. As discussed above, to overcome this limitation, sequencing of entire genomes is performed through either whole genome shotgun sequencing or the clone-by-clone hierarchical shotgun sequencing. In either case, DNA is sheared into smaller fragments whose

ends are then sequenced. The generated sequences need to be assembled using computer programs called assemblers. The output of assembly programs consists in a collection of contiguous sequence pieces (contigs). They are rarely, if ever, entire chromosomes reconstructed into a single contig, but many smaller pieces. Additional computer programs called the scaffolder use the information linking together sequencing reads from the ends of fragments to order and orient the contigs with respect to each other along a chromosome.

Sequence assembly is essentially a set of contigs, each contig being a multiple alignment of reads (Dear et al. 1998). The assembler relies on the basic assumption that two sequence reads that share the same string of letters originated from the same place in the genome. Mathematical modeling indicated that an 8–10 genome coverage should allow the vast majority of sequences to be assembled into large contigs (more than 200,000 bp) (Lander and Waterman 1988).

Ideally, genome sequence assembly should produce one contig for every chromosome of the genome being sequenced. In reality, however, many contigs are produced as a result of a combination of factors. Nonrandom shearing and the presence of repeats are the two greatest challenges. Even at eightfold to tenfold coverage, some portions of the genome remain unsequenced as gaps.

A number of different strategies have been proposed to deal with genome sequence assembly (Peltola et al. 1984, Huang 1996, Parsons et al. 1993, Bonfield et al. 1995, Notredame and Higgins 1996, Zhang and Wong 1997, Ewing and Green 1998, Ewing et al. 1998). Currently, there are mainly two different existing approaches for assembling sequences: (1) the iterative and (2) the All-in-One-Step approach. The first type of assembly is essentially derived from the fact that the data analysis and reconstruction approximation algorithms can be parametrized differently, ranging from very strict assembly of only the highest quality parts to very 'bold' assembly of even lowest quality stretches. An assembly starts with the most strict parameters, having the output edited manually by highly trained personnel, or by software and then the process is reiterated with less strict parameters until the assembly is finished (http://www.bioinfo.de/isb/gcb99/talks/chevreux/main.html). The second approach has been made popular by the PHRAP assembler presented by Phil Green (http://www.phrap.org/). This assembler uses low and high-quality sequence data from the start and generates a consensus by puzzling its way through an assembly using the highest quality parts as reference, giving the result to a human editor for finishing. An integrated approach of the two approaches has been developed by Chevreux and others (http://www.bioinfo.de/isb/gcb99/talks/chevreux/main.html).

The fundamentals of genome sequence assembly are the same as assembly of short DNA sequence of a few kb. The quality of the sequences in base calling is the most important (Figure 26.3). However, the genome sequence assembly is much more complex because of the large sizes and similar sequences exist here and there in the genome by chance. In highly repetitive genomic regions, sequence assembly can be a disaster. For accurate genome assembly, the clone-by-clone hierarchical shotgun sequencing is superior to the whole genome shotgun sequencing (Figure 26.4). If nothing else, two aspects of the clone-by-clone approach make it easier for sequence assembly. First, the region under consideration is a BAC clone, most often less than 200 kb in size, as compared to the whole genome of billions of bps. Second, the repetitive regions within a single BAC can be much simpler than the situation of facing a whole genome. The coincidence of similar sequences by chance is much smaller with a

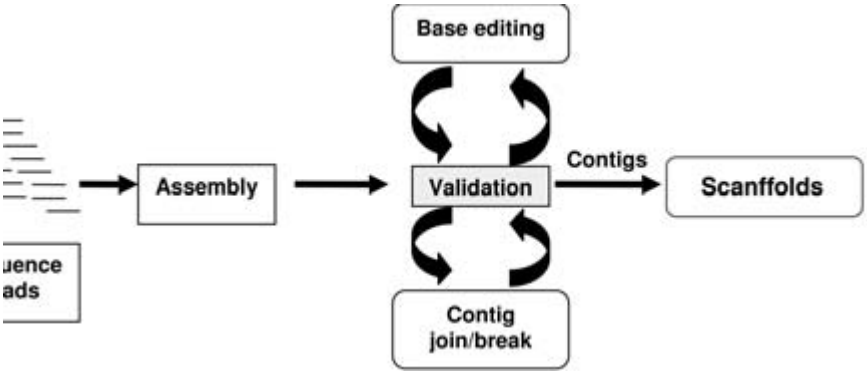


Figure 26.3. Schematic flow chart for sequence assembly processes.

single BAC than with a whole genome. In addition, gap filling is much easier for the clone-by-clone approach with known regions for higher levels of genome coverage in additional sequencing. In contrast, just increasing sequencing genome coverage may or may not easily fill the gaps.

Draft Genome Sequence and Finishing

Sequencing the bases equal to the genome size is defined as one genome coverage. Obviously, the higher the genome coverage, the greater the possibility for assembly of

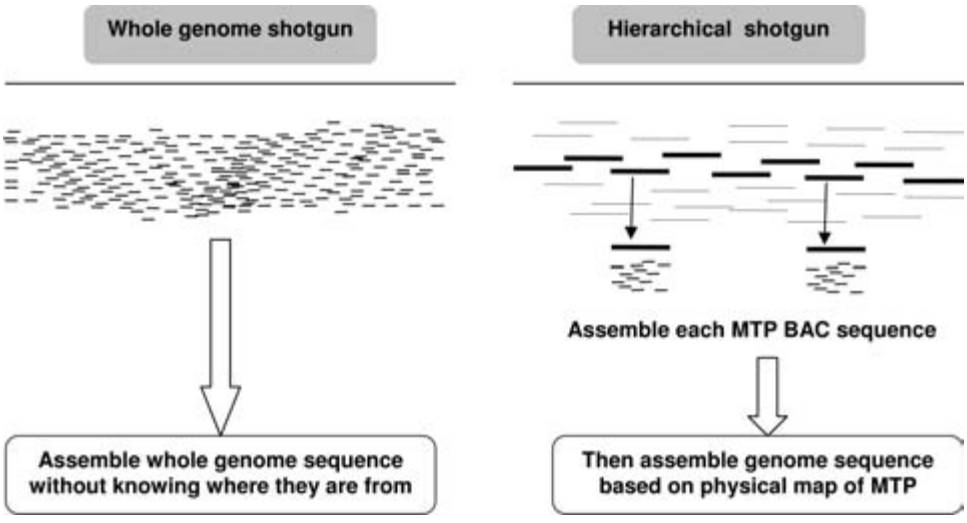


Figure 26.4. The comparison of the whole genome shotgun and the hierarchical shotgun sequencing approaches in sequence assembly. Short lines are individual sequence reads. Long lines are BAC clone inserts. Thick long lines are selected BAC clones with minimal overlapping (minimal tiling path) for sequencing.

the complete genome sequence. However, the greater the coverage is, the higher the costs. This dilemma must be balanced to provide a reasonable accurate assembly, but keep the costs as low as possible. Usually, it takes 4–5× genome coverage to assemble genome sequence into reasonably large segments (usually larger than 10,000 bp). Such an assembly is referred to as the draft genome sequence, or in other words, the sequence assembly is preliminary. Using a genome sequencing coverage of 4–5×, the human draft sequence was released in June 2000. To generate the high-quality human reference sequence, additional sequencing was needed to close gaps, reduce ambiguities that allow for only a single error every 10,000 bases. A high-quality sequence is critical for recognizing regulatory components of genes that are very important in understanding human biology and such disorders as heart disease, cancer, and diabetes. After sequencing of 8–9 genome coverage, the human genome sequence was finished in 2003.

The ultimate goal of any sequencing project is to determine every single bp of the original set of chromosomes. However, it is rare for an assembly program to be able to reconstruct a single piece of DNA per chromosome, leading to gaps in the reconstruction of the genome. These gaps are filled in through directed sequencing in a process called finishing or gap closure. Gap filling allows merge of sequence contigs into supercontigs and scaffolds (Figure 26.3). Finishing is the process whereby additional sequences are obtained to achieve full coverage and continuity over very long distances by filling in the gaps. Only the Human Genome Project is undertaking finishing. Gap filling sounds easier than it is, but requires great effort and effective strategies.

Gap Filling Using Expressed Sequence Tag (EST) Resources

In most cases, a large EST database exists for the genome being sequenced. After initial sequencing for the generation of draft sequence, many smaller segments of sequence reads are left unassembled because there are no overlapping sequences, or in other words, there are gaps. One of the strategies for gap filling is to use the leftover unassembled sequences to search the EST database of the species, allowing identification of exon segments. The exon sequences can be used to design the primer for direct BAC sequencing.

Gap Filling by Creating Libraries with Variable Sizes

In many cases, the existence of gaps could be attributed to the methods used for the construction of genome sequencing libraries, either because of the restriction enzymes used, or because of the selection of certain insert sizes of the libraries. By changing the restriction enzymes used for the construction of sequencing plasmid libraries, or increasing the average insert sizes, it is possible to close the gaps, allowing continuous assembly of the genome sequence. Therefore, planning the use of several libraries with various size distributions is beneficial.

Gap Filling by Primer Walking Sequencing or by Transposon Insertion into BACs

Primer walking sequencing is highly useful for gap filling, especially for sequences generated with the clone-by-clone approach. In the approach, the regions with obvious gaps are known as BAC clones, and primers can be designed immediately following the known sequences to extend sequencing.

The BAC clones can also be further sequenced by creating transposon insertion sublibraries. In this case, transposon sequences are randomly inserted into the BAC where gaps exist, creating a sublibrary. The transposon sequences provide primer-binding sites for sequencing the neighbor regions. Kits for making transposon insertion libraries are available (e.g., the Tn5 transposon insertion system of EPICENTRE (http://www.epibio.com/a_simple_invitro_transposition_reaction.asp). The transposon insertion clone is particularly useful for genomic regions containing highly repetitive or simple sequences prohibiting the design of sequencing primers. Obviously, such regions can also be dealt with PCR amplification followed by cloning into a plasmid vector and sequencing.

Scaffolding

Scaffolding is the contig merging processes using not only sequencing information, but also the mate pair sequencing reads (Figure 26.5). The contigs produced by an assembly program can be ordered and oriented along a chromosome using mate pair sequencing reads. Mate pair sequencing reads are sequencing reads obtained from both ends of a single clone. For instance, sequencing of a BAC from both ends would generate mate pair reads that are physically linking them together by the distance of the BAC insert size. Similarly, mate pair reads are obtained by sequencing both ends

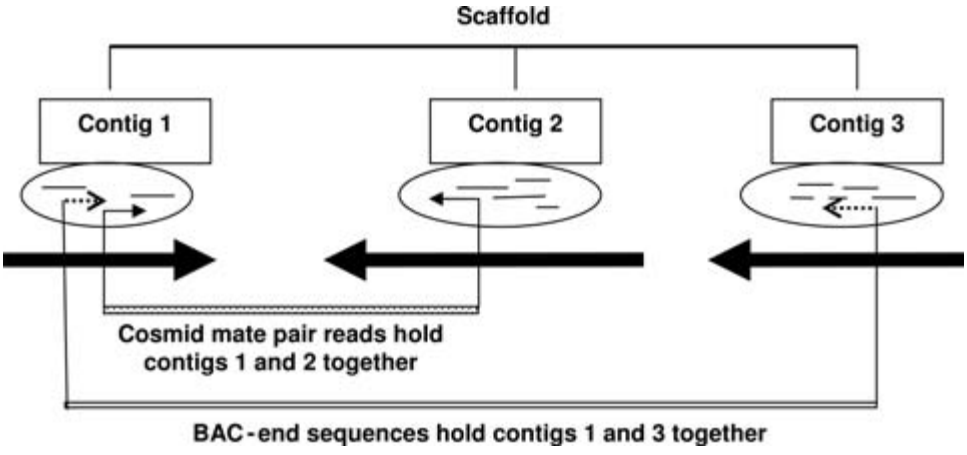


Figure 26.5. Schematic presentation of scaffolding using mate pair sequences. Shown are examples of mate pair sequences from a BAC and a Cosmid clone.

of cosmid libraries or plasmid libraries. Within the assembly, the paired end reads must be placed at a distance consistent with the size of the library from which they originate and must be oriented toward each other. Within an assembly, each read is assigned an orientation corresponding to the DNA strand from which the read was generated. The constraints provided by mate pairs lead to constraints on the relative order and orientation of the contigs. In such a process, the mate pair reads are used to order and orient the contigs along a chromosome, combining them into scaffold.

Sequence Annotation

After the initial generation of A, C, G, Ts, the raw sequence data provide very little biological insight. Annotation can greatly enhance the biological value of these sequences. Useful annotations include homologies to known genes, possible gene locations, gene signals such as promoters, etc. To use any sequence it must be interpreted in the context of other biological knowledge. This is the process of annotation, the task of adding explanatory notations to the sequence text. We define an annotation as the biological evaluation and explanation of a specific region on a nucleic acid sequence that includes any feature that can be anchored to the sequence, such as gene transcripts, an exon, a promoter, a transposable element, a regulatory region, mutations, cut sites, start and stop signals, transcription factor binding sites, probe or primer binding sites, or a CpG island (Lewis et al. 2002).

Sequence annotations are performed using powerful computer programs. Many programs were developed for the annotation of DNA sequences. Some examples include Genotator (Harris 1997), Artemis (Rutherford et al. 2000), AceDB (http:// www.acedb.org), Apollo (Lewis et al. 2002), Genamics Expression (http://genamics.com/expression/annotating.htm), EAnnot (Ding et al. 2004), the distributed annotation system (DAS) (http://stein.cshl.org/das/), etc. For the aquaculture genome community, much can be learned from the human genome research community and the recent sequencing projects of livestock animals such as chicken, bovine, and swine sequencing projects. In most cases, however, genome annotation is a complex task requiring assistance from large genome sequencing centers, or bioinformatics experts. The biological information, however, most often comes from the research community.

What is Next with A, C, G, Ts?

After the generation of genome sequence, it is not the end of genome biology. As described at the U.S. Department of Energy's (DOE) Genome Project Web page, "The words of Winston Churchill, spoken in 1942 after 3 years of war, capture well the HGP era: 'Now this is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning.' The avalanche of genome data grows daily. The new challenge will be to use this vast reservoir of data to explore how genes are expressed and their networking, and interactions with one another and with the environment to create complex, dynamic living systems. Systematic studies of function on a grand scale, i.e., functional genomics, will be the focus of biological explorations in

this century and beyond. Deriving meaningful knowledge from DNA sequence will define biological research through the coming decades and require the expertise and creativity of teams of biologists, chemists, engineers, and computational scientists, among others.

Ethical, Legal, and Social Issues (ELSI) of Genome

Sequencing coming into the genomics era also brings with it many ethical, legal, and social issues (ELSI). This is sometimes also referred to as the GE3LS (genomics, ethics, environment, economics, law, and society). The ELSI is serious enough that the National Institutes of Health (NIH) and DOE spend 3–5% of their genome research budget to address such issues. Many questions arise as a result of genome sequence availability. Although we may not have answers for many of these questions, some examples of the questions posted on DOE's Genome Project Web site (http://doegenomes.org/) are relevant to many: *Who should have access to personal genetic information, and how will it be used? Who owns and controls genetic information? How does personal genetic information affect an individual and society's perceptions of that individual? How does genomic information affect members of minority communities? Do healthcare personnel properly counsel parents about the risks and limitations of genetic technology? How reliable and useful is fetal genetic testing? What are the larger societal issues raised by new reproductive technologies? How will genetic tests be evaluated and regulated for accuracy, reliability, and utility? (Currently, there is little regulation at the federal level). How do we prepare healthcare professionals for the new genetics? How do we prepare the public to make informed choices? How do we as a society balance current scientific limitations and social risk with long-term benefits? Should genetic testing be performed when no treatment is available? Should parents have the right to have their minor children tested for adult-onset diseases? Are genetic tests reliable and interpretable by the medical community? Do people's genes make them behave in a particular way? Can people always control their behavior? What is considered acceptable diversity? Where is the line between medical treatment and enhancement? Are GM foods and other products safe to humans and the environment? How will these technologies affect developing nations' dependence on the West? Who owns genes and other pieces of DNA? Will patenting DNA sequences limit their accessibility and development into useful products? etc., etc.*

Unfortunately, we do not have any genomes sequenced today from aquaculture species, but fortunately, the aquaculture genome sequence will not involve an equal number of challenges as the human genome sequence in relation to the ethical, legal, and social issues. However, many of the ethical, legal, and social issues will be similar to the questions above facing the human genome situation. By the time the genomes are sequenced from aquaculture species, hopefully answers will be found for many of the above questions. On the other hand, because the aquatic environments and many of the unique characteristics of aquaculture species, and their close relationship with the environment, the generation of aquaculture species genome sequence likely will pose challenging questions related to ethical, environmental, economic, legal, and social issues.

Conclusion

Although no genomes have been sequenced from aquaculture species, it is likely that some important aquaculture species will soon be subjected to genome sequencing. (After this chapter was written, NIH made a committment to sequence the tilapia genome.) The major reasons for this assessment include their economic importance, their close relationship with the environment, their roles in the study of genome evolution, and many unique characteristics of aquaculture species. In addition, given the declining world fisheries, aquaculture needs to be developed and sustained, which demands genome-based technologies. Many biological and production problems are unique to aquaculture species, and sequencing of related model species will not provide relevant information. For instance, the enteric septicemia of catfish is unique in catfish, and the zebrafish genome information may not help in elucidation of the disease resistance genes in catfish.

However, the low research funding, the relatively small research community, and the large numbers of species involved in aquaculture make the initial funding of genome sequencing projects difficult. There are no limitations in technology, biology, or readiness for genome sequencing, because basic genome resources are well prepared from several important aquaculture species including the salmonids, catfish, tilapia, shrimp, and the oysters. The limiting factor is financial. The hard part is to obtain the first pot of money. Once the first pot of funds is obtained, international collaboration is expected to meet the remaining demands.

The emerging sequencing technology such as the 454 sequencing platform brings greater hope to the aquaculture species for genome sequencing, as sequencing with a genome coverage of 5× cost several hundred thousand dollars that may be justified with funding agencies. The problem is with the ability of sequence assembly with the short sequencing reads of the 454 system. Let's hope the technology will improve soon. On the other hand, many "sequence islands" as produced by the 454 sequencing platform, when anchored by well-developed physical maps and genetic maps, may provide sufficient tools for studies of agricultural related issues such as performance and production traits.

References

Bonfield JK, KF Smith, and R Staden. 1995. A new DNA sequence assembly program. Nucleic Acids Res, 23, pp. 4992–4999.

Coulson A, J Sulston, S Brenner, and J Karn. 1986. Towards a physical map of the genome of the nematode *C. elegans.* Proc Natl Acad Sci, 83, 7821–7825.

Dear S, R Durbin, L Hilloier, G Marth, J Thierry-Mieg, and R Mott. 1998. Sequence Assembly with CAFTOOLS. Genome Res, 8, pp. 260–267.

Ding L, A Sabo, N Berkowicz, RR Meyer, Y Shotland, MR Johnson, KH Pepin, RK Wilson, and J Spieth. 2004. EAnnot: a genome annotation tool using experimental evidence. Genome Res, 14, pp. 2503–2509.

Ewing B and P Green. 1998. Basecalling of automated sequencer traces using Phred. II. Error probabilities. Genome Res, 8, pp. 186–194.

Ewing B, L Hillier, MC Wendl, and P Green. 1998. Basecalling of automated sequencer traces using Phred. I. Accuracy assessment. Genome Res, 8, pp. 175–185.

Harris NL. 1997. Genotator: a workbench for sequence annotation. Genome Res, 7, pp. 754–762.
Huang X. 1996. An improved sequence assembly program. Genomics, 33, pp. 21–31.
Lander ES and MS Waterman. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics, 2, pp. 231–239.
Lewis SE, SM Searle, N Harris, M Gibson, V Lyer, J Richter, C Wiel, L Bayraktaroglir, E Birney, MA Crosby, JS Kaminker, BB Matthews, SE Prochnik, CD Smithy, JL Tupy, GM Rubin, S Misra, CJ Mungall, and ME Clamp. 2002. Apollo: a sequence annotation editor. Genome Biol, 3, RESEARCH0082.
Luo MC, C Thomas, FM You, J Hsiao, S Ouyang, CR Buell, M Malandro, PE McGuire, OD Anderson, and J Dvorak. 2003. High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. Genomics, 82, pp. 378–389.
Notredame C and DG Higgins. 1996. SAGA: sequence alignment by genetic algorithm. Nucleic Acids Res, 24, pp. 1515–1524.
Parsons R, S Forrest, and C Burks. 1993. Genetic algorithms for DNA sequence assembly. ISMB, pp. 310–318.
Peltola H, H Soderlund, and E Ukkonen. 1984. SEQAID: a DNA sequence assembling program based on a mathematical model. Nucleic Acids Res, 12, pp. 307–321.
Rutherford K, J Parkhill, J Crook, T Horsnell, P Rice, MA Rajandream, and B Barrell. 2000. Artemis: sequence visualization and annotation. Bioinformatics, 16, pp. 944–945.
Soderlund C, S Humphray, A Dunham, and L French. 2000. Contigs built with fingerprints, markers, and FPC V4.7. Genome Res, 10, pp. 1772–1787.
Soderlund C, I Longden, and R Mott. 1997. FPC: a system for building contigs from restriction fingerprinted clones. Comput Appl Biosci, 13, pp. 523–535.
Venter JC, HO Smith, and L Hood. 1996. A new strategy for genome sequencing. Nature, 381, pp. 364–366.
Xu P, SL Wang, L Liu, E Peatman, B Somridhivej, J Thimmapuram, G Gong, and Z Liu. 2006. Channel catfish BAC-end sequences for marker development and assessment of syntenic conservation with other fish species. Animal Genetics, 37, pp. 321–326.
Zhang C and AK Wong. 1997. A genetic algorithm for multiple molecular sequence alignment. Comput Appl Biosci, 13, pp. 565–581.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 27
Bioinformatics

Lei Liu

Introduction

Bioinformatics emerged from the revolutionary development of "omics" in biotechnology. Various high throughput technologies such as automated DNA sequencing and microarray enable the production of very high volume and extremely complex dataset. Analysis of such huge and complicated data becomes a big challenge and "bottleneck." Combining biological studies with computer science, mathematics, and statistics, bioinformatics develops methods, solutions, databases, and software to discover patterns, generate models, and facilitate bench biologists on understanding of genomic data.

Bioinformatics touches almost all levels of high throughput biological studies including genomics, which deciphers the complete DNA sequences of all of the genetic information in an organism. Functional genomics, which measures the messenger RNA (mRNA) levels using microarray technologies to monitor the gene expression of all the genes in a genome known as transcriptome. Proteomics dissects the complete set of proteins in a cell at a certain stage. Metabolomics monitors all the metabolic compounds in a cell known as metabolome. Many more "omics" technologies exist as well. Above all, one major task of bioinformatics is to integrate data from different levels and prior biological knowledge to achieve system level understanding of biological phenomena. Since it is impossible to cover every aspect in bioinformatics in a short chapter, I will provide a general overview of the field and focus on several key areas that may be highly related to aquagenomics including sequence alignment, phylogenetic analysis, genome analysis, and microarray analysis. In the Sequence Alignment section, I will first introduce the basic concept and techniques of sequence alignment. This area deals with problems such as how to make an optimal alignment between two sequences and how to search sequence databases quickly with an unknown sequence. I then will introduce several widely used programs. In the Phylogenetic Trees section, I will discuss multiple alignment and building phylogenetic trees to infer evolution history. In the Genome Analysis section, I will mainly deal with the data analysis at genome scale. The problems include contig assembly, gene prediction, expressed sequence tag (EST) analysis, gene function annotation, and comparative genomics. Many techniques of sequence analysis are used in genome analysis, but many new methods were developed for the unique problems. In the Microarray Analysis section, I will introduce some basic concepts and techniques that are widely used. The problems in this area are completely different from sequence analysis. Many statistical and data mining techniques are applied in the field. Nonetheless, bioinformatics has become an established program at many major universities, and covering such a new but broad area by a single chapter can be difficult.

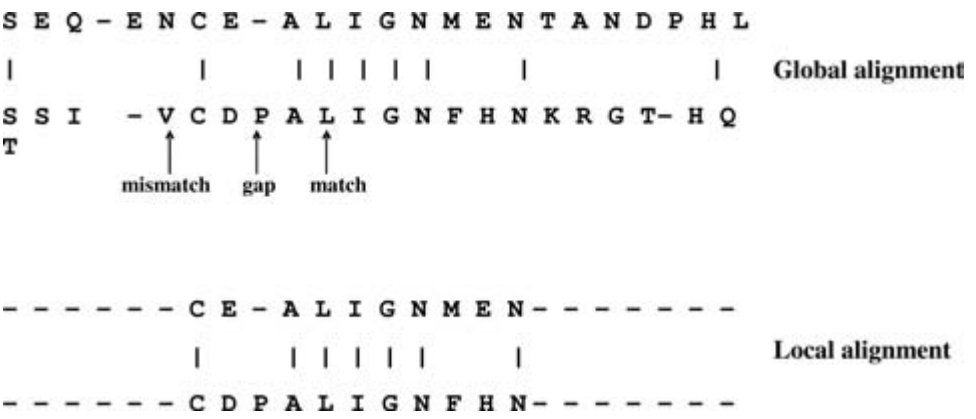


Figure 27.1. Pair-wise alignment and the distinction between global and local alignment.

This chapter is written to serve as an introduction of bioinformatics relevant to aquaculture genomics. Interested readers are referred to a few excellent books on bioinformatics (e.g., Mount 2004, Baxevanis and Ouellette 2004, Claverie and Notredame 2003, Gentleman et al. 2005, Baldi and Brunak 2001, Jones and Pevzner 2004).

Sequence Alignment

Sequence alignment is used to compare two (pair-wise alignment) or more (multiple sequence alignment) DNA or protein sequences. In this section, we mainly discuss pair-wise alignment. The basic idea of sequence alignment is to align as many identical characters as possible when putting the two sequences parallel on a page. For a given character in one sequence, there are three possible situations in the other sequence: a matching character, a mismatching character, or a gap as shown in Figure 27.1. Gaps are introduced to maximize the number of matched character pairs. There are two types of alignment, global and local as shown in Figure 27.1. Global alignment tries to align the entire sequence, up to the end of each sequence. Local alignment identifies a stretch of high-density matching pairs, which forms subalignments in the aligned sequences.

Pair-wise Alignment

Two sequences can be aligned in many ways. The sequence alignment problem is actually a problem to find the optimal alignment between given sequences, which can be described as the following: Given two strings of text, X and Y, which may be DNA or amino acid sequences, find the optimal way of inserting dashes into the two sequences to maximize a given scoring function between them. The scoring function depends on both the length of the regions of consecutive dashes and on the pairs of characters that are in the same position when gaps have been inserted. Figure 27.2 gives an example of a scoring scheme.

```
Sequence 1    A  C  G  G  A  -  T  C  A
              |     |     |     |  |
Sequence 2    A  G  G  A  A  T  T  C  G
Score (4)     1  0  1  0  1 -1  1  1  0

Total score = Sum of all the scores of aligned pairs
```

Figure 27.2. A simple scoring scheme in pair-wise alignment. Each match scores 1; each mismatch scores 0; and each gap scores −1.

Usually we use scoring matrix to define the scores between a pair of letters. For DNA sequences, the scoring matrix can be as simple as shown in Figure 27.2. For protein, the scoring matrix can be more complicated because there are 20 alphabets in the sequences. The scoring matrix needs to reflect the fact that some of the amino acids are similar to one another and some are different based on their physical-chemical properties. We want to assign large positive scores for two identical residues, small positive scores for similar residues (e.g., D and E), and small negative scores for different residues (e.g., D and L). Many scoring matrices have been developed for protein sequence alignment such as PAM and BLOSUM matrices. The alignment algorithms work the same way no matter what scoring matrix is used, but the optimal alignment produced depends on the scoring system. Another part of the scoring system is the gap penalty, which is kept quite simple in many different alignment programs. In the example of Figure 27.2, we used −1 as a gap penalty.

The "optimal alignment" between two sequences depends on the scoring function that is used. As we shall see, an optimal sequence alignment for a given scoring function may not be unique. A dynamic programming approach to sequence alignment was proposed by Needleman and Wunsch (1970). The idea behind the dynamic programming approach can be explained using the two sequences, CCGAT and CA-AT, of Figure 27.3a. If we break this alignment into two parts (Figure 27.3b), we have two alignments: the left is the alignment of the two sequences CCGA and CA-A, and the right is the alignment of the last elements T-T. If the scoring system is additive, then the score of the alignment of Figure 27.3b is the sum of the scores of the four base-alignment on the left plus the score of the alignment of the pair T-T on the right. If the alignment in Figure 27.3a is optimal then the four-base alignment in the left side of Figure 27.3b must also be optimal. If this were not the case (for example if a better alignment would be obtained by aligning A with G), then the optimal alignment of Figure 27.3c would lead to a higher score than the alignment shown in Figure 27.3a. The optimal alignment ending at any stage is therefore equal to the total (cumulative) score of the optimal alignment at the previous stage plus the score assigned to the aligned elements at that current stage.

The optimal alignment of two sequence ends with either of the last two symbols aligned, the last symbol of one sequence aligned to a gap, or the last symbol of the other sequence aligned to a gap. In our analysis x_i refers to the i^{th} symbol in sequence 1 and y_j refers to the j^{th} symbol in sequence 2 before any alignment has been made. We

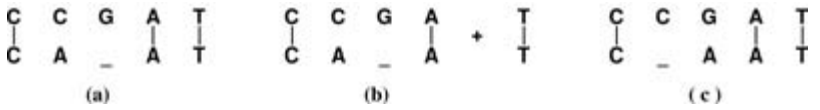


Figure 27.3. An overview of the dynamic programming approach.

will use the symbol $S(i, j)$ to refer to the cumulative score of the alignment up until symbols x_i and y_j, and the symbol $s(x_i, y_j)$ to refer to the score assigned to matching elements x_i and y_j. We will use d to refer to the cost associated with introducing a gap.

1. If the current stage of the alignment matches two symbols, x_i and y_j, then the score, $S(i, j)$, is equal to the previous score, $S(i - 1, j - 1)$, plus the score assigned to aligning the two symbols, $s(x_i, y_j)$.
2. If the current match is between symbol x_i in sequence 1 and a gap in sequence 2 then the new score is equal to the score up until symbol x_{i-1} and the same symbol y_j, $S(i-1, j)$, plus the penalty associated with introducing a gap, $-d$.
3. If the current match is between symbol y_j in sequence 2 and a gap in sequence 1 then the new score is equal to the previous score up until symbol y_{j-1} and the same symbol x_i, $S(i, j-1)$, plus the gap penalty $-d$.

The optimal cumulative score at symbols x_i and y_j is:

$$S(i, j) = \max \begin{cases} S(i - 1, j - 1) + S(x_i, y_j) \\ S(i - 1, j) - d \\ S(i, j - 1) - d \end{cases} \tag{27.1}$$

The previous equation determines the new elements at each stage in the alignment by successive iterations from the previous stages. The maximum at any stage may not be unique. The optimal sequence alignment (s) is the one that provides the highest score. This is usually performed using a matrix representation, where the cells in the matrix are assigned an optimal score, and the optimal alignment is determined by a process called trace back (Gotoh 1982, Durbin et al. 1998).

The optimal alignment between two sequences depends on the scoring function that is used. This brings the need for a score that is biologically significant and relevant to the phenomenon being analyzed. Substitution matrices present one method of achieving this using a "log-odds" scoring system. One of the first substitution matrices used to score amino acid sequences was developed by Dayhoff and others (1978). Other matrices such as the BLOSUM50 matrix were also developed using databases of more distantly related proteins.

The Needleman-Wunsch (N-W) algorithm and its variation (Gotoh 1982) provide the best global alignment for two given sequences. Smith and Waterman (1981) presented another dynamic programming algorithm that deals with finding the best local alignment for smaller subsequences of two given sequences rather than the best global alignment of the two sequences. The local alignment algorithm identifies a pair of subsegments, one from each of the given sequences, so that there is no other pair of subsegments with greater similarity.

Heuristic Alignment Methods (BLAST)

The dynamic programming based alignment can produce optimal alignments, but in practice, the algorithm can be very slow. Heuristic search methods for sequence alignment have gained popularity and extensive use in practice because of the complexity and large number of calculations in the dynamic programming approach, especially for database searches. Heuristic approaches search for local alignments of subsegments and use these alignments as "seeds" in which to extend out to longer sequences. The most widely used heuristic search method available today is the Basic Local Alignment Search Tool (BLAST) by Altschul and others (1990). BLAST alignments define a measure of similarity called Maximal Segment Pair (MSP) as the highest scoring pair of identical length subsegments from two sequences. The lengths of the subsegments are chosen to maximize the MSP score. It then tries to extend an alignment from the matching words in each direction along the sequences, continuing for as long as the score continued to increase. At the point the extension stops, a larger stretch of sequence, called high-scoring segment pair (HSP) is found. BLAST next determines the significance of each HSP score by calculating the probability of two sequences being aligned by random chance. People often use the e-value to determine the significance of a database hit by BLAST. The smaller the e-value, the more significant the hit is. But the e-value is calculated based on database size (roughly the total length of all the sequences in the database). Therefore, we may not be able to compare e-values across completely different databases. BLAST is now a suite of tools provided by the National Center for Biotechnology Information (NCBI). Users can download the BLAST tools for local use or go to the NCBI BLAST site (http://www.ncbi.nlm.nih.gov/BLAST/) for online use. When it is installed locally, users need to format the sequence databases using the 'formatdb' program in the BLAST tools to set up a searchable database. The input of the 'formatdb' program is a FASTA format file, which is a text file with the following format:

```
>gi|1195552|gb|AAA87954.1| phosphoinositide-specific phospholipase C
KLIDKYEPDLTAKQNKHMTKDGFLAYLTHLEGFIFNPAHINVYQDM
RQPINHYFISSSHNNYLMQEQLKGASSTEGYIRALLKSCRCVELDCW
DGPNGEPVIYHGHTLTSKVLFKDVIKAINEYAFKTSEYPVILSLENHC
SVEQQKLMAHHMTSILGSALLTRPLGKEIPSSLPSPEELKGRILIKGK
RLNKLEAAFNNNAEDADSVSEEDEAAEAKQGEQETKPKKKKMKL
AKELSDLVIYCKSVHFNSFEHSREKQACYEMSSFKESKAKNLAEHSA
TEFIQHNMDKLSRIYPAGTRTDSSNYSPVTMWNAGCQIVALNFQTY
GKEMDVNQGRFLPNGKCGYILKPEFMRNPDSRFDPNAPTEGRWFR
KSILHVMVISAQQLPKLNKDKPKSIVDPFVKWEIFGVPGDQARAQTS
YINNNGFNPAWNKTFKFPIHVPDLAIVRFVVEDYDAASHNDFIGQYT
LPFTSIKNGYRHIPLLTKGGDVIPSAKLFVHIMFMDP
```

Each entry starts with '>' at the beginning of the first line. The first line is the comment line. Sequences start from the second line. A FASTA file can contain multiple sequence entries.

In BLAST tools, there are five different BLAST programs for different types of searches (Table 27.1). BLASTN compares DNA sequences. BLASTP compares protein sequences. BLASTX translates query DNA sequences into six reading frames

Table 27.1. A summary of the usage of different BLAST programs.

Programs	Query Sequence	Database	Formatdb option
BLASTN	DNA	DNA	Nonprotein
BLASTP	Protein	Protein	Protein
BLASTX	DNA	Protein	Protein
TBLASTN	Protein	DNA	Nonprotein
TBLASTX	DNA	DNA	Nonprotein

and then compares them to protein sequences. TBLASTN translates the DNA database into six reading frames and then compares with protein query sequence. TBLASTX compares DNA sequences, but at protein level by translating both query and database sequences into six reading frames. Translated BLAST tools (TBLASTN and TBLASTX) are very computing intensive and thus are very slow.

Multiple Sequence Alignments

Multiple sequence alignments are alignments of more than two sequences. The inclusion of additional sequences can improve the accuracy of the alignment, find protein motifs, identify related protein sequences in a database, and predict protein secondary structure. Multiple sequence alignments are also the first step in constructing phylogenetic trees.

The most common approach for multiple sequence alignments is progressive alignment, which involves choosing two sequences and performing a pair-wise alignment of the first to the second. The third sequence is then aligned to the first and the process is repeated until all the sequences are aligned. The score of the multiple alignments is the sum of scores of the pair-wise alignments. Pair-wise dynamic programming can be generalized to perform multiple alignments using the progressive alignment approach; however, it is computationally impractical even when only a few sequences are involved (Lipman et al. 1989). The sensitivity of progressive alignment was improved for divergent protein sequences using CLUSTALW (Thompson et al. 1994), available at the Web page (http://clustalw.genome.ad.jp/). Many other approaches to sequence alignment have been proposed in the literature. For example, a Bayesian approach was suggested for adaptive sequence alignments (Lawrence et al. 1993, Zhu et al. 1998). The data now available from the human genome project have suggested the need for aligning whole genome sequences where large-scale changes can be studied as opposed to single-gene insertions, deletions, and nucleotide substitutions. MuMMer (Delcher et al. 1999) follows this direction and performs alignments and comparisons of very large sequences.

Phylogenetic Trees

The main objectives of phylogenetic tree studies are (1) to reconstruct the genealogical ties between organisms and (2) to estimate the time of divergence between organisms

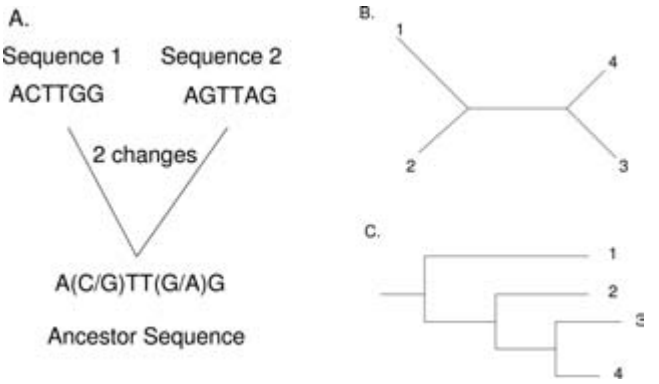


Figure 27.4. A. Concept of a phylogenetic tree. B. Unrooted tree. C. Rooted tree.

since they last shared a common ancestor. See Chapters 8 and 9. With the explosion of genomic data in the last few years, tree building has become more popular, where molecular-based phylogenetic studies have been used in many applications such as the study of gene evolution, population subdivisions, analysis of mating systems, paternity testing, environmental surveillance, and the origins of diseases that have transferred among species.

When two sequences (DNA or protein) found in two different organisms are similar, they are likely to have been derived from a common ancestor sequence, as shown in Figure 27.4A. Before we can build a phylogenetic tree, alignment is needed to show which positions in sequences were conserved. For a set of sequences, multiple alignment programs such as CLUSTALW are used. There are different ways to draw a phylogenetic tree. The unrooted tree as shown in Figure 27.4B shows the evolutionary relationships among sequences 1–4 without the location of the ancestral nodes. The rooted tree as shown in Figure 27.4C shows the steps in evolution with the end nodes as the present day sequences 1–4. In the rooted tree, if the length of the branches is made proportional to the number of changes, the branch length can then represent the evolutionary distance. If we assume that the evolutionary change happens at a constant rate, the branch length in a rooted tree can also represent the evolutionary time, which is known as a molecular clock.

There are many ways to construct the phylogenetic tree, which can generally be classified into one of four types: distance-based methods, parsimony methods, maximum likelihood methods, and Bayesian methods. For a detailed discussion of each of these methods, see Li (1997). Many software programs have been developed for the phylogenetic analysis based on these methods including PHYLIP (Felsenstein 1993) and Mr. Bayes (Huelsenbeck and Ronquist 2002).

Genome Analysis

Analysis of completely sequenced genomes has been one of the major driving forces for the development of the bioinformatics field. The major challenges in this area include genome assembly, gene prediction, function annotation, EST analysis, and

comparative genomics of conserved regions. For a genome project, one must ask several fundamental questions: How can we put the whole genome together from many small pieces of sequences? Where are the genes located on a chromosome? What are the functions of the genes? And so on. In this section, we will discuss the sequence assembly problem, gene prediction, EST analysis, functional annotation, and comparative genomic mapping.

Contig Assembly

In a high throughput DNA genome sequencing project, large DNA molecules with billions of base pairs, such as a human chromosome, are broken into smaller fragments (approximately 100 kilobase [kb]) and cloned into vectors such as bacterial artificial chromosome (BAC). These BAC clones can be tiled together by physical mapping techniques (see Chapter 14). Individual BACs can be further broken down into smaller random fragments of 1–2 kb. These fragments are sequenced and assembled based on overlapping fragments. With more fragments sequenced, there will be enough overlaps to cover most of the sequence. This method is often referred to as "shotgun sequencing." Computer tools were developed to assemble the small random fragments into large contigs based on the overlapping ends among the fragments using similar algorithms as the ones used in the basic sequence alignment. The widely used programs include PHRAP/Consed (Gordon et al. 1998, 2001) and CAP3 (Huang and Madan 1999).

The assembly programs first make sequence alignments. Based on the stringency of criteria for assembly including the length of overlapping region, the percentage of identity within the overlapping regions and the quality of the base calling, different contigs may be obtained. There is always a trade off between stringency and the length of the consensus sequences. If too stringent, it may produce too many short contigs. If too relaxed, it may produce too many misassemblies or mistakes. Usually the default parameters in PHRAP and CAP3 are optimized to balance the trade off. One criterion we can consider is the coverage of the overlap regions (how many reads in the overlapped region). A good contig should not have many single read covered areas. Most of the regions in a contig should be covered by multiple reads. Consed is a visualization and editing tool for PHRAP (as well as CAP3). One can edit the contigs based on human observation and correct misassembled regions. Most prokaryotic genomes can be sequenced directly by the "shotgun sequencing" strategy with special techniques for gap closure. For large genomes, such as the human genome, there are two strategies (see Chapter 26). One is to assemble large contigs first and then tile together the contigs based on the physical map to form the complete chromosome (Waterston et al. 2002). The other strategy is called whole genome shotgun (WGS) sequencing strategy, which assembles the genome directly from the "shotgun sequencing" data in combination with mapping information (Myers et al. 2000). WGS is a faster strategy to finish a large genome, but the challenge of WGS is how to deal with the large number of repetitive sequences in a genome. Nevertheless, WGS has been successfully used in completing the *Drosophila* and human genomes (Adams et al. 2000, Venter et al. 2001).

EST Analysis

ESTs represent partial descriptions of the transcribed portions of genomes. They are generated from single-pass complementary DNA (cDNA) library sequencing in high throughput manner (see Chapter 20). ESTs provide a quick survey of the gene contents in an organism for which genomic information is lacking. They are valuable resources for microarray production and genome annotation. Even though EST sequencing is similar to genomic sequencing, the analysis of ESTs has some unique features. EST analysis pipelines include many steps: Base calling is the step to generate the sequences from chromatograph. Each base called is accompanied with a score to indicate the quality or confidence of the call. Quality trimming is to trim off low quality bases based on the quality scores. Vector trimming is to trim off vector and adaptor sequences in the read. After trimming, some of the sequences may become very short. One can set a cut-off length and remove the short sequences. Repeat masking is to identify and mask the low complexity and known repetitive elements in the sequences. Filter for unwanted sequences is to remove ribosomal RNA (rRNA), vector, *E. coli* genomic sequences, and any other unwanted sequences, such as viral sequences. This step is usually carried out using the BLASTN program. After the removal of sequences from the above steps, the resulting sequences are considered to be "clean" sequences and used for further analysis of clustering and assembly. Clustering and assembly is to put the sequences representing the same gene or gene family together. In this step, we first cluster the sequences together based on their identity (most often incorrectly referred to as homology) identified by the sequence alignment programs such as BLAST. Next, we use the same programs that are used in contig assembly in genome sequences to assemble the sequences for each cluster and generate consensus sequences. These steps can be chained using programming languages such as Perl scripts to form a pipeline program. Many of these programs have been packaged into software such as ESTAP (Mao et al. 2003), EST-Web (Paquola et al. 2003), and ESTIMA (Kumar et al. 2004). The functional annotations of ESTs are mostly accomplished by inference from sequences whose functions are known.

Genome Annotation and Gene Prediction

After complete assembly of a genome, the next task is to decipher the information coded in the genome, which is often called genome annotation. The process includes the prediction of gene structures and other features on a chromosome and the function annotation of the genes. There are two basic types of genes in a genome: RNA genes and protein-encoding genes. The majority of genes in a genome are protein-encoding genes. Therefore, the big challenge is how to find protein-encoding regions in a genome. The simplest way to search for a protein-encoding region is to search for open reading frames (ORF), which is a contiguous set of codons between two stop codons. There are six possible reading frames for a given DNA sequence. Three of them start at the first, second, and third base, respectively. The other three reading frames are at the complementary strand. The longest ORFs between the start codon

and the stop codon in the same reading frame provide good, but not sufficient evidence of a protein-encoding region. Gene prediction is generally easier and more accurate in prokaryotic than eukaryotic organisms due to the intron/exon structure in eukaryote genes. Computational methods of gene prediction based on the Hidden Markov Model (HMM) have been quite successful, especially in prokaryote genome. These methods involve training a gene model to recognize genes in a particular organism. Because of the variations in codon usage, a model must be trained for each new genome. In prokaryote genome, genes are packed densely with relatively short intergenic sequences. The model reads through a sequence with unknown gene composition and finds the regions flanked by start and stop codons. The codon composition of a gene is different from that of an intergenic region and can be used as a discriminator for gene prediction. Several software tools, such as GeneMark and GLIMMER (Delcher et al. 1999) are widely used HMM methods in prokaryotic genome annotation. Similar ideas are also applied to eukaryote gene prediction. Because of the intron/exon structure, the model is much more complex with more attention on the boundary of intron and exon. Programs such as GeneScan (Burge and Karlin 1997) and GenomeScan (Yeh et al. 2001) are HMM methods for eukaryote gene prediction. Neural network-based methods have also been applied in eukaryote gene prediction, such as Grail (Xu and Uberbacher 1997). Additional information for gene prediction can be found using ESTs. Because cDNA is derived from mRNA, a match to an EST is a good indication that the genomic region encodes a gene. Functional annotation of the predicted genes is another major task in genome annotation. This process can also be viewed as gene classification with different functional classification systems such as protein families, metabolic pathways, and Gene Ontology. The simplest way is to infer annotation from the sequence similarity to a known gene, for example, BLAST search against a well-annotated protein database such as SWISS-PROT. A better way can be a search against protein family databases (e.g., Pfam), which are built based on profile HMMs. The widely used HMM alignment tools include HMMER (Eddy 1998) and SAM (Krogh et al. 1994). All automated annotation methods can produce mistakes. More accurate and precise annotation requires manual checking and a combination of information from different sources.

Besides the gene structures, other features such as promoters can be better analyzed with a finished genome. In prokaryotic organisms, genes involved in the same pathway are often organized in an operon structure. Finding operons in a finished genome provides information on the gene regulation. For eukaryotic organisms, the completed genomes provide upstream sequences for promoter search and prediction. Promoter prediction and detection has been a very challenging bioinformatics problem. The promoter regions are the binding sites for transcription factors. Promoter prediction is to discover the sequence patterns that are specific for transcription factor binding. Different motif finding algorithms have been applied including scoring matrix method (Stomo and Hartzell 1989), Gibbs sampling, and Multiple EM for Motif Elicitation (MEME) (Bailey and Elkan 1994). The results are not quite satisfactory. Recent studies using comparative genomics methods on the problem have produced some promising results and demonstrated that the promoters are conserved among closely related species. In addition, microarray studies can provide additional information for promoter discoveries. See the Microarray Analysis section.

Comparative Genomics

With more and more genomes being completely sequenced, comparative analysis becomes increasingly valuable and provides more insights of genome organization and evolution. One comparative analysis is based on the orthologous genes, called clusters of orthologous groups (COG). Two genes from two different organisms are considered orthologous genes if they are believed to come from a common ancestor gene. Another term, paralogous genes, refers to genes in one organism and related to each other by gene duplication events. In COG, proteins from all completed genomes are compared. All matching proteins in all the organisms are identified and grouped into orthologous groups by speciation and gene duplication events. Related orthologous groups are then clustered to form a COG that includes both orthologs and paralogs. These clusters correspond to classes of functions. Another type of comparative analysis is based on the alignment of the genomes and studies the gene orders and chromosomal rearrangements. A set of orthologous genes that show the same gene order along the chromosomes in two closely related species is called a synteny group. The corresponding region of the chromosomes is called synteny blocks (O'Brien et al. 1999). In closely related species, such as mammalian species, the gene orders are highly conserved. The gene orders are changed by chromosomal rearrangements during evolution including the inversion, translocation, fusion, and fission. By comparing completely sequenced genomes, for example, human and mouse genomes, we can reveal the rearrangement events. One challenging problem is to reconstruct the ancestral genome from the multiple genome comparisons and estimate the number and types of the rearrangements (Bourque and Pevzner 2002). Completion of sequencing of many genomes has produced a wealth of resources and a strong basis for comparing ordered gene maps of other closely related species (Waterston et al. 2002). Information from gene-rich species maps can be transferred to moderate-resolution gene maps of other species. Those comparative genome maps can provide a basis for understanding the rates and patterns of genome evolution.

Based on the comparative genome maps, traits mapped in one species may be mapped in other species *in silico* (O'Brien et al. 1999). Here, we introduce the comparative mapping by annotation and sequence similarity (COMPASS) strategy to predict map locations of sequences in genomes that have not been sequenced (Liu et al. 2004). The prerequisite of COMPASS is a comparative map table between the reference genome and the predicting genome. To implement the COMPASS strategy, Perl or other script pipelines can be written to automate the steps of process, including sequence similarity searches using the basic local alignment search tool for nucleotides (BLASTN), parsing BLASTN output, querying comparative map tables, calculating the predicted position, and finally composing the prediction reports. The chromosome location of a query sequence (e.g., an EST or a BAC-end sequence) can be predicted on the map of the species of interest, by BLASTN search against the reference genome sequence and the data from the synteny table. For instance, for any catfish sequence with significant BLASTN match with the zebrafish genome sequence, the position of the closest marker mapped in both human and cattle genomes can be found (Figure 27.5). If the position of the sequence of interest falls into a conserved segment on the catfish and zebrafish comparative map, then the position of this sequence can be predicted in the cattle genome. If the position of the sequence falls outside of a

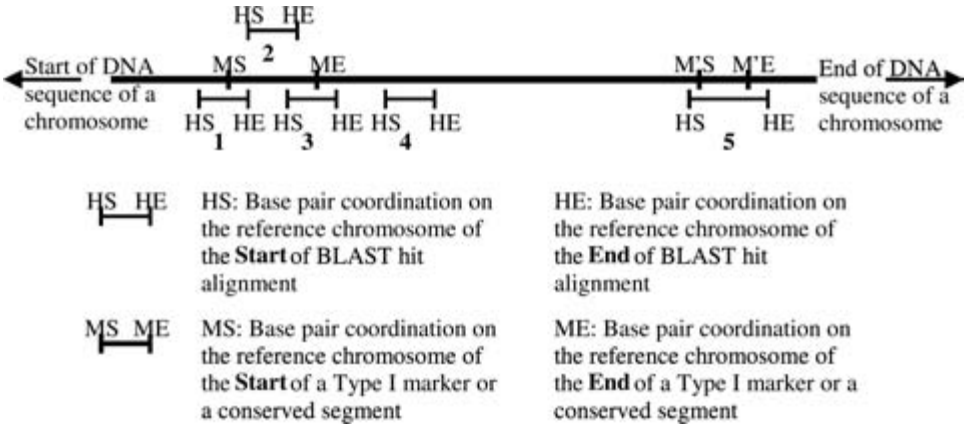


Figure 27.5. Illustration of different scenarios of BLASTN hits on reference chromosome and the definition of overlap between a marker and its closest marker. M and M' represent anchored markers and H represents a BLASTN hit alignment of a query sequence. Scenario 1, 2, 3, and 5: overlap with an anchored marker; Scenario 4: close to an anchored marker. (Reprint from Liu et al. 2004.)

conserved segment, then no prediction can be made at this time. The closest marker can be defined as the marker for which the BLASTN hit overlaps with or using the following calculation in the case that the BLASTN hit lies between two markers:

If $(HS - ME) < (M'S - HE)$, M is the closest marker;
If $(HS - ME) > (M'S - HE)$, M' is the closest marker.

HS, HE, ME, and M'S are denoted as shown in Figure 27.5.

Microarray Analysis

Microarray technologies allow biologists to monitor genome-wide patterns of gene expression in a high throughput fashion (see Chapters 21 and 22). DNA microarrays are typically composed of thousands of DNA sequences, called features, fixed to a glass or silicon substrate. The DNA sequences can be long (500–1,500 base pairs [bp]) cDNA sequences or shorter (25–70 mer) oligonucleotide sequences. The features can be deposited with a pin or piezoelectric spray on a glass slide, known as spotted array technology. Oligonucleotide sequences can also be synthesized *in situ* on a silicon chip by photolithographic technology (i.e., Affymetrix GeneChip). Relative quantitative detection of gene expression can be carried out between two samples on one array (spotted array) or by single samples comparing multiple arrays (Affymetrix GeneChip). In spotted array experiments, samples from two sources are labeled with different fluorescent molecules (Cy3 and Cy5) and hybridized together on the same array. The relative fluorescence between each dye on each spot is then recorded and a composite image may be produced. The relative intensities of each channel represent the relative abundance of the RNA or DNA product in each of the two samples. In

Affymetrix GeneChip experiments, each sample is labeled with the same dye and hybridized to different arrays. The absolute fluorescent values of each spot may then be scaled and compared with the same spot across arrays. For an example of a composite image from spotted arrays, readers are referred to Chapter 22.

Microarray analyses usually include several steps: image analysis and data extraction, data quantification and normalization, identification of differentially expressed genes, and knowledge discovery by data mining techniques such as clustering and classification. Image analysis and data extraction are fully automated and mainly carried out using a commercial software package or a freeware depending on the technology platforms. For example, Affymetrix developed standard data processing procedures and software for its GeneChips (for detailed information http://www.affymetrix.com); GenePix is widely used image analysis software for spotted arrays. For the rest of the steps, the detailed procedures may vary depending on the experiment design and goals. We will discuss some of the procedures below.

Normalization

The purpose of normalization is to adjust for systematic variations, primarily for labeling and hybridization efficiency so that we can discover true biological variations as defined by the microarray experiment (Bolstad et al. 2003). For example, as shown in the self-hybridization scatter plot (Figure 27.6) for a two-dye spotted array, variations (dye bias) between dyes are obvious and related to spot intensities. To correct the dye bias, one can apply the following model:

$$\log_2(\mathrm{R/G}) \Rightarrow \log_2(\mathrm{R/G}) - \mathrm{c(A)} \tag{27.2}$$

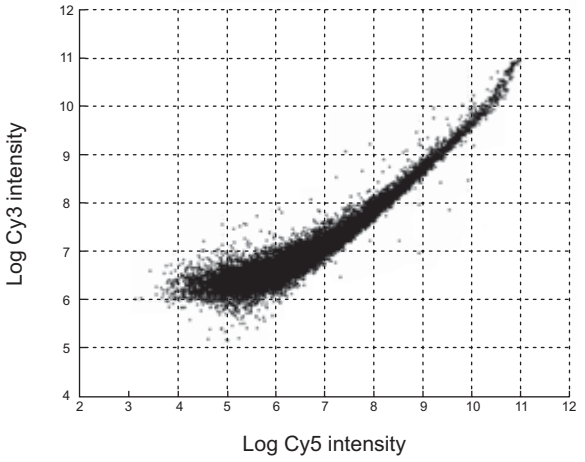


Figure 27.6. Self-hybridization scatter plot. Y-axis is the intensity from one dye; X-axis is the intensity from the other dye. Each spot is a gene.

Where R and G are the intensities of the dyes; A is the signal strength [$\log_2$(R*G)/2); M is the logarithm ratio ($\log_2$(R/G)]; c(A) is the locally weighted polynomial regression (LOWESS) fit to the MA plot (Yang et al. 2002, Yang and Thorne 2003).

Statistical Analysis for Discovering Differentially Expressed Genes

After correction of systematic variations, we want to determine which genes are significantly changed during the experiment and to assign appropriately adjusted p-values to the genes. Linear model approaches will be used to study how the genes respond over conditions (tissue, time, genotype, or infection status) while accounting for the various sources of experimental noise. Response variables include the absolute and relative (ratio of) fluorescence levels from the competing mRNA samples as described by Kerr and Churchill (2000). Considering the typical distribution of expression levels, a logarithmic transformation of the expression values is expected to better correspond to the model assumptions. Other transformations will be considered if the logarithmic transformation was not suitable. To discover the effects among experimental conditions, pair-wise comparisons were performed among all treatments by fitting the analysis of variance (ANOVA) models gene by gene.

Suppose that y_{ij} is the log ratio between treatment channel and control channel of ith condition and jth array, the model we use follows:

$$y_{ij} = \mu + A_i + B_j + e_{ij} \qquad (27.3)$$

where μ is the overall mean of all conditions; A_i is the effect of ith condition; B_j is the effect of jth array and is considered to be the random effect, which follows normal distribution N (0, σ_b^2); e_{ij} is identically independently distributed error term and follows Normal distribution N (0, σ_e^2). For example, to compare condition i and condition l treatments, we test whether $A_i - A_l$ is significantly different from zero. For T = $(A_i - A_l)$/se$(A_i - A_l)$, where se is the standard error, if absolute value of T is very large, we will decide that i and l conditions are significantly different, otherwise, there is no significant difference between them. Since many genes are tested, multiple test adjustment is needed. The false discovery rate (FDR) adjustment is often used, which adjusts p-values using the method of Benjamini and Hochberg (1995).

For Affymetrix GeneChips analysis, even though the basic steps are the same as spotted microarrays, because of the difference in technology, different statistical methods were developed. Besides the statistical methods provided by Affymetrix, several popular methods are packaged into software such as dChip (Li and Wong 2001) and RMA (Bolstad et al. 2003) in Bioconductor (http://www.bioconductor.org). With rapid accumulation of microarray data, one challenging problem is how to compare microarray data across different technology platforms. Some recent studies on data agreements have provided some guidance (Wang et al. 2005, Culhane et al. 2003).

Clustering and Classification

Once a list of significant genes has been obtained from the statistical test, we would apply different data mining techniques to find interesting patterns. At this step, the

microarray data set is organized as a matrix. Each column represents a condition; each row represents a gene. An entry is the expression level of the gene under the corresponding condition. If a set of genes exhibits a similar fluctuation under all of the conditions, it may indicate that these genes are co-regulated. One way to discover the co-regulated genes is to cluster genes with similar fluctuation patterns using various clustering algorithms. Hierarchical clustering was the first clustering method applied to the problem (Eisen et al. 1998). The result of hierarchical clustering forms a two-dimensional dendrogram as shown in Figure 27.7. The measurement used in the clustering process can be either a similarity such as Pearson's correlation coefficient or a distance such as Euclidian distance.

Many different clustering methods have been applied later on, such as, K-means (Ben-Dor et al. 1999), self-organizing map (SOM) (Tamayo et al. 1999), and support vector machine (Alter et al. 2000). In the K-means approach, all of the genes are compared to all of the vectors that correspond to each partition. Each gene is partitioned

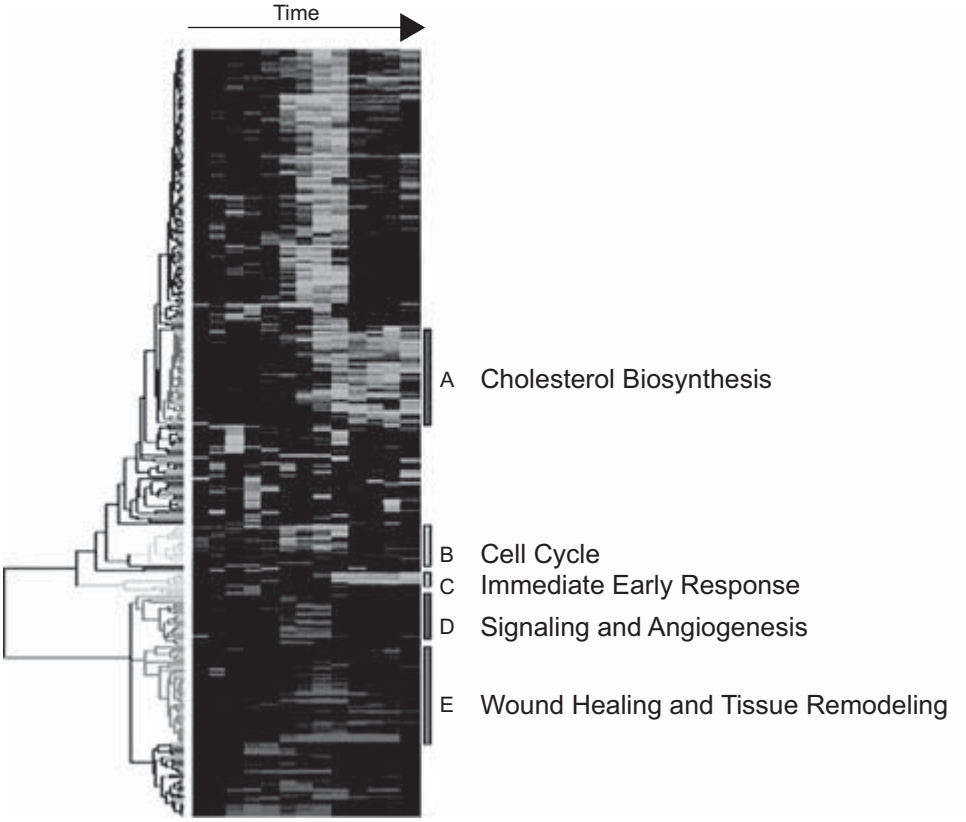


Figure 27.7. Hierarchical clustering of microarray data. Rows are genes. Columns are RNA samples at different time points. Values are the signals (expression levels), which are represented by the color spectrum. The bars beside the dendrogram show the clusters of genes, which exhibit similar expression profiles (patterns). The bars are labeled with letters and description of possible biological processes involving the genes in the clusters. (Reprinted from Eisen et al. 1998.) (Also see color plate.)

to the partition that has the most similar associated vector. After partitioning of all of the genes, the vectors of each partition are calculated as the mean of the genes that are partitioned to that vector. This process is repeated iteratively, with repartitioning of the genes, and recalculation of the vectors until all genes map to the same partitions twice in a row. In SOMs, the user must provide x and y dimensions that determine the number of partitions to segregate the data. In K-means and SOM, the total number of partitions is provided by the user, which influences the final clusters. Hence, multiple partition numbers will be tested to evaluate the sensitivity of the results to the partition specification.

Another type of microarray study involves classification techniques. For example, we can use the gene expression profile to classify cancer types. Golub and others (1999) first reported using classification techniques to classify two different types of leukemia. Many commercial software packages (e.g., GeneSpring and Spotfire) offer the use of these algorithms for microarray analyses.

Conclusion

In this chapter, we have discussed some basic concepts and techniques in bioinformatics. Because of the limitation of the space, we have emphasized several key areas that closely related to aquagenomics. Sequence alignment is one of the fundamental techniques in bioinformatics, which has been applied in genome analysis. Microarray analysis uses a completely different set of techniques mostly from statistics and data mining areas. Bioinformatics is a very broad area that touches many aspects of modern biology. Many other areas are not mentioned in this chapter including biological network analysis, proteomic analysis, structural biology, and dynamic modeling of biological processes. Once again, because of the limitation of the scope of this chapter, interested readers are referred to highly specialized bioinformatics books.

References

Adams MD, SE Celniker, RA Holt, CA Evans, JD Gocayne, et al. 2000. The genome sequence of Drosophila melanogaster. Science, 287, pp. 2185–2195.

Alter O, PO Brown, and D Bostein. 2000. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci USA, 97, pp. 10101–10106.

Altschul SF, W Gish, W Miller, E Myers, and J Lipman. 1990. Basic Local Alignment Search Tool. J Mol Biol, 215, pp. 403–410.

Bailey LT and C Elkan. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28–36.

Baldi P and D Brunak. 2001. Bioinformatics: The Machine Learning Approach. 2nd Ed. MIT Press, Cambridge, MA.

Baldi P, Y Chauvin, T Hunkapillar, and M McClure. 1994. Hidden Markov Models of Biological Primary Sequence Information. Proc Natl Acad Sci USA, 91, pp. 1059–1063.

Baldi P and GW Hatfield. 2001. Microarrays and Gene Expression. Cambridge University Press, Cambridge, UK.

Baxevanis AD and BF Ouellette. 2004. Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. Wiley-Interscience, New York, NY.

Ben-Dor A, R Shamir, and Z Yakhini. 1999. Clustering gene expression patterns. J Comp Biol, 6, pp. 281–297.

Benjamini Y and Y Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Royal Statistical Society, Series B, 57, pp. 289–300.

Bolstad BM, RA Irizarry, M Astrand, and TP Speed. 2003. A Comparison of Normalization Methods for High Density oligonucleotide Array Data Based on Bias and Variance. Bioinformatics, 19, pp. 185–193.

Bork P, T Dandekar, Y Diaz-Lazcoz, F Eisenhaber, M Huynen, and Y Yuan. 1998. Predicting Function: From Genes to Genomes and Back. J Mol Biol, 283, pp. 707–725.

Bourque G and AP Pevzner. 2002. Genome-scale evolution: reconstructing gene orders in the ancestral species, Genome Res, 12, pp. 26–36.

Bower J and H Bolouri. 2001. Computational Modeling of Genetic and Biochemical Networks. MIT Press, Cambridge, MA.

Bray N, I Dubchak, and L Pachter. 2003. AVID: A global alignment program. Genome Res, 13, pp. 97–102.

Brown PO and D Botstein. 1999. Exploring the new world of the genome with DNA microarrays. Nature Genetics, 21, pp. 33–7.

Brudno M, CB Do, GM Cooper, MF Kim, E Davydov, ED Green, A Sidow, and A Batzoglou. 2003a. LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res, 13, pp. 721–31.

Brudno M, S Malde, A Poiakov, C Do, O Couronne, I Dubchak, and A Batzoglou. 2003b. Glocal Alignment: Finding Rearrangements during alignment. Bioinformatics Special Issue on the Proceedings of the ISMB, 19, pp. 54i–62i.

Burge C and S Karlin. 1997. Prediction of complete gene structures in human genomic DNA. J Mol Biol, 268, pp. 78–94.

Claverie JM and C Notredame. 2003. Bioinformatics for Dummies. Wiley-Interscience, New York, NY.

Cohen FE. 1999. Protein misfolding and prion diseases. J Mol Biol, 293, pp. 313–320.

Culhane AC, G Perriere, and DG Higgins. 2003. Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. BMC Bioinformatics, 4, p. 59.

Dayhoff MO, RM Schwartz, and BC Orcutt. 1978. A model of evolutionary change in proteins. Atlas of Protein Sequence and Structure. Vol 5, supplement 3. National Biomedical Research Foundation, Washington DC, pp. 345–352.

Delcher AL, D Harmon, S Kasif, O White, and SL Salzberg. 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res, 27, pp. 4636–4641.

Dudoit S, J Fridlyand, and TP Speed. 2002. Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association, 97, pp. 77–87.

Durbin S, S Eddy, A Krogh, and G Mitchison. 1998. Biological Sequence Analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge, UK.

Eddy S. 1998. Profile hidden Markov models. Bioinformatics, 14, pp. 755–763.

Eddy S, G Mitchison, and R Durbin. 1995. Maximum Discrimination Hidden Markov Models of Sequence Consensus. J Comput Biol, 2, pp. 9–23.

Efron B, E Halloran, and S Holmes. 1996. Bootstrap confidence levels for phylogenetic trees. Proc Natl Acad Sci USA, 93, pp. 13429–34.

Eisen MB, PT Spellman, PO Brown, and D Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA, 95, pp. 14863–14868.

Farris JS. 1983. The logical basis of phylogenetic analysis, In Advances in Cladistics, vol. 2, N Platnick and V Funk, Eds., pp. 7–36.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol, 17, pp. 368–76.

Felsenstein J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c, Distributed by the author. Department of Genetics, University of Washington, Seattle. http://evolution.genetics.washington.edu/phylip.html.

Fitch WM and E Margoliash. 1967. Construction of phylogenetic trees. Science, 155, pp. 279–284.

Gardner M. 1997. The Last Recreations, Copernicus-Springer Verlag, New York.

Gentleman R, V Carey, W Huber, R Irizarry, and S Dudoit. 2005. Bioinformatics and Computational Biology Solutions Using R and Bioconductor, Statistics for Biology and Health. Springer, New York, NY.

Golub TR, DK Slonim, P Tamayo, C Huard, M Gaasenbeek, JP Mesirov, H Coller, ML Loh, JR Downing, MA Caligiuri, CD Bloomfield, and ES Lander. 1999. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science, 286, pp. 531–537.

Gordon D, C Abajian, and P Green. 1998. Consed: a graphical tool for sequence finishing. Genome Res, 8, pp. 195–202.

Gordon D, C Desmarais, and P Green. 2001. Automated finishing with autofinish. Genome Res, 11, pp. 614–625.

Gotoh O. 1982. An Improved Algorithm for Matching Biological Sequences. J Mol Bio, 162, pp. 705–708.

Gribaldo S and P Cammarano. 1998. The root of the universal tree of life inferred from anciently duplicated genes encoding components of the protein-targeting machinery. J Mol Evol, 47, pp. 508–516.

Helden JV, B Andre, and J Collado-Vides. 1998. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Bio, 281, pp. 827–842.

Holmes S. 2003. Bootstrapping Phylogenetic Trees: Theory and Methods. Statist Sci, 18, pp. 241–255.

Huang X and A Madan. 1999. CAP3: A DNA sequence assembly program. Genome Res, 9, pp. 868–877.

Huelsenbeck J and F Ronquist. 2002. Mr. Bayes. Bayesian Inference of Phylogeny. http://morphbank.ebc.uu.se/mrbayes/links.php.

Jones NC and PA Pevzner, Eds. 2004. An Introduction to Bioinformatics Algorithms (Computational Molecular Biology), The MIT Press, Boston, MA.

Karlin S and SF Altschul. 1990. Methods for assessing the statistical significance of molecular sequences features by using general scoring schemes. Proc Nat Ac Sci USA, 87, pp. 2264–2268.

Keith JM, P Adams, D Bryant, DP Kroese, KR Mitchelson, DAE Cochran, and GH Lala. 2002. A simulated annealing algorithm for finding consensus sequences. Bioinformatics, 18, pp. 1494–1499.

Kerr M and G Churchill. 2000. Analysis of variance for gene expression microarray data. J Comp Biol, 7, pp. 819–837.

Krogh A, M Brown, IS Mian, K Juolander, and D Haussler. 1994. Hidden Markov models in computational biology applications to protein modeling. J Mol Biol, 235, pp. 1501–1531.

Kumar CG, R LeDuc, G Gong, L Roinishivili, HA Lewin, and L Liu. 2004. ESTIMA, a Tool for EST Management in a Multi-Project Environment. BMC Bioinformatics, 5, p. 176.

Lawrence CE, SF Altschul, MS Boguski, JS Liu, AN Neuwald, and J Wootton. 1993. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. Science, 262, pp. 208–14.

Li C and WH Wong. 2001. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci USA, 98, pp. 31–36.

Li S, DK Pearl, and H Doss. 2000. Phylogenetic tree construction using MCMC. J Am Stat Ass, 95, pp. 493–503.

Li WH. 1997. Molecular Evolution. Sinauer Assoc, Boston, MA.
Lipman JD, SF Altschul, and JD Kececioglu. 1989. A Tool for Multiple Sequence Alignment. Proc Natl Acad Sci USA, 86, pp. 4412–4415.
Liu L, G Gong, Y Liu, S Natarajan, DM Larkin, A Everts-van der Wind, M Rebeiz, and JE Beever. 2004. Multi-species comparative mapping in silico using the COMPASS strategy. Bioinformatics, 20, pp. 148–154.
Ma B, J Tromp, and M Li. 2002. PatternHunter: Faster And More Sensitive Homology Search. Bioinformatics, 18, pp. 440–445.
Mao C, JC Cushman, GD May, and JW Weller. 2003. ESTAP—An automated system for the analysis of EST data. Bioinformatics, 19, pp. 1720–1722.
Mount DW. 2004. Bioinformatics: Sequence and Genome Analysis. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York.
Muckstein U, IL Hofacker, and PF Stadler. 2002. Stochastic pairwise alignments. Bioinformatics, 18, sup. 2, pp. S153–S160.
Myers EW, GG Sutton, AL Delcher, IM Dew, DP Fasulo, et al. 2000. A whole-genome assembly of Drosophila. Science, 287, pp. 2196–2204.
Needleman SB and CD Wunsch. 1970. A General Method Applicable to the Search for Similarities in Amino Acid Sequence of Two Proteins. J Mol Biol, 48, pp. 443–453.
Notredame C, D Higgins, and J Heringa. 2000. T-Coffee: A novel method for multiple sequence alignments. J Mol Bio, 302, pp. 205–217.
O'Brien SJ, M Menotti-Raymond, WJ Murphy, WG Nash, J Wienberg, R Stanyon, NG Copeland, NA Jenkins, JE Womack, and JAM Graves. 1999. The promise of comparative genomics in mammals. Science, 286, pp. 458–481.
Overbeek R, N Larsen, GD Pusch, M D'Souza, E Selkov, N Kyrpides, M Fonstein, N Maltsev, and E Selkov. 2000. WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucleic Acids Res, 28, pp. 123–125.
Paquola AC, MY Nishyiama Jr, EM Reis, AM da Silva, and S Verjovski-Almeida. 2003. EST-Web: bioinformatics services for EST sequencing projects. Bioinformatics, 19, pp. 1587–1588.
Peitsch MC. 1996. ProMod and Swiss-Model: Internet-based tools for automated comparative protein modeling. Biochem Soc Trans, 24, pp. 274–279.
Pevzner PA. 2000. Computational molecular biology, an algorithmic approach. MIT Press, Cambridge, MA.
Rannala B and Z Yang. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. J Mol Evol, 43, pp. 304–311.
Smith TF and MS Waterman. 1981. Identification of Common Molecular Subsequences. J Mol Biol, 147, pp. 195–197.
Smyth GK, YH Yang, and TP Speed. 2003. Statistical issues in microarray data analysis. In: Functional Genomics: Methods and Protocols, MJ Brownstein and AB Khodursky Eds., Methods in Molecular Biology Volume 224, Humana Press, Totowa, NJ, pp. 111–136.
Swofford DL. 2001. PAUP (Phylogenetic analysis using parsimony) V4.0. Available from Sinauer Associates, Boston, MA.
Tamayo P, D Solni, J Mesirov, Q Zhu, S Kitareewan, E Dmitrovsky, ES Lander, and TR Golub. 1999. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA, 96, pp. 2907–2912.
Thompson JD, DG Higgins, and TJ Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res, 22, pp. 4673–4680.
Venter JC, et al. 2001. The sequence of the human genome. Science, 29, pp. 1304–1351.
Wang H, X He, M Band, C Wilson, and L Liu. 2005. A study of inter-lab and inter-platform agreement of DNA microarray data. BMC Genomics, 6, p. 71.
Wang LS, R Jansen, B Moret, L Raubeson, and T Warnow. 2002. Fast phylogenetic methods for the analysis of genome rearrangement data: An empirical study. Proc. of 7th Pacific Symposium on Biocomputing.

Waterston RH, ES Lander, and JE Sulston. 2002. On the sequencing of the human genome. Proc Natl Acad Sci USA, 99, pp. 3712–3716.

Xu Y and CE Uberbacher. 1997. Automated Gene Identification in Large-Scale Genomic Sequences. J. Comp. Biol., 4, pp. 325–338.

Yang YH, S Dudoit, P Luu, DM Lin, V Peng, J Ngai, and TP Speed. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucl Acids Res, 30, p. e15.

Yang YH and N Thorne. 2003. Normalization for Two-color cDNA Microarray Data. Science and Statistics: A Festschrift for Terry Speed, D Goldstein, Eds., IMS Lecture Notes, Monograph Series, Vol. 40, pp. 403–418.

Yang Z and B Rannala. 1997. Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. Mol Biol Evol, 14, pp. 717–724.

Yeh RF, LP Lim, and CB Burge. 2001. Computational inference of homologous gene structures in the human genome. Genome Res, 11, pp. 803–816.

Zhu J, JS Liu, and CE Lawrence. 1998. Bayesian Adaptive Sequence Alignment Algorithms. Bioinformatics, 14, pp. 25–39.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Part 5
Dealing with the Daunting Genomes of Aquaculture Species

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 28
Dealing with Duplicated Genomes of Teleosts

Alan Christoffels

Genome duplication has been postulated as the driving force behind evolutionary novelty in an organism. For example, early vertebrate evolution has been characterized by a minimum of one and possibly two rounds of whole genome duplication. In addition, teleosts have undergone a fish-specific genome duplication after fish and land vertebrates split about 450 million years ago. Understanding the contribution of duplications to the shaping of genomes over 450 Mya requires computational tools that can adequately cope with the plethora of fragmented and assembled genomic data that are available, particularly for teleost genomes. In this chapter techniques for analyzing duplicate genomes are described with respect to (1) the identification of duplicate genes, (2) dating duplicated genes, (3) identifying duplicate segments or paralogons within a sequenced genome, (4) identifying duplicate genes using Expressed Sequence Tag (EST) data, and (5) assessing the functional divergence of duplicate genes.

Background

As early as 1970, Susumu Ohno proposed that two rounds of whole genome duplications occurred during the early stages of vertebrate evolution, based on genome size differences in chordates and evidence of recent tetraploidization in some fish lineages (Ohno 1970). Today, Ohno's hypothesis is more commonly known as the 2R hypothesis that states that one round of duplication occurred at the root of the vertebrate lineage, followed by another around Agnatha and Gnathostomata split (Figure 28.1) (Makalowski 2001). Overwhelming evidence has subsequently mounted in support of the 2R hypothesis in shaping the genomes of vertebrates (Postlethwait et al. 2000, McLysaght et al. 2002, Gu et al. 2002, Panapoulou et al. 2003).

Recently, genome-scale analysis of two teleost genomes, namely, *Fugu rupribes* and *Tetraodon nigroviridis* (Christoffels et al. 2004, Vandepoele et al. 2004, Jallion et al. 2004) have provided confirmation of a third round of genome duplication within teleosts (3R) after ray-finned fish diverged from lob-finned fish about 450 million years ago (Figure 28.1), confirming numerous studies on subsets of duplicate gene families. See references in Van de Peer (2004). In fact, the role of genome duplication via polyploidization has been studied in other eukaryotes including yeast and plants (Kellis et al. 2004, Blanc and Wolfe 2004a, respectively).

These large-scale duplication events, namely 1R, 2R, or 3R, can be hidden in a genome because of exponential loss of duplicate genes in a genome over time (Seioghe and Wolfe 1999; Lynch and Conery 2000, 2003; Wong et al. 2002; Blanc

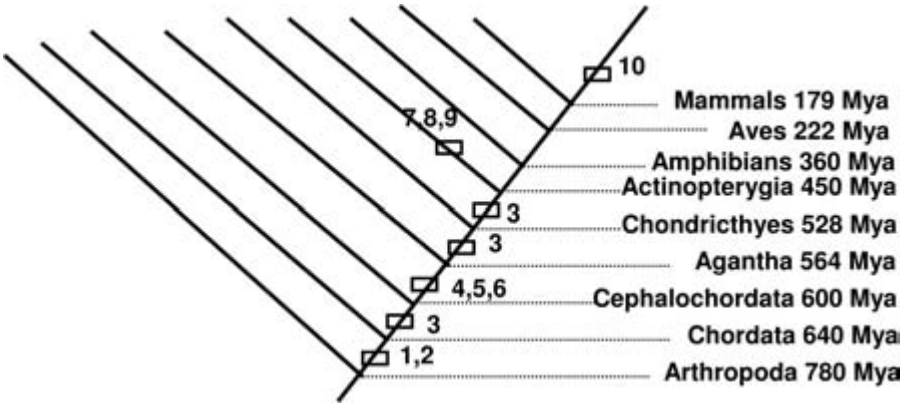


Figure 28.1. Schematic representation of the phylogeny of the main vertebrate groups rooted with arthropods.
Proposed duplication events are indicated with open rectangles. Divergence times are indicated in millions of years ago (Mya). References include (1) McLysaght et al. 2002, (2) Gu et al. 2002, (3) Ohno 1970, (4) Lundin et al. 2003, (5) Holland 2003, (6) Panopoulou et al. 2003, (7) Christoffels el. 2004, (8) Vanderpoele et al. 2004, (9) Jallion et al. 2004, (10) Bailey et al. 2002. Additional references can be obtained within the material listed above.

et al. 2003). In addition, chromosomal rearrangements scramble pieces of duplicated chromosomes around the genome hindering the identification of duplication signatures in a genome (Wolfe 2001, Gu and Huang 2002, Seioghe 2003). Identifying these remnants of duplication evidence has been made possible with the development and maturation of computational tools. Here I will describe methodologies used for the identification of duplicate genes, dating duplicated genes, identifying duplicate segments or paralogons within a sequenced genome, identifying duplicate genes using EST data, and assessing the functional divergence of duplicate genes.

Identification of Duplicate Genes

A phylogenetic tree, hereafter referred to as a 'tree,' captures the evolutionary relationships between genes. Trees provide a reliable method of distinguishing orthologs (genes arising from a common ancestor) from paralogs (genes arising from a duplication event). The tree is composed of branches and nodes (Figure 28.2) where nodes represent the points at which two branches diverge. The branch pattern of a tree is called the tree topology and distinguishes duplication events specific to one species (Figure 28.2a) or shared by multiple species (Figure 28.2b). The root of a phylogenetic tree is its base and implies the branching order of a tree, namely, which sequences share a common ancestor. The tree can only be rooted with an outgroup or an external point of reference (invertebrate branch in Figure 28.2c).

The following strategy is recommended to identify duplicated genes: protein family clustering, multiple sequence alignment, phylogenetic tree construction, and high-throughput screening of tree topologies.

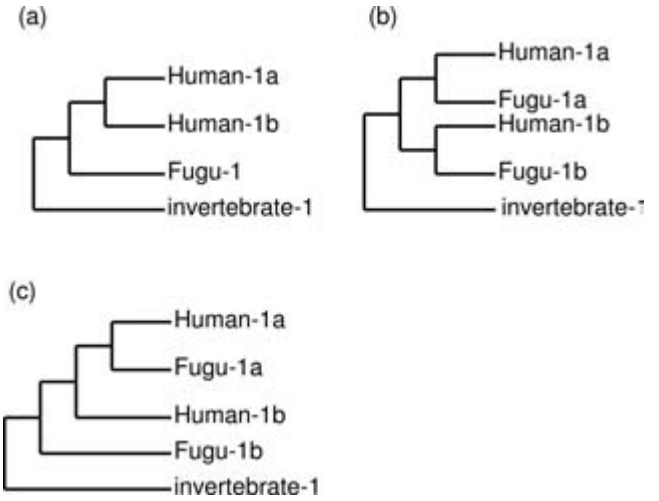


Figure 28.2. Schematic representation of duplication topologies. (a) Gene 1 duplicated in human. (b) Two duplication events occurred before human and fish diverged depicting (A+B)(C+D) topology, also known as the 2+2 topology. (c) Two duplication events in human and Fugu but the two ancestral events cannot be inferred from the tree, possibly due to changes in evolutionary rates.

Protein Family Clustering

A critical consideration before embarking on any of the sections outlined above requires the assignment of putative protein families for the genome under investigation. Homology detection is usually carried out by an all-against-all similarity search within a species using the Basic Local Alignment Search Tool (BLAST) (Altschul et al. 1990). However, before clustering, a number of preprocessing steps are required to "clean" the protein data before clustering. Failure to remove splice variants will result in the detection of false positive duplication events. Annotated genomes such as those present in Ensembl are somewhat easier to filter because alternative splice variants are annotated in the Ensembl database (www.ensembl.org). The splice variants should be reduced to one transcript per gene, and very often the choice of a splice variant depends on the length of the transcript.

Note that an analysis of duplication events should focus on functional genes. To this end, potential pseudo-genes should be removed. In practice, protein-coding genes that do not start with a methionine are removed together with genes annotated as putative pseudo-genes or transposable elements. Subsequent filtering based on substitution rates as a proxy for genes under positive selection can be introduced when comparing paralogs.

All proteins in the final "cleaned" dataset are compared using BLAST. Identifying homologous or paralogous genes using BLAST require certain thresholds. Expectation scores or e-values are often used as a threshold to decide on the relationship between sequences. However, when sequence divergence increases too far, for example identity of 20% between proteins, then the true extent of homology is difficult to interpret. For this reason, additional measures are incorporated, such as

increasing the sequence similarity cut-off with decreasing alignable regions (Rost 1999). Proteins are erroneously included into gene families when short proteins share domains with longer proteins. To overcome this problem, thresholds are chosen based on the fraction of the BLAST query and database hit that is contained in the aligned sequence pair (Li et al. 2001, Christoffels et al. 2004). To avoid the inclusion of multigene families, putative transposable elements are discarded together with sequences that identify more than five matching sequences at an e-value threshold of $1e^{-10}$.

Selecting the appropriate threshold for protein family size can be influenced by the aim of the study. For example, protein families containing vertebrate-specific duplicate genes can be generated using a matching invertebrate gene as a cut-off of protein family assignment (McLysaght et al. 2002). After similarity scores are assigned to homologous genes, a single linkage algorithm is used to group proteins into families. Simply put, if protein A hits protein B and protein B hits protein C then protein A, B, and C are assigned to the same cluster even if protein A and C do not share sufficient identity. Recently a new approach, TRIBE-MCL has been developed to cluster protein sequences and is implemented in Ensembl protein family assignment (Enright et al. 2002).

Identifying duplication events and particularly large-scale events need to be distinguished from tandem duplications, which are single duplication events that create tandem repeats in the genome. As a result, proteins within the same genome with an e-value less than $1e^{-15}$ and separated by at most 20 unrelated genes on a genome segment are classified as tandem duplicates and are removed from the dataset.

The above approach is restricted to identifying paralogs within a species. However, in practice, protein families are often expanded to include orthologs from other species. Orthologs are very often identified using a reciprocal BLAST search. Briefly, genes in one species (S1) are searched against the collection of genes in a second species (S2). An orthologous pair is identified when gene-A in S1 is the best hit for gene-B in S2 and gene-B in S2 is the best hit for gene-A in S1.

Multiple Sequence Alignments

Multiple sequence alignments are generally constructed by a progressive sequence alignment approach such as implemented in Clustal W (Thompson et al. 1994). Insertions or deletions (indels) within a gene can change the reading frame or introduce stop codons. Indels are presented as gaps in an alignment, and the corresponding size of a gap is less important than the fact that the gap is present. These gaps have an undue influence on the resulting tree because there are separate penalties for inserting a gap and for making the gap bigger. In practice, alignments can be adjusted by 'eye' using software such as BioEdit (http://www.mbio.ncsu.edu/BioEdit/bioedit.html). Automated strategies for gap removal in an alignment have proven successful for large-scale analysis (Lynch and Conery 2003, Vandepoele et al. 2004). See the Identifying Duplicates Using EST Data section for a description.

Phylogenetic Tree Reconstruction

A range of phylogenetic methods apply to molecular data and include (1) distance methods, (2) parsimony methods, and (3) likelihood methods, which are packaged

either individually or together in software such as PHYLIP (Felsenstein 1989) or PAML (Yang 1997). Trees are tested with more than one phylogenetic method for consistency before extrapolating any information. Reconstructing trees is hampered by the observation that the true evolutionary difference between two sequences is obscured by multiple mutations at the same site (Graur and Li 2000). These rapidly evolving sites would cause estimates of evolution to be underestimated. However, various models have been developed to estimate the true difference between sequences based on amino acid substitution matrices such as PAM and BLOSUM or gamma correction (Gu and Zhang 1997) where more weight is given to changes at slowly evolving sites.

The resulting trees are usually tested for accuracy using bootstrapping (Felsenstein 1985). Bootstrapping tests whether the whole dataset supports the tree or whether there is a slightly better tree compared to many alternatives. This is done by taking random subsamples of the dataset while allowing repeat sampling of sites. The frequency with which various parts of a tree are reproduced in each of the random samples is calculated and reported as a percentage. Bootstrap values of 70% or higher are likely to indicate reliable groupings (Hills 1994). Another problem encountered with tree accuracy is that of long-branch attraction and refers to the tendency of highly divergent sequences to group together in a tree regardless of their true relationships. Reasons for, and possible solutions to long-branch attraction have been reviewed by Baldauf (2003) and Felsenstein (1985).

High-throughput Screening of Tree Topologies

The large numbers of trees that are generated by a genome scale analysis of duplicated genes require adequate storage in a reusable format. Trees are represented in a computer-readable form using the NEWICK format. The rooted tree in Figure 28.2a can be represented in NEWICK format as follows: (Fly:4.0, (Fugu:2.0, (Human-1a,Human-1b):7.0)) where branch lengths are indicated by numbers preceded by colons, and internal nodes are represented by a pair of matched parentheses. The bioperl open source initiative (Stajich et al. 2002) maintains a collection of PERL wrappers that facilitates automated phylogenetic tree construction for high throughput analysis. Additional PERL routines exist to manipulate and extract data from the NEWICK trees. Many of these routines can be modified to identify species-specific topologies as depicted in Figure 28.2a.

Dating Duplicate Genes

A large proportion of duplicate genes created around the same time is strong evidence that the gene duplicates were created by a single event. A variety of methods can be harnessed to date duplicate genes, which are described in the following sections.

Dating Based on Synonymous Substitution Rates (Ks)

Given that natural selection acts mainly on protein sequences, synonymous codon positions are thought to be free from selection and therefore accumulate changes at a

neutral rate similar to the mutation rate (Li 1993). By assuming that the number of silent substitutions increases approximately linearly with time for two homologous sequences, the relative age distribution of gene duplicates within a genome can be inferred indirectly from the distribution of Ks values (Blanc and Wolfe 2004a).

Time of divergence (T) since duplication between sequences is calculated using the formula:

$$\begin{aligned} T &= Ks/2\lambda \quad where, \\ Ks &= synonymous\ substitutions\ per\ synonymous\ site \\ \lambda &= mean\ rate\ of\ synonymous\ substitution. \end{aligned} \tag{28.1}$$

Synonymous substitutions have been used extensively to estimate when duplication events occurred (Simillion et al. 2002, Blanc et al. 2003). However, one caveat needs to be considered, namely, substitutions used in dating require small values for Ks (low level of sequence divergence). Ks values > 1 for a gene pair means that there are potentially large inaccuracies in divergence estimates due to multiple substitutions at the same site.

The interpretation of a distribution of Ks values can be described by a L-shaped curve because the process of duplications and gene losses occur randomly and at a steady state (Figure 28.3) (Blanc and Wolfe 2004a). An initial peak of duplicates is contained in the younger class of genes. Over time, there is an elimination of duplicates resulting in an exponential decrease of density along with increasing age (Lynch and Conery 2000, 2003; Blanc and Wolfe 2004a). Finally, the distribution tapers off accounting for those duplicate pairs that are evolving under selective constraints (Prince and Pickett 2002). In the case of *Arabidopsis*, a secondary peak would be observed due to a second and possibly a third ploidy event. Similar, Fugu-specific duplications would be separated from ancient vertebrate duplicates.

Calculating the Ks values required to plot the above distribution curve requires a maximum likelihood (ML) procedure in the PAML package (Yang 1997). These rates can be sensitive to rates of transitions and transversions when the sequence divergence is high. Therefore, species-specific estimates of transition/transversion bias are

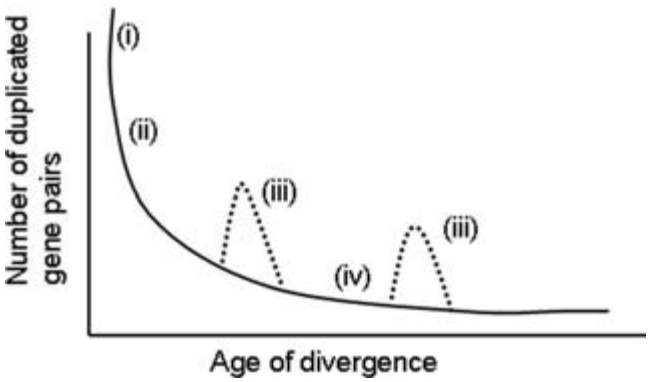


Figure 28.3. Theoretical age distributions of pairs of duplicated genes in a genome. (i) Most recent duplicated genes, (ii) Exponential decrease of duplicated genes due to gene loss, (iii) Secondary peaks corresponding to ancient large-scale duplication events, and (iv) Flattened tail due to older duplicate genes under selective constraints. An expansion of this distribution and supporting data can be obtained from Blanc and Wolfe (2004b).

first obtained from fourfold redundant sites in pairs of sequences that are similar enough to exclude multiple substitutions at the same site.

Dating by Analysis of Phylogenetic Trees

Duplicates that are mapped onto a phylogenetic tree can pinpoint whether the duplication events happened before or after a speciation event. For example, evidence in favor of duplicate genes created after a speciation event (S1) but before a second speciation event (S2) suggest that large-scale duplication events occurred between these two speciation events (Figure 28.1). This relative dating approach has been applied to ray-finned fish (Taylor et al. 2003, Van de Peer et al. 2003, Christoffels et al. 2004).

Genes that were created by two rounds of duplications show a topology as shown in Figure 28.2b, namely [(A+B) (C+D)]. There are many examples where the tree topology does not resemble [(A+B) (C+D)] because phylogenetic tree reconstruction methods are susceptible to genes that show unequal rates of evolution (Figure 28.2c). For example, Friedman and Hughes (2001) showed that 76% of four-member families had topologies as depicted in Figure 28.2c. They concluded that two rounds of genome duplications were unlikely. However, Gibson and Spring (1998) argue that the absence of a 2 + 2 topology could reflect the absence of sufficient phylogenetic signals particularly when two duplication events are separated by 10 million years.

Dating Using Linearized Trees

Linearized trees can be used when the timing of a speciation event is known. Trees are reconstructed under the assumption of equal rates of evolution in different lineages (Takezaki et al. 1995). Relative-rate and branch-length tests for rate heterogeneity are applied to the trees to test for deviations from those assumptions of a constant molecular clock. Slower or faster evolving sequences are removed so that trees are reconstructed using data from sequences that evolve at the same rate. The divergence node for a pair of duplicate genes is compared with a calibration point or speciation event, to infer the age of the paralogous genes (Figure 28.4). This methodology has been

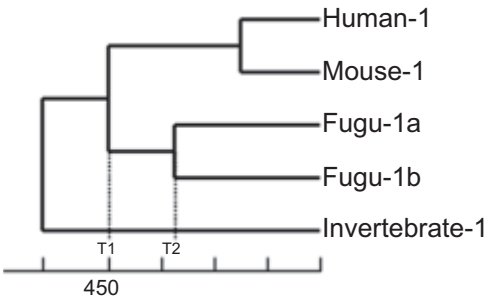


Figure 28.4. Linearized tree displaying a duplicated gene in *Fugu rubripes.* All species in the linearized tree are evolving at the same rate. Divergence of the two Fugu duplicate genes (T1) can be inferred from the land and fish vertebrate split (450 Mya = T2).

recently applied to the dating of Fugu-specific duplicate genes, and clearly showed a large-scale event around 350 million years ago (Christoffels et al. 2004, Vandepoele et al. 2004).

Paralogon Detection

Paralogons represent different chromosomal regions that share a set of paralogs. Specifically, paralogons can be defined according to the number of duplicated genes they contain (*sm*) and the number of unduplicated intervening genes allowed (*d*) (Figure 28.5). An analysis of the arrangement of paralogons can shed light on the type of duplication events that have arisen in a duplicated genome. For example, well-studied paralogous chromosomal segments include the Hox clusters. The existence

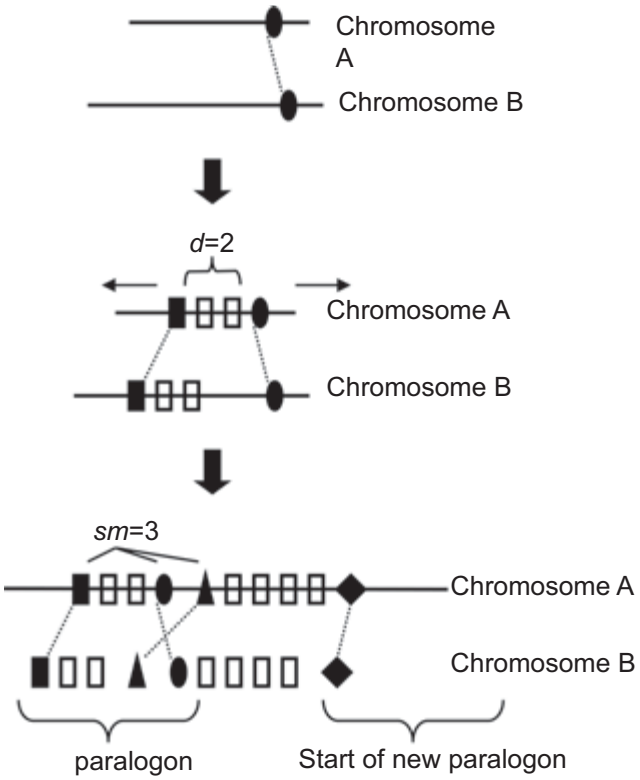


Figure 28.5. Algorithm for identifying paralogons using only gene content. (i) Algorithm starts by identifying a pair of homologous genes on two chromosomal segments that serve as the first anchor point (filled shapes). (ii) Search in both directions from the first anchor point for two more homologous genes that are separated from the first homologous pair by a distance $d=2$ (i.e., two unrelated genes away from the first anchor point). (iii) A paralogon is generated when the distance between homologous genes on the same chromosome segment exceed a threshold $d=2$. The process is then repeated with the search for a new anchor point to mark the start of a new paralogon. In practice d ranges from 10–30 unrelated genes.

of four Hox clusters in mammals and one cluster in invertebrates supports the model that a duplication event occurred before the origin of the vertebrates. Extrapolating these duplication events to a whole genome duplication event require paralogons to be identified throughout the genome. Paralogons should comprise duplicates that are not randomly distributed in the genome. Algorithms to identify paralogons can be divided into those algorithms that consider gene content (McLysaght et al. 2002, Cavalcanti et al. 2003, Christoffels et al. 2004) and those that rely on both gene order and gene content (Vandepoele et al. 2004).

Algorithm Relying on Gene Content Only

The algorithm proceeds as follows: Two homologous genes on different chromosomal segments also referred to as anchor points are identified (Figure 28.5). The segments are searched in both directions for two other homologous genes that are located on the same segments at a specified distance = *d* from the first gene pair (Figure 28.5 where $d = 2$). The second pair of genes is then added to the first to increase the cluster of genes. These clusters are used to pinpoint homologous segments in the genome. A paralogon is delineated when the distance between homologous genes exceeds threshold = *d* (Figure 28.5). Genome segments are sensitive to distances between homologous gene pairs. At least 500 statistically significant paralogons of $sm >= 3$ with *d* ranging from 10–30 were identified in 79% of the human genome (McLysaght et al. 2002, Panapoulou et al. 2003). The even distribution of paralogons in the human genome lends itself to have arisen by whole genome duplication rather than segmental duplications that are found to be preferentially located near telomeres (Bailey et al. 2002). Similarly, 7% of the Fugu genome was covered by Fugu-specific paralogons (Christoffels et al. 2004). The accuracy of paralogons can be enhanced by mapping phylogenetically defined duplicate genes to the paralogons.

Algorithm Relying on Gene Content and Gene Order

This algorithm has been used successfully on the *Arabidopsis* and Fugu genomes (Vandepoele et al. 2002, Vandepoele et al. 2004). The algorithm also starts with the identification of homologous gene pairs close to two different segments in the genome. The order of genes in this segment together with the distance covered by the homologous gene pairs affects the statistical significance of the identified segments (i.e., whether these segments could be identified by chance). The algorithm is as follows:

1. Chromosomes are represented as gene lists, and lists are sorted according to chromosome positions.
2. Within a list, identify homologous genes on other chromosomal segments using similarity scores (identified by BLAST).
3. Homology information is stored in a matrix.
4. Identify collinear regions on the diagonal lines in the matrix.

Identifying the diagonals or paralogons for these algorithms requires consideration of two parameters. First, the intervening genes refer to the number of nonhomologous

genes that separate two anchor points or homologous genes on a chromosomal segment (parameter *d*, Figure 28.5). Segments corresponding to ancient duplications are affected by gene loss and rearrangements resulting in variable distances between homologous chromosomal segments. Second, the probability of finding putative paralogons has to be tested for statistical significance (i.e., can these paralogons appear by chance in a randomized genome?).

The homology matrix (Step 3 above) provides a visual method of identifying useful information. Tandem duplicates appear as horizontal or vertical lines. Inversions are identified on the diagonals with opposite direction, and insertions correspond to gaps in the diagonal regions (result of translocations or gene loss).

Identifying Duplicates Using EST Data

ESTs represent single-pass reads that are fragmented and as a result require cleaning prior to construction of gene indices or EST clusters (Adams et al. 1991). In practice, vector contamination and low quality sequences are removed from ESTs followed by clustering with tools such as TIGR assembler (Quakenbush et al. 2001) or STACKPACK (Miller et al. 1999, Christoffels et al. 2004). A consensus sequence for each EST cluster is used to identify paralogs within a species and orthologs between species.

In the absence of genomic sequences, EST data have been leveraged successfully in identifying gene duplicates and calculating the relative age distribution of these duplicates (Van der Hoeven et al. 2002, Blanc and Wolfe 2004a). Using EST data to calculate the age of duplicate events can result in overestimation of the fraction of recent duplication events due to the fragmented nature of the EST data. However, these relative age distributions provide information on the occurrence and timing of large-scale duplication events without comparing the distributions in a quantitative way (Blanc and Wolfe 2004a).

Paralog identification, using EST sequences, within a species is carried out by an all-against-all nucleotide sequence similarity search using BLASTN. Sequences that are aligned over >300 base pairs (bp) and showing at least 40% identity are defined as pairs of paralogs. Putative orthologs between species A and B are identified using reciprocal BLAST and requiring sequences to be aligned over >300 bp.

Calculating Substitution Rates Using ESTs

The EST consensus sequences contain sequencing errors that can result in frameshifts and these sequences might not have any annotated open reading frames (ORF). Prior to estimating the level of silent substitutions, EST alignments are improved as follows: For each paralogous sequence pair, BLASTX searches are used to find protein hits in GenBank. Significant hits are identified as alignment lengths longer than 99 amino acids with an e-value smaller than $1e^{-15}$. Nucleotide sequences are then translated with GeneWise (Birney et al. 1996) using the protein sequence as a guide. The translations for the two paralogs are then aligned using a Smith-Waterman algorithm with the resulting alignment used as a guide to align the nucleotide sequences. Gaps are removed from the alignment including N-containing codons and the Ks values calculated on the final alignment.

A caveat concerning the use of paralogs derived from ESTs is that multiple entries for the same gene can be present which results in redundant Ks values. ESTs for the same gene could be splice variants, and the corresponding Ks would be zero. In practice, one sequence of a pair of paralogs is discarded when Ks = 0 together with all Ks values attached to this sequence.

Considering the alignment quality is important because of the inherent noise in poorly aligned EST sequences. A number of heuristics must be implemented to filter poorly aligned regions so that larger datasets can be analyzed. These heuristics are based on the principle that positions flanking an aligned gap are potentially of low quality and therefore these flanking positions are removed together with alignment gaps.

The following heuristic can be applied to large-scale analysis of alignments of ESTs. Aligned regions in the BLASTX protein alignments (described above) are scanned until a gap is encountered. Sequence on either side of the gap is discarded until a pair of identical amino acids is encountered. Sequences are retained provided at least one out of six amino acids in the alignment are identical. The final alignments are used to retrieve nucleotide regions that are then aligned.

Assessing the Functional Divergence of Duplicate Genes

A number of possible outcomes exist for duplicate genes that are retained in a genome including (1) new function acquired by a gene copy (i.e., neofunctionalization) or (2) copies retain a subset of the functions of the ancestral gene (i.e., subfunctionalization) or (3) one copy is lost over time (i.e., nonfunctionalization). See Lynch and Conery (2000) and references therein. Quantifying these outcomes can be difficult as genes exert their biological effects in different ways including interactions with other DNA or proteins, expression in different locations for genes with the same function. However, the Gene Ontology (GO) classification scheme provides a method of relating gene functions to loss or retention of duplicate genes. In particular, the GO consortium provides a standardized vocabulary to describe gene functions as defined in three categories namely molecular function, biological process, and cellular compartment (Ashburner et al. 2000). These GO terms can be mapped to genes within an annotated genome to show that a subset of genes (i.e., duplicated genes are not evenly distributed in GO functional categories, suggesting nonrandom loss or retention of duplicated genes that is influenced by the function of the proteins they encode). For example, several regulatory functions and transporter activities were found overrepresented suggesting a mechanism of increasing complexity of regulatory networks (Blanc and Wolfe 2004b).

An alternative method of addressing functional divergence for duplicate genes estimates the rate of asymmetrical protein evolution. In particular, an outgroup is identified for each duplicate pair of genes using reciprocal BLAST as outlined in the Identification of Duplicate Genes section. The triplet sequences are aligned, and the maximum likelihood estimates are computed under two competing models of evolution. The first model allows all sequences to evolve at their rate while the second model restricts sequences to evolving at clock-like rates (i.e., the length of branches leading from one node to the duplicate proteins are forced to be equal). Twice the difference of the log likelihood [$2 \times \mathrm{Ln}_{(\text{no constraint})} - \mathrm{Ln}_{(\text{clock})}$] follows a chi-square distribution with one degree of freedom (Goldman and Yang 1994). A probability is attached to the null hypothesis that the two duplicate protein sequences are evolving at the same rate.

Exploiting the wealth of gene expression data available through microarray analyses, Blanc and Wolfe (2004b) tested whether gene expression diverged between genes that were retained as duplicates. In their study, Blanc and Wolfe analyzed 1,557 expression profiles for *Arabidopsis*. The Pearson correlation coefficient (r) was calculated as a proxy for divergence of gene expression patterns in gene duplicates. In theory, paralogous genes will have a correlation coefficient of 1.0 immediately after a ploidy event. As expression profiles diverge, the r-value decreases. Firstly, r was measured for all duplicate gene pairs. Next, they measured r for 10,000 randomly chosen, nonduplicated genes to determine a statistically significant cut-off point. For the 10,000 randomly chosen gene pairs, 95% of the correlation coefficients were <0.52. Therefore, any pair of duplicate genes with an expression profile of $r<0.52$ was considered to have a diverged expression profile. These diverged expression profiles combined with asymmetrical protein evolution data outlined in the previous paragraph suggests that duplicates have acquired divergent functions.

Conclusion

Techniques to analyze duplicated genomes, as outlined in this chapter, have matured over the years and provide a toolkit for the analysis of sequenced or partially sequenced genomes to better understand the impact of genome duplication in vertebrate evolution. Currently, four teleost genomes have been completely sequenced. However, with some 24,000 species, teleosts represent the largest group of vertebrates and are unlikely to be sampled by many completely sequenced genomes. Leveraging fragmented sequence data such as ESTs will, in the future, become important when trying to understand the impact of duplicated genomes within teleost species as more and more fish genomes are sampled by EST technologies.

References

Adams MD, et al. 1991. Complementary DNA sequencing: Expressed sequence tags and the human genome project. Science, 252, pp. 1651–1656.

Altschul S, W Gish, W Miller, EW Meyers, and DJ Lipman. 1990. Basic local alignment search tool. J Mol Biol, 215, pp. 403–410.

Ashburner M, et al. 2000. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 25, pp. 25–29.

Bailey J, et al. 2002. Recent segmental duplications in the human genome. Science, 297, pp. 1003–1007.

Baldauf SL. 2003. Phylogeny for the faint of heart: a tutorial. Trends in Genet, 19, pp. 345–351.

Birney E, JD Thompson, and TJ Gibson. 1996. PairWise and SearchWise: Finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. Nucleic Acids Res, 24, pp. 2730–2739.

Blanc G, K Hokamp, and KH Wolfe. 2003. A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. Genome Res, 13, pp. 137–144.

Blanc G and KH Wolfe. 2004a. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell, 16, pp. 1667–1678.

Blanc G and KH Wolfe. 2004b. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. The Plant Cell, 16, pp. 1679–1691.

Cavalcanti AR, R Ferreira, Z Gu, and WH Li. 2003. Patterns of gene duplication in *Saccharomyces cerevisiae* and *Caenorhabditis elegans.* J Mol Evol, 56, pp. 28–37.

Christoffels A, EG Koh, J-M Chia, S Brenner, S Aparicio, and B Venkatesh. 2004. *Fugu* genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fish fishes. Mol Biol Evol, 21, pp. 1146–1151.

Enright AJ, S Van Dongen, and CA Ouzounis. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res, 30, pp. 1575–1584.

Felsenstein J. 1985. Confidence limits in phylogenies: An approach using the bootstrap. Evolution, 39, pp. 783–791.

Felsenstein J. 1989. PHYLIP (phylogeny inference package), Version 3.2. Cladistics, 5, pp. 164–166.

Friedman R and AL Hughes. 2001. Gene duplication and the structure of eukaryotic genomes. Genome Res, 11, pp. 373–381.

Gibson TJ and J Spring. 1998. Genetic redundancy in vertebrates: polyploidy and persistence of genes encoding multidomain proteins. Trends Genet, 14, pp. 46–49.

Goldman N and Z Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol, 11, pp. 725–736.

Graur D and WH Li. 2000. Fundamentals of molecular evolution. Sinauer Associates Inc.

Gu X and W Huang. 2002. Testing the parsimony test of duplications: a counterexample. Genome Res, 12, pp. 1–2.

Gu X, Y Wang, and J Gu. 2002. Age distribution of human gene families show significant roles for both large and small scale duplications in vertebrate evolution. Nat. Genet., 31, pp. 205–209.

Gu X and J Zhang. 1997. A simple method for estimating the parameter of substitution rate variation among sites. Mol Biol Evol, 14, pp. 1106–1113.

Hills DM. 1994. Application and accuracy of molecular phylogenies. Science, 264, pp. 671–677.

Holland PW. 2003. More genes in vertebrates? J Struct Funct Genomics, 3, pp. 75–84.

Jallion O, J-M Aury, F Brunet, J-L Petit, N Strange-Thomann, E Mauceli, L Bouneau, C Fischer, C Ozouf-Costaz, A Bernot, S Nicaud, et al. (61 co-authors). 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature, 431, pp. 946–957.

Kellis M, BW Birren, and ES Lander. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae.* Nature, 428, pp. 617–624.

Koch MA, B Hauboid, and T Mitchell-Olds. 2000. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis* and related genera (Brassicaceae). Mol Biol Evol, 17, pp. 1483–1498.

Li WH. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol, 36, pp. 96–99.

Li WH, Z Gu, H Wang, and A Nekrutenko. 2001. Evolutionary analyses of the human genome. Nature, 409, pp. 847–849.

Lundin L, D Larhammar, and F Hallbook. 2003. Numerous groups of chromosomal regional paralogies strongly indicate two genome doublings at the root of the vertebrates. Journal of Struct and Funct Genomics, 3, pp. 53–63.

Lynch M and JS Conery. 2000. The evolutionary fate and consequences of duplicate genes. Science, 290, pp. 1151–1155.

Lynch M and JS Conery. 2003. The evolutionary demography of duplicate genes. J Struct Funct Genomics, 3, pp. 35–44.

Makalowski W. 2001. Are we polyploids? A brief history of one hypothesis. Genome Res, 11, pp. 667–670.

McLysaght A, K Hokamp, and KH Wolfe. 2002. Extensive genomic duplication during early chordate evolution. Nature Genet, 31, pp. 200–204.

Miller RT, AG Christoffels, C Gopalakrishnan, J Burke, AA Ptitsyn, TR Broveak, and WA Hide. 1999. A comprehensive approach to clustering of expressed human gene sequence: The sequence tag alignment and consensus knowledge Base. Genome Res, 9, pp. 1143–1155.

Ohno S. 1970. Evolution by Gene Duplication. Springer, New York.

Panopoulou G, S Hennig, D Groth, A Krause, AJ Poustka, R Herwig, M Vingron, and H Lehrach. 2003. New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. Genome Res, 13, pp. 1056–1066.

Postlethwait JH, IG Woods, P Ngo-Hazelett, YL Yan, PD Kelly, F Chu, H Huang, A Hill-Force, and WS Talbot. 2000. Zebrafish comparative genomics and the origins of vertebrate chromosomes. Genome Res, 10, pp. 1890–1902.

Prince VE and FB Pickett. 2002. Splitting pairs. The diverging fates of duplicated genes. Nat Rev Genet, 3, pp. 827–837.

Quakenbush J, J Cho, D Lee, F Liang, I Holt, S Karamycheva, B Parvizi, G Pertea, R Sultana, and J White. 2001. The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic Acids Res, 29, pp. 159–164.

Rost B. 1999. Twilight zone of protein sequence alignments. Protein Eng, 12, pp. 85–94.

Seioghe C. 2003. Turning the clock back on ancient genome duplication. Curr Opin Genet Devel, 13, pp. 636–643.

Seioghe C and KH Wolfe. 1999. Updated map of duplicated regions in the yeast genome. Gene, 238, pp. 253–261.

Simillion C, K Vandepoele, MCE Van Montagu, M Zabeau, and Y Van de Peer. 2002. The hidden duplication past of Arabidopsis thaliana. Proc Natl Acad Sci, 99, pp. 13627–13632.

Stajich J, et al. 2002. The Bioperl Toolkit: Perl modules for the life science. Genome Res, 12, pp. 1611–1618.

Takezaki N, A Rzhetsky, and M Nei. 1995. Phylogenetic test of the molecular clock and linearized trees. Mol Biol Evol, 12, pp. 823–833.

Taylor JS, I Braasch, T Frickey, A Meyer, and Y Van der Peer. 2003. Genome duplication, a trait shared by 22000 species of ray-finned fish. Genome Res, 13, pp. 382–390.

Thompson JD, DG Higgins, and TJ Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res, 22, pp. 4673–4680.

Van de Peer Y. 2004. Computational approaches to unveiling ancient genome duplications. Nature Rev Genet, 5, pp. 752–763.

Van de Peer Y, J Taylor, and A Meyer. 2003. Are all fishes ancient polyploids? J Struct Funct Genomics, 2, pp. 65–73.

Vandepoel K, Y Saeys, C Simillion, J Raes, and Y Van de Peer. 2002. The automatic detection of homologous regions (ADHoRe) and its application to microcolinearity between *Arabidopsis* and rice. Genome Res, 12, pp. 1792–1801.

Vandepoele K, W De Vos, JS Taylor, A Meyer, and Y Van de Peer. 2004. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. Proc Natl Acad Sci USA, 101, pp. 1638–1643.

Van der Hoeven R, C Ronning, J Giovannoni, G Martin, and S Tanksley. 2002. Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. Pant Cell, 14, pp. 1441–1456.

Wolfe KH. 2001. Yesterday's polyploids and the mystery of diploidization. Nature Rev Genet, 2, pp. 333–341.

Wong S, G Butler, and KH Wolfe. 2002. Gene order evolution and paleopolyploidy in hemiascomycete yeasts. Proc Natl Acad Sci USA, 99, pp. 9272–9277.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Computer Applications in BioSciences, 13, pp. 555–556.

Aquaculture Genome Technologies
Zhanjiang (John) Liu

Chapter 29
Bivalve Genomics: Complications, Challenges, and Future Perspectives

Jason P. Curole and Dennis Hedgecock

"Looked at as a question in natural history, the oyster problem is very simple. The demand has outgrown the natural supply, but it is easy to increase the supply indefinitely by oyster culture, and this is all that is needed."

William K. Brooks on the overharvesting of oysters in Chesapeake Bay, 1905

Introduction

The world's marine fisheries are in severe decline, primarily as a result of overharvesting. As worldwide population expands, stress on resources will only increase; indeed, because of overfishing, some fisheries are now tightly regulated and others have been closed altogether. In addition to the loss of a fishery, severe declines in fishery species can have radiating impacts on ecosystems (Pauly and Maclean 2003). One iconic example of a species that has suffered from overfishing and whose loss has led to significant impacts on its native habitat is that of the eastern oyster, *Crassostrea virginica*, in Chesapeake Bay.

The overwhelming significance of the oyster—economically, ecologically, and as a cultural icon—is highlighted by the devastation of the oyster population in Chesapeake Bay to what is currently estimated as 1% of its original abundance. The collapse of the oyster industry in Maryland and Virginia has led to controversial proposals to introduce a nonnative oyster species as a solution to the environmental and economic crises (National Research Council 2004). The loss of oysters is a loss of their capacity to filter and to help control algal populations in these coastal estuarine waters (Newell 1988, Wetz et al. 2002). Although fishing pressure has decreased from its peak in the early twentieth century, diseases (primarily Dermo and MSX) have maintained an intense pressure on populations, in such a way that abundance has continually declined. The reestablishment of a sustainable, vigorous oyster fishery on the East Coast of the United States may be aided by genomic approaches to understand the devastating effects of diseases and stress.

As early as 1891, Brooks argued for the use of aquaculture as a means to relieve the fishing pressure on the natural populations in Chesapeake Bay (Brooks 1905). Although growth in molluscan aquaculture production has slightly lagged behind that of overall aquaculture growth, in 2002 worldwide molluscan aquaculture accounted for 11.8 million tons of production, second only to freshwater fish, with a value of $10.5 billion (FAO Fisheries Department 2004). Altogether, bivalve aquaculture accounted for 26.1% of total aquatic production (Table 29.1), with other molluscs adding an additional 9.4%. In addition, freshwater molluscs accounted for 633,000

Table 29.1. Aquaculture production of various categories of molluscs as reported by FAO Fisheries Department.

Species	Production (million tonnes)	Proportion of total (%)	Annual growth (APR)
Oysters	4.32	10.8	3.9
Misc. marine molluscs	3.74	9.4	14.3
Clams, cockles, arkshells	3.43	8.6	14.1
Mussels	1.44	3.6	2.7
Scallops, pectens	1.23	3.1	3.1
Freshwater molluscs	0.013	0.03	14.6

tons of inland capture fisheries, nearly equivalent to that of cichlids (including tilapia) and freshwater crustaceans. Clearly there is a substantial economic incentive for increasing molluscan aquaculture production.

In the paragraph immediately following Brooks' simple answer to the oyster problem, he suggests that, as a practical matter, increasing oyster production is anything but simple. Despite this difficulty, aquaculture production of oysters has clearly grown. The Pacific oyster *Crassostrea gigas* has the highest annual worldwide production of any marine or freshwater species, at about 4.2 million tons in 2002. As worldwide capture fisheries plateau (or decline), growth in aquaculture is necessary if worldwide seafood demands are to be met without destroying fisheries stocks (FAO Fisheries Department 2004).

Aquaculture and restoration are two of the primary motivating factors underlying the use of genomics for the oyster model. As additional bivalve fisheries grow, they will face the same questions and difficulties faced by the oyster community. Indeed, the market for geoduck (*Panope abrupta*) has recently and rapidly grown (Davis 2006), but genetic and genomic resources for this species lag. In this chapter, we will review past and present genomic efforts and issues unique to bivalves and make suggestions for future researchers to avoid pitfalls.

Complications as a Result of Extremely High Fecundity

The enormous fecundity (10^6–10^8 eggs per female per season) and high larval mortality of most marine animals, including bivalves, makes them fundamentally different from the more familiar and well examined animal models (10–10,000 eggs per female, lifetime). Thirty years ago, Williams (1975) argued in his Elm-Oyster model that sexual reproduction and genetic diversity are favored to a much greater degree in high fecundity species than in low fecundity species. Species with high fecundity likely generate more mutations than low fecundity species, because of the large number of cell divisions required to produce millions or billions of gametes (*cf.* to the argument for male-driven evolution in humans) (Li et al. 2002). The consequences of high fecundity were observed in the earliest genetic studies of bivalve populations, which discovered high levels of protein polymorphism, heterozygote deficiency, and correlations of individual heterozygosity with fitness-related traits, such as size-at-age. These unique biological phenomenon remain at the heart of bivalve genetics, with recessive

lethals leading to distortion of Mendelian ratios in controlled crosses and a high frequency of null alleles leading to inaccurate genotyping.

Challenges as a Result of High Rate of Polymorphism

Protein polymorphism for bivalves was long ago found to be among the highest for animals: average allozyme heterozygosity is greater than 20%, three to four times the mammalian average (Buroker et al. 1979, Fujio 1982, Hedgecock and Sly 1990). The complications of high genetic polymorphism manifest as nonreactive, null allozyme alleles (Mallet et al. 1985, Foltz 1986), and abundant nonamplifying, PCR-null alleles for many DNA markers (Hu et al. 1993, McGoldrick et al. 2000, Vadopalas and Bentzen 2000, Launey and Hedgecock 2001, Reece et al. 2001, Sekino et al. 2003, Reece et al. 2004, Astanei et al. 2005, Sato et al. 2005). In the Pacific oyster, null alleles are segregating at over half of the approximately 100 loci tested in mapping families, even though microsatellites and families were from the same natural population (Li et al. 2003, Hedgecock et al. 2004a). In addition, cross species transfer of DNA markers is limited, as cross-specific polymerase chain reaction (PCR) amplification decays rapidly with evolutionary distance, so that only one in eight Pacific oyster markers amplifies from the eastern oyster, which diverged >5 million years ago. This decay in cross-specific amplification exceeds that observed across genera or even families of vertebrates (Schlötterer et al. 1991, Garza et al. 1995, Pépin et al. 1995, Fitzsimmons et al. 1995, Rico et al. 1996) or species of *Drosophila* (Harr et al. 1998, Colson et al. 1999, Noor et al. 2001, Huttunen and Schötterer 2002), suggesting rapid rates of sequence evolution in PCR primer binding sites.

In the context of controlled crosses, null alleles are a minor complication. Nonamplifying alleles are often identified when the expected Mendelian ratio (based on parental genotypes) does not fit with the predicted genotypes of the offspring from a controlled cross. A simple and common case is the progeny of AB × Aø parents. The parents present as an AB heterozygote and an AA homozygote, and, as expected, one-half of the progeny are phenotypically A (AA or Aø); however, only one-quarter are AB and the remaining quarter show only the B allele indicating that a null allele must be segregating. In contrast, null alleles are a significant issue for population genetic work in bivalves, because failure to identify null heterozygotes (e.g., the Aø heterozygote mentioned above) leads to inaccurate genotyping and an, often significant, excess of homozygous classes. Thus, they will likely be an issue for natural population association studies that are common for model organisms. Whether single nucleotide polymorphisms (SNP) present in coding loci will prove more robust is unclear at this time, but many oyster biologists hope that SNPs will overcome the uncertainties of null alleles.

Inbred Lines, Phenotypes and Pedigree Analysis

Controlled crosses (particularly using inbred lines) provide an opportunity to manipulate the genotype and, combined with genomic approaches, provide powerful means for dissecting complex, multifactorial phenotypes, such as disease and stress resistance, growth, and survival. Disease resistance is of great interest in restoration of the

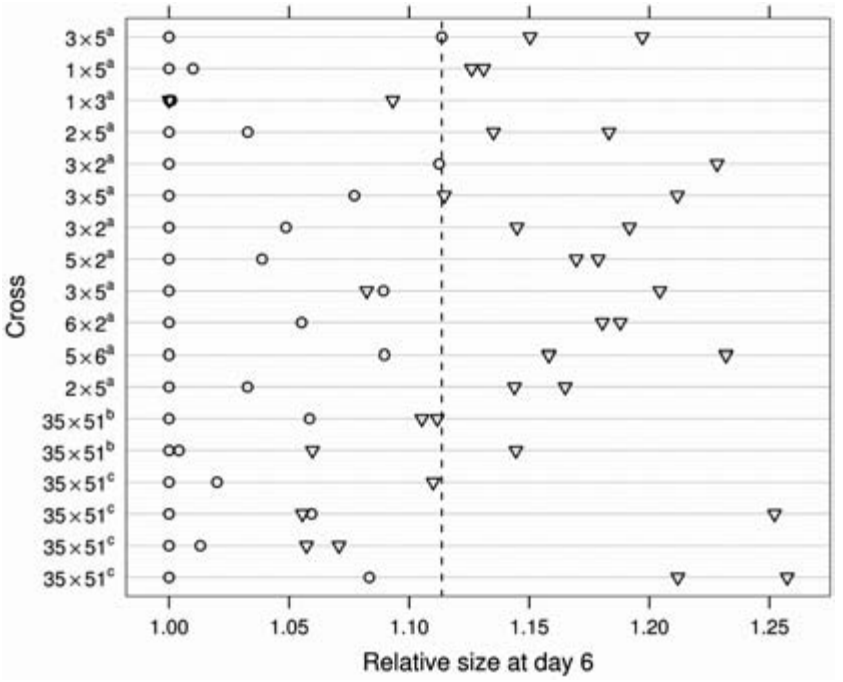


Figure 29.1. Size variation in inbred and hybrid *C. gigas* larvae. Upper graph: Size at day 6 for 18 crosses relative to the slowest growing family in each cross. ∇-hybrid larvae, o-inbred larvae. Dashed line represents the fastest growing inbred family across all crosses. Data from Pace et al. (2006); Curole unpublished data; Meyer and Curole unpublished data.

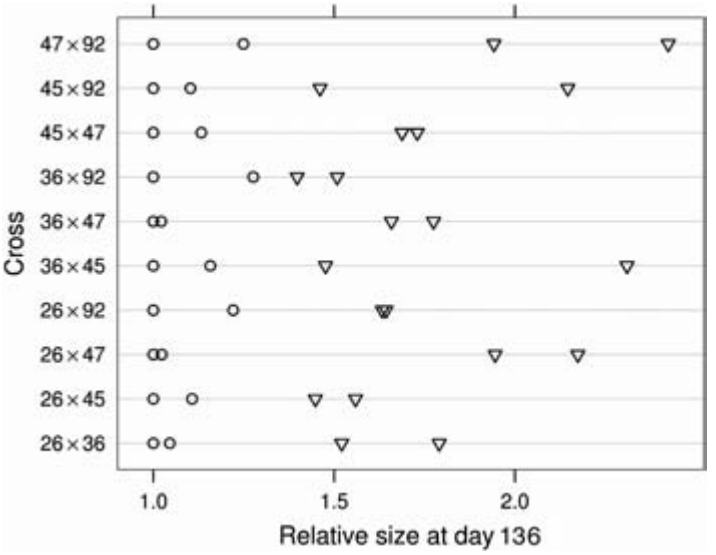


Figure 29.2. Size at day 136 post-fertilization for 10 families of inbred and hybrid *C. gigas* relative to the slowest growing family for in cross. ∇-hybrids, o-inbreds. Data from Hedgecock and Davis (2006).

American oyster; stress resistance, growth, and survival are of greatest interest in the Pacific oyster, the primary bivalve aquaculture species.

Growth heterosis, or hybrid vigor, in bivalve molluscs has a long and storied history, founded primarily on early observations of correlations between allozyme heterozygosity and particular fitness-related traits in wild-caught populations (Zouros and Foltz 1987, Britten 1996). Size-at-age is the most commonly used metric, with significant correlations between heterozygosity and shell length or gross weight at a particular time point in the life cycle (Fujio 1982, Gaffney and Scott 1984, Koehn and Gaffney 1984, Foltz and Chatry 1986, McAndrew et al. 1986, Koehn et al. 1988, Zouros et al. 1988, Gosling 1989, Gaffney 1990, Pogson and Zouros 1994, Zouros and Pogson 1994). The Pacific oyster provides an animal model for studying heterosis, a phenomenon more evident in plants and underlying the improvement of most crops (Gowen 1952, Crow 1998); however, the fundamental genetic cause and underlying mechanisms of heterosis have eluded researchers for decades (Birchler et al. 2003). An experimental genomic approach now offers the possibility to understand this phenotype and harness this understanding to increase production.

In an effort to establish a breeding program, investigators have developed approximately 50 inbred lines for experimental crosses from the naturalized population of *C. gigas* in Dabob Bay, Washington (Hedgecock 1994), using self and brother-sister mating (Hedgecock et al. 1995). The third inbred generation (G_3) of inbred lines was propagated in 2004 to produce the G_4, which has an expected inbreeding coefficient of 0.59. Factorial crosses of inbred lines showed that oysters exhibit heterosis for both growth and survival (yield), experimentally validating the aforementioned years of work associating growth with heterozygosity (Hedgecock et al. 1995). Heterosis for yield and size-at-age has been observed repeatedly, across generations and inbred lines (Figures 29.1 and 29.2) (Hedgecock and Davis 2006). In addition, 24 of the 34 hybrids (71%) presented in Figure 29.1 show a greater relative size at day 6 than the largest inbred across all crosses. Controlled crosses have revealed other similarities with agricultural crops (in particular, maize). Notably, hybrid populations show significantly less variance in size (analysis of variance [ANOVA] of standard deviations at day 6, $F = 23.2$, $P = 5.8 \times 10^{-6}$; Figure 29.3), which means that for larvae a majority of the population is simultaneously reaching the settlement stage, therefore maximizing yield at seed-set. Controlled crosses have also facilitated the elucidation of the physiological mechanisms of heterosis (Hedgecock et al. 1996, Pace et al. 2006), an aspect of this phenotype that has eluded researchers.

Crossbreeding inbred lines of oysters to produce hybrids holds great promise for increasing the yields of farmed Pacific oysters (Hedgecock and Davis 2000). Factorial crosses among inbred lines produce F_1 hybrids for yield traits (progress reports at http://www.hmsc.orst.edu/projects/wrac) (Hedgecock and Davis 2006), and the most promising crosses are then reproduced at a commercial scale. Hybrids are currently in commercial production on the U.S. West Coast (JP Davis, Taylor Resources, personal communication).

Disease-resistant selected strains have been developed for the eastern oyster. Several strains are available (National Research Council 2004), including the DEBY strain, which has been deployed in the Chesapeake. Many of these strains show significant increases in survival after disease challenge, with the greatest increase following the first generation of selection. In field trials, the DEBY strain shows anywhere from one-quarter to one-half the mortality level of local controls (Ragone Calvo et al.

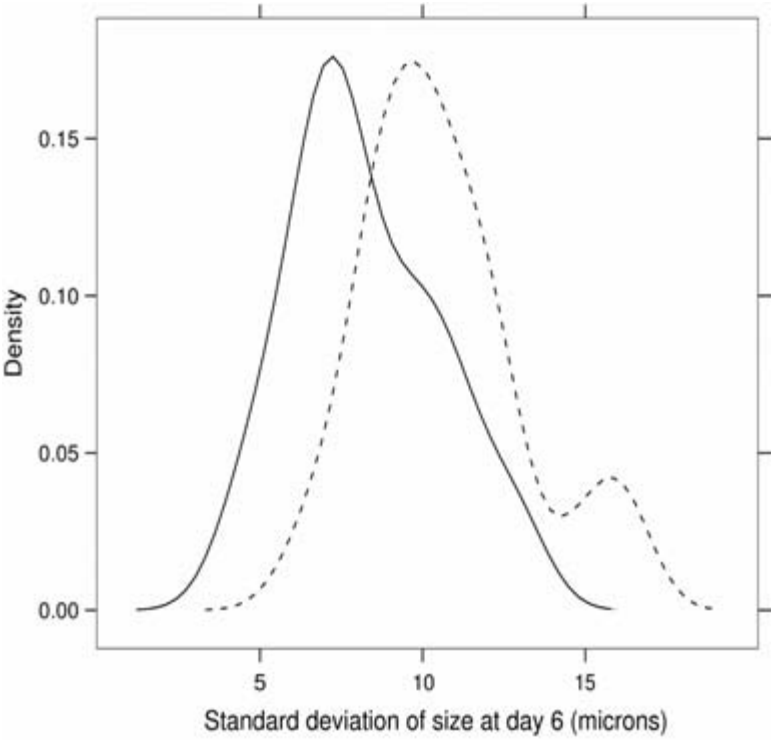


Figure 29.3. Density estimate of standard deviation of size-at-age for crosses presented in Figure 29.1. Solid line-hybrids, dashed line-inbreds. Standard deviations were subjected to kernel density estimation using the density function in the statistical program R (R Development Core Team 2006).

2003). Strains are generally maintained through a mass spawn of selected individuals as opposed to pairwise crosses. This approach is haphazard and can lead to the unintentional loss of genetic diversity in these lines (Hedgecock et al. 1992), which may ultimately result in inbreeding depression.

The greatest difficulty in dealing with inbred oyster lines or selected broodstock is maintaining separate populations. Although great effort is taken to keep lines separate by growing the animals in different tanks, wellers, and bags, cross contamination is a fact (Mallet et al. 1985, Foltz 1986, Zouros et al. 1994, Li and Guo 2004). Over 4 years of genotyping for WRAC broodstock, an average of 14.7% of broodstock animals were excluded from crosses because they were either verified contaminants or insufficient genotype information was available to determine provenance (Table 29.2). Based on these data, it should be clear that an unlucky choice of animals could result in crossing of full-sibs when a hybrid cross is intended, leading to a significantly reduced yield. Even worse, contamination of a line or broodstock is a possible outcome; this has serious implications for a broodstock program, especially if lines have been specifically bred for disease resistance as they are for the eastern oyster. The solution to this issue is broodstock genotyping, a must for any hatchery breeding program. Ideally, a set of highly variable hatchery markers is established, genotyping methods are standardized, and tissues are preserved for all broodstock. Prior to

Table 29.2. Broodstock contamination. Contaminants have genotypes at several loci that are incompatible with parent or control genotype. Uncertain are individuals with insufficient genotype information to determine provenance. Six microsatellite markers were used in 2001 and 2002. Eleven microsatellite markers were used in years thereafter.

Year	Tested	No. of contaminants	Uncertain
2001	164	8	1
2002	203	11	11
2003	339	37	53
2004	272	23	0
Total	978	79	65
Percent rejected		8.1%	6.6%

spawning, potential broodstock candidates are genotyped to validate their pedigree and only spawned after validation. As a rule of thumb, 10–12 individuals per line should be screened and genotyped to guarantee at least one ripe male and female. The importance of pedigree validation cannot be overstated.

Linkage Maps

One of the most fundamental genomic approaches is the generation of a genetic linkage map, as described in Chapter 10, because linkage maps are ideal for identifying quantitative trait loci (see Chapter 11). Prior to the advent of DNA markers, allozymes were used in the hope that an allozyme locus might be linked with a particular phenotype of interest. Minimal linkage groups were generated from the handful of allozymes that co-segregated; full linkage maps were not feasible with allozymes because of the paucity of loci. Early efforts with allozyme loci in crosses of wild bivalves discovered significant distortions of Mendelian segregation ratios. These distortions are the result of a high genetic load at least in the Pacific oyster, but likely in other bivalves as well with a minimum of 15–20 recessive lethal mutations per oyster, about 5 times the genetic load of a human or fruit fly (Launey and Hedgecock 2001, Bucklin 2002). A high genetic load resolves two aforementioned issues in bivalve genetics: distortions of Mendelian inheritance ratios in lab-reared progeny of wild parents (Wada 1975, Beaumont et al. 1983, Foltz 1986, Hu et al. 1993, McGoldrick and Hedgecock, 1997, McGoldrick et al. 2000, Reece et al. 2004) and the correlation of heterozygosity with fitness related measures in natural populations. These phenomena are uncommon or nonexistent in terrestrial animals (Houle 1989, Britten 1996). High mutational load further accounts for severe inbreeding depression and its converse, hybrid vigor (heterosis), in experimental crosses (Lannan 1980; Hedgecock et al. 1995, 1996; Evans et al. 2003).

Distortion of Mendelian ratios (the result of recessive lethals) at individual loci does not significantly affect linkage mapping, but in combination, two linked loci that affect viability can cause difficulties (see Hackett and Broadfoot 2003 for details). Linked lethal recessives have been observed in the Pacific oyster (Bucklin 2002), but are less frequent than unlinked lethals. Despite these potential complications, framework linkage maps of more than 100 microsatellite DNA markers have been

published for the Pacific oyster (Li et al. 2003, Hubert and Hedgecock 2004). Consensus maps have 10 linkage groups, in accord with the haploid chromosome number, cover 70–80% of the Pacific oyster genome, and have marker densities so that the expected distance of a new gene to the nearest marker on a map is 4 to 6 map units (centiMorgans [cM]). Consistent with the high levels of genetic polymorphism, linkage maps for the Pacific oyster show significant differences among families in recombination rate and, more significantly, gene order. These differences suggest polymorphism for the distance between markers as well as for chromosomal rearrangements in the Pacific oyster (Hubert and Hedgecock 2004). An advantage of maps from multiple families is the ability to map a locus by synteny (e.g., if marker x is between markers y and z in family 1, then one can infer that it is between markers y and z in other families where marker x is not segregating); however, because of the polymorphism in marker location, one should be cautious in inferring the location of markers across families.

Amplified fragment length polymorphism (AFLP) maps of 341 markers (Hedgecock et al. 2004b) and 119 markers have also been generated (Li and Guo 2004). Because of the high levels of Mendelian distortion in the Pacific oyster, AFLPs are not as useful as in species with limited distortion (Li and Guo 2004). An AFLP map with three microsatellites and one Type I marker is also available for the eastern oyster (Yu and Guo 2003). This map has good coverage because of the use of AFLPs, and revealed lower levels of segregation distortion than in the Pacific oyster. Unfortunately, AFLPs have major weaknesses because they are anonymous dominant markers that are specific to a cross and, because of their anonymous nature, are nontransferable to other crosses. Each cross also produces two maps, one for each parent, which can be difficult to reconcile. The dominant nature of these markers also makes it impossible to estimate genetic effects at quantitative trait loci, a significant complication if the goal is understanding complex production traits. Despite the relative ease of generation of a large number of AFLP markers (compared with microsatellites), codominant microsatellite loci are superior for map construction.

Tissue Preservation

Preservation of tissue has emerged as an important issue, because genomic work requires the use of intact, good-quality DNA. Adult tissues are easily preserved in ethanol for later DNA extraction, but it is best to perform high-quality DNA extractions as soon as possible because the tissue degrades and loses structure over time. Although more expensive, commercially available kits produce high-quality DNA extractions that are stable over time; this cost may be justified if the sample is important (e.g., broodstock parents or mapping family tissues).

Larvae and spat present a much greater obstacle. If placed directly in ethanol, larvae and spat will simply close their valves and undergo anaerobic metabolism, essentially digesting themselves until death, leaving poor quality tissue and DNA. This effect is minimized to some extent if larvae are killed immediately by adding formalin or bleach directly to seawater containing swimming larvae and then preserved in ethanol. Again, it is best to extract DNA from larvae as soon as possible, because tissue degradation likely begins immediately after preservation, and after time,

extractions produce little to no DNA. We have had the greatest success by killing larvae and then immediately aliquoting individual larvae into extraction buffer in a 96-well plate and preserving at −80°C. Plates prepared in this manner show greater PCR amplification success.

QTL Mapping of Performance Traits

By consolidating mapping efforts, the oyster community has initiated the establishment of F_2 families as reference material for genotyping. As indicated above, the order and location of markers may vary across families, so that a map should be made for each family. The availability of DNA and a linkage map for a reference family, onto which loci can be mapped, should reduce the number of orphan loci (loci that have been genotyped but can not be placed on a linkage group).

With the development of genetic linkage maps, mapping of quantitative trait locus has moved forward. QTL have been identified for three phenotypes in the Pacific oyster: shell shape, shell color, and growth. A single, additive quantitative trait locus (QTL) explains 32% of the variance in pigment saturation of the left valve (Hedgecock et al. 2006a), supporting the observation that shell and mantle edge color are heritable in *C. gigas* (Brake et al. 2004). Two QTL, one of which also explains 32% of the variance, were identified for a left bend in the dorsal-ventral growth axis (Hedgecock et al. 2006a). Four QTL residing on three linkage groups have been identified for initial size, growth curve inflection, and growth rate (Hedgecock et al. 2004b). Dominance appears to play a role in growth, as it does in survival via recessive lethals, with all growth QTL showing dominance. There is also some evidence for epistasis between QTL on two different linkage groups. At least two significant QTL for growth map to linkage groups with significant Mendelian distortion. Although Mendelian distortion does not appear to impact the ability to identify QTL, separating the effects of a QTL for the phenotype of interest from the effects of recessive lethals may prove a formidable task. At loci tightly linked to the recessive lethal, the frequency of one homozygote class is significantly less than the other genotypic classes (or possibly completely absent), reducing the power to detect differences between genotypic classes. Although it might be reasonable to expect that the recessive lethal is the locus responsible for differences in growth, proving this may be very difficult. Additionally, this is probably not a reasonable expectation for other phenotypes, such as shell color or sex.

Disease-resistance QTL have also been identified for the eastern oyster (Yu and Guo 2006). At this time, phenotyping for disease resistance involves genotyping a subset of animals prior to exposure, exposing them to disease in the field, and genotyping a subset of survivors. Deviations from Mendelian ratios before and after disease exposure are then used to identify putative disease-resistance QTL. Using this approach, 12 QTL across two families were identified. MSX infections were light for these families, but Dermo infections were heavy; as such, it is likely that these are disease-resistance QTL for Dermo. A majority of these QTL were identified in a cross between a wild individual and a selected male from the Rutgers NEH stock. In contrast, only two QTL were identified in a cross between parents both from the NEH stock, suggesting that most disease-resistant QTL are fixed in this line. This possibility is disconcerting

given that the NEH × NEH family experienced 53% mortality. Lab-based challenges would be ideal, as well as the genotyping of all individuals prior to field exposure and monitoring of individual mortality.

Cytological Maps

Aquaculture bivalve species generally have low chromosome numbers, with *Mytilus* species having 14 pairs of chromosomes (Thiriot-Quiéveux and Ayraud 1982, Moynihan and Mahon 1983) and cupped oysters having 10 pairs of metacentric or submetacentric chromosomes (Ahmed and Sparks 1967, Longwell et al. 1967, Leitão et al. 1999). In *C. gigas* there are 1.0–1.3 chiasmata per chromosome, implying a genetic map length of 500–650 cM (X Guo, personal communication). Chromosomal banding and FISH techniques with P1 clones, ribosomal RNA (rRNA) genes and repetitive sequences have recently been applied to chromosome identification and mapping (Wang et al. 2001). A repeat that accounts for 1–4% of the genome has been isolated and mapped to centromeric regions of several chromosomes in the Pacific oyster (Clabby et al. 1996, Wang et al. 2001). Major rRNA genes have been mapped to 10q in the Pacific oyster and 2p in the eastern oyster (Xu et al. 2001, Wang and Guo 2004).

Physical Maps

Bacterial artificial chromosome (BAC) libraries have been constructed by the Clemson University Genomics Institute and are publicly available. These are deep coverage libraries (12× and 10× coverage for the Pacific and eastern oysters, respectively), with average insert sizes of 134 kilobase (kb) and 150 kb, respectively. The libraries were constructed from sperm cells, which, in the case of *C. gigas,* were taken from two F_1 hybrid males (Cunningham et al. 2006b).

Assays of gene copy number for *C. gigas* suggest the possibility of gene duplication in this species (Cunningham et al. 2006b). Estimates of gene copy number for *C. virginica* indicate one to two copies in the BAC library. In contrast, copy numbers of four to five were estimated for two genes in the *C. gigas* library. Whether the high estimates for copy number are a sampling artifact or truly represent duplications is of great interest, because gene duplications can complicate genomic efforts.

With these resources in hand, the oyster community is poised to generate a physical map. Efforts are under way to secure funding for BAC fingerprinting and the generation of a physical map. In addition, a large number of ESTs that will be sequenced by the Department of Energy's Joint Genome Institute (JGI) (see EST libraries below) will be from the same inbred lines crossed to make the hybrid males whose DNA was represented in the BAC library. These lines have also been the subject of extensive genetic mapping work. JGI is also sequencing approximately 60 clones from these BAC libraries, which hybridized to probes for eight selected genes (Cunningham et al. 2006a). The BAC sequencing effort will also quantify single nucleotide and gap nucleotide polymorphism, which could interfere with assembly of whole genome shotgun sequences.

Expressed Sequence Libraries: Gene and SNP Discovery

Cloning and sequencing of ESTs are in progress, with the ultimate goal being gene discovery. Two large-scale projects are being performed by the JGI. The larger of these is the agreement by the JGI to sequence 150,000 complementary DNAs (cDNA) from several *C. gigas* libraries. This sequencing effort will produce a wealth of data, including SNPs for at least a couple of inbred lines. One technical issue regarding the construction of the library is the presence of mitochondrial transcripts. Transcriptomic analysis (see below) revealed that the mitochondrial, large ribosomal subunit accounts for 25% of transcripts in 6-day larvae, although this may change with life stage (Anisimov 2005). Removing these transcripts prior to library construction is important for reducing the redundancy of the library. The smaller-scale effort is the sequencing of 25,000 clones from *Mytilus* cDNA libraries. This is directed at identifying genes important in thermal tolerance.

On a smaller scale, pilot EST collection programs for the Pacific and eastern oysters using hemocyte and embryo cDNA libraries from *C. virginica* and a hemocyte cDNA library from *C. gigas* have been completed (Jenny et al. 2002, Gueguen et al. 2003, Tanguy et al. 2004). Although EST collections are small, the gene discovery rate is excellent for genes important in stress and immune challenge function. In addition to the traditional oyster EST collections described, a library of 4.6 million Pacific oyster ESTs is available from the Massively Parallel Signature Sequence (MPSS) comparative transcriptomic analysis described below. A subset of these ESTs has been sequenced to identify candidate loci for heterosis (Meyer 2006).

Several labs are also working to identify SNPs in a subset of previously sequenced loci. In particular, SNP discovery is being done for a subset of loci in both the Pacific and eastern oysters for the purpose of comparative mapping, which has thus far been limited by poor cross-amplification of microsatellite loci (see the Challenges as a Result of High Rate of Polymorphism section). An additional effort is taking place in the Pacific oyster to identify SNPs for candidate loci identified in the transcriptomic analysis of hybrid vigor (see below). SNP frequencies are relatively high (1 every 50 base pairs [bp]), and gap nucleotide frequencies are very high (1 every 33 bp) (Curole and Hedgecock 2005).

Transcriptomics

Transcriptome analyses have rapidly grown in recent years, and expression analyses of bivalves have been propelled by this explosion in growth. The early genetic and physiological work on hybrid vigor in the Pacific oyster attracted collaboration with Lynx Therapeutics, Inc. (now Solexa), which provided MPSS profiles of gene expression in inbred and hybrid larval oysters. These profiles quantify genomic expression with great depth, to the equivalent of a few messenger RNA (mRNA) molecules per cell, for all expressed genes simultaneously (Jongeneel et al. 2003). Expression analysis by MPSS produced 4.6 million 17-bp sequences, comprising 23,275 unique signatures whose expression is greater than 3 transcripts per million. Expression less than this level is not significantly different from zero (Hedgecock et al. 2006b). Statistical contrasts among genotypes of MPSS expression data identified approximatley 350 candidate heterosis

genes for further genetic or functional analysis. Observed patterns of gene expression are more complex than predicted from the classical dominance and overdominance explanations of heterosis (Gowen 1952, Crow 1998), in that hybrids show dominance for low expression and even underexpression. These expression patterns are consistent, on the other hand, with the metabolic efficiency hypothesis for growth heterosis (e.g., reduced rates of protein turnover in hybrid compared to inbred oysters) (Hawkins et al. 1986, Hedgecock et al. 1996, Meyer 2006). These results are similar to work in maize, which has identified over- and underdominant expression in hybrids relative to inbreds (Swanson-Wagner et al. 2006).

In addition to the MPSS work, the Oyster Microarray Consortium has constructed a microarray with *C. gigas* and *C. virginica* ESTs (Jenny et al. 2006). The microarray includes duplicate spots for 4,460 *C. virginica* clones and 2,320 *C. gigas* clones, as well as 188 various control ESTs, with two arrays per slide (for 27,496 features per slide). Approximately 70% of the sequences show no significant match to the Gene Ontology database; of the remaining 30%, approximately 70% are annotated as biological processes, 7% as molecular function, and approximately 20% as cellular component for the two *Crassostrea* species (Jenny et al. 2006). The availability of a standardized microarray platform should significantly advance transcriptomic work. Disease-resistance efforts will likely receive a large boost, as many of the clones are derived from libraries made with disease-challenged animals. One note of caution is necessary; although microarrays have the ability to generate large amounts of data relatively easily, their use for bivalves must be evaluated carefully. The high levels of DNA polymorphism, in particular indels, may lead to biases in hybridization signal. Tempering this caution is the observation that indels are less frequent in coding regions (Curole unpublished data).

Linking Candidate Genes to QTL

Several factors have motivated the identification of SNPs in coding loci and their use in genetic mapping. High levels of marker polymorphism are ideal for intraspecific work because of a greater number of segregating loci, but the breakdown in interspecific cross-amplification of highly polymorphic loci has hindered comparative genomic efforts. The importance of comparative genomic approaches has been exemplified by the success of the vertebrate community, where syntenic maps facilitate the identification of QTL candidate genes for nonmodel organisms. With just a few coding loci mapped, one can quickly compare model and nonmodel organisms, identify the region of synteny, and then scan for possible candidate genes. Lastly, the identification of candidate genes for several phenotypes of interest has motivated an effort to genetically map these loci to test candidacy.

Efforts are under way to map factors responsible for growth in the Pacific oyster and disease in the eastern oyster (see the QTL Mapping of Performance Traits section above). In addition to mapping microsatellite loci to localize QTL, the goal of one SNP discovery effort is testing the candidates for linkage to QTL. Approximately half of the candidates for hybrid vigor that have been sequenced have SNPs segregating between the two lines crossed to generate the F_2 mapping family, despite the relatively short sequences (generally 300 bp). Thus, with short sequences one can expect

that half of the loci will have SNPs segregating in any particular family. Seven candidates have been genotyped in the F_2 population, two of which show linkage to QTL for growth (Curole and Hedgecock 2006). Using these preliminary data, approximately 100 (28%) of the MPSS candidates will require additional evaluation as possible genetic contributors to the heterosis phenotype.

Tools for the Future

Advances in the genomics field are occurring so rapidly that it is likely any specifics given in this section will be out of date by the time it is published (or shortly thereafter). With this in mind, we focus our discussion on principles that are likely not to be out of date. Primary among these is the development of SNPs, because these represent the most fundamental level of genetic polymorphism in an organism. (See Chapter 6.) SNP discovery is set to take a large step forward with the extensive EST sequencing that is ongoing for the Pacific oyster, and the smaller-scale efforts for the American oyster and the California mussel (*Mytilus californianus*) will help to round out the Pacific oyster work. Of interest are exonic SNPs and, in particular, SNPs located in protein-coding regions of genes. There is hope that these SNPs will prove more stable and show a much lower frequency of null alleles.

Sequencing technologies are rapidly advancing, and at some point, genome sequencing will likely require small-scale efforts as opposed to the large-scale sequencing efforts currently required. (Also see Chapters 25 and 26.) Such a change in efficiency would clearly open up vast opportunities in bivalve aquaculture and restoration. What is currently a slow process of moving from candidate EST, to gene identification, SNP discovery, and finally genetic mapping, would move much more rapidly. In addition, comparative genomics efforts would be significantly bolstered by inexpensive genome sequencing of candidate species.

Multiple genome sequences would facilitate SNP discovery, enabling high-density genetic maps so that QTL could be narrowed to much smaller candidate regions as opposed to the current tens-of-centiMorgans. This would also require high throughput SNP genotyping, but platforms are currently available. As with most other molecular techniques, development will be driven by model species and then, with time, will trickle down to nonmodel species; systems that are well developed, such as the oyster, will be in an excellent position to capitalize on these techniques.

Probably the greatest challenge facing genomics in nonmodel organisms is bioinformatics (see Chapter 27). The ability to produce large amounts of data requires the development of tools to organize and analyze this data. Again, tools developed for model organisms will have to be adapted to these nonmodel species. In contrast to molecular tools, adapting bioinformatic tools should be relatively inexpensive because they only require adequate computing power.

Understanding the System

The primary goal of genomic work, and in particular genomic work in oysters, is focused on understanding the genetic, biochemical, and physiological system that

produces a particular phenotype. In the case of the Pacific oyster, there is great interest in understanding the hybrid vigor phenotype to capture this potential for aquaculture. Disease resistance is of interest for the American oyster, with the ultimate goal being the reestablishment of a healthy oyster fishery, particularly in the Chesapeake Bay.

Transcriptomic studies offer the potential to identify the genes contributing to these phenotypes, but with one particular caveat. If these phenotypes are the result of a complex set of genic interactions or pathways, as they likely are, transcriptomics will produce candidates that may be important in development of the phenotype (proximate causal factors), but are not necessarily the segregating genetic variation leading to the phenotype (the ultimate causal factor). Recent developments, in particular the approach of mapping expression as a quantitative trait, offer a potential bridge between these proximate and ultimate factors. See Gibson and Weir (2005) for a review.

Predicting the Future: Biomarkers

The ultimate goal of much of the oyster genomics work is predictive. The large number of lines and potential hybrid crosses make it impractical to grow and to harvest all possible pairwise hybrid combinations. Ideally, identifying the specific pairwise cross, or crosses, that will produce the greatest yield would be done early in the larval life stage (e.g., at day 2 or 3 when feeding begins). Early prediction of yield would allow a hatchery to separate from the hundreds of possible crosses (Hedgecock 2005), the handful that will produce the greatest growth, and to focus its efforts on these crosses, significantly increasing the efficiency of oyster production and moving one step closer to fulfilling Brook's desire to increase the supply indefinitely.

Acknowledegments

Jason P. Curole is supported by a USDA Postdoctoral Fellow grant in Genomics and Genetic Mechanisms. Research on oyster genomics in the Hedgecock lab has been supported by grants from the USDA National Research Initiative Competitive Grants Program in Animal Genomics, by the USDA Western Regional Aquaculture Center, and the National Science Foundation.

References

Ahmed M and AK Sparks. 1967. A preliminary study of chromosomes of two species of oysters (*Ostrea lurida* and *Crassostrea gigas*). J Fish Res Board Can, 24, pp. 2155–2159.

Anisimov S. 2005. A large-scale screening of the normalized mammalian mitochondrial gene expression profiles. Genet Res, Camb, 86, pp. 127–138.

Astanei I, E Gosling, J Wilson, and E Powell. 2005. Genetic variability and phylogeography of the invasive zebra mussel, *Dreissena polymorpha* (Pallas). Mol Ecol, 14, pp. 1655–1666.

Beaumont AR, CM Beveridge, and MD Budd. 1983. Selection and heterozygosity within single families of the mussel, *Mytilus edulis* (L.). Mar Biol Letters, 4, pp. 151–161.

Birchler JA, DL Auger, and NC Riddle. 2003. In search of the molecular basis of heterosis. The Plant Cell, 15, pp. 2236–2239.

Brake J, F Evans, and C Langdon. 2004. Evidence for genetic control of pigmentation of shell and mantle edge in selected families of pacific oysters, *Crassostrea gigas.* Aquaculture, 229, pp. 89–98.

Britten HB. 1996. Meta-analyses of the association between multilocus heterozygosity and fitness. Evolution, 50, pp. 2158–2164.

Brooks WK. 1905. *The Oyster.* The Johns Hopkins University Press, Baltimore.

Bucklin KA. 2002. *Analysis of the genetic basis of inbreeding depression in the Pacific oyster* Crassostrea gigas. PhD thesis, University of California Davis.

Buroker NE, WK Hershberger, and KK Chew. 1979. Population-genetics of the family Ostreidae. I. Intraspecific studies of *Crassostrea gigas* and *Saccostrea commercialis.* Mar Biol, 54, pp. 157–169.

Clabby C, U Goswami, F Flavin, NP Wilkins, JA Houghton, and R Powell. 1996. Cloning, characterization and chromosomal location of a satellite DNA from the Pacific oyster, *Crassostrea gigas.* Gene, 168, pp. 205–209.

Colson I, SJ Macdonald, and DB Goldstein. 1999. Microsatellite markers for interspecific mapping of *Drosophila simulans* and *D. sechellia.* Mol Ecol, 8, pp. 1951–1955.

Crow JF. 1998. 90 years ago: the beginning of hybrid maize. Genetics, 148, pp. 923–928.

Cunningham C, J Hakima, M Jenny, G Fang, C Saski, R Chapman, M Lundqvist, R Wing, P Cupit, P Gross, G Warr, and J Tomkins. 2006a. New resources for marine genomics: BAC libraries for the Eastern and Pacific oysters (*Crassotrea virginica* and *C. gigas).* Mar Biotechnol, 8, pp. 521–533.

Cunningham C, M Jenny, R Chapman, GW Warr, J Almeida, A Chen, D McKillen, and H Trent. 2006b. Design and characterization of a multi-species oyster cDNA microarray. In: National Shellfisheries Association, Programs and Abstracts, volume 98.

Curole JP and D Hedgecock. 2005. High frequency of SNPs in the Pacific oyster genome. In: Plant and Animal Genome Meetings XIII, http://www.intl-pag.org/13/abstracts/PAG13_W026.html.

Curole JP and D Hedgecock. 2006. Genetic mapping of candidate loci for growth heterosis in the Pacific oyster. In: Plant and Animal Genome Meetings XIV, http://www.intl-pag.org/14/abstracts/PAG14_W186.html.

Davis JP. 2006. Geoduck aquaculture in North America. In: National Shellfish Association 98 Annual Meeting.

Evans F, JB Matson, and C Langdon. 2003. The effects of inbreeding on performance traits of adult Pacific oysters (*Crassostrea gigas*). Aquaculture, 230, pp. 89–98.

FAO Fisheries Department. 2004. *The state of wold fisheries and aquaculture.* United Nations.

Fitzsimmons NN, C Moritz, and SS Moore. 1995. Conservation and dynamics of microsatellite loci over 300-million years of marine turtle evolution. Mol Biol Evol, 12, 432–440.

Foltz DW. 1986. Null alleles as a possible cause of heterozygote deficiencies in the oyster *Crassostrea virginica* and other bivalves. Evolution, 40, pp. 869–870.

Foltz DW and M Chatry. 1986. Genetic heterozygosity and growth rate in Louisiana oysters. Aquaculture, 57, pp. 261–269.

Fujio Y. 1982. A correlation of heterozygosity with growth rate in the Pacific oyster, *Crassostrea gigas.* Tohoku J Agric Res, 33, pp. 66–75.

Gaffney PM. 1990. Enzyme heterozygosity, growth rate, and viability in *Mytilus edulis:* another look. Evolution, 44, pp. 204–210.

Gaffney PM and TM Scott. 1984. Genetic heterozygosity and production traits in natural and hatchery populations of bivalves. Aquaculture, 42, p. 289.

Garza JC, M Slatkin, and NB Freimer. 1995. Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. Mol Biol Evol, 12, pp. 594–603.

Gibson G and B Weir. 2005. The quantitative genetics of transcription. Trends Genet, 21, pp. 616–623.

Gosling EM. 1989. Genetic heterozygosity and growth rate in a cohort of *Mytilus edulis* from the Irish coast. Mar Biol, 100, pp. 211–215.

Gowen JW. 1952. *Heterosis: a record of researches directed toward explaining and utilizing the vigor of hybrids.* Iowa State University Press, Ames.

Gueguen Y, JP Cadoret, D Flament, C Barreau-Roumiguiere, AL Girardot, J Garnier, A Hoareau, E Bachere, and JM Escoubas. 2003. Immune gene discovery by expressed sequence tags generated from hemocytes of the bacteria-challenged oyster, *Crassostrea gigas.* Gene, 303, pp. 139–145.

Hackett C and Broadfoot L. 2003. Effects of genotyping errors, missing values and segregation distortion in molecular marker data on the construction of linkage maps. Heredity, 90, pp. 33–38.

Harr B, B Zangerl, G Brem, and C Schlötterer. 1998. Conservation of locus-specific microsatellite variability across species: a comparison of two *Drosophila* sibling species, *D. melanogaster* and *D. simulans.* Mol Biol Evol, 15, pp. 176–184.

Hawkins A, B Bayne, and A Day. 1986. Protein-turnover, physiological energetics and heterozygosity in the blue mussel, *Mytilus edulis*—the basis of variable age-specific growth. P Roy Soc B-Biol Sci, 229, pp. 161–176.

Hedgecock D. 1994. Does variance in reproductive success limit effective population size in marine organisms? In: Beaumont A, Ed. Genetics and evolution of aquatic organisms. Chapman & Hall, London, pp. 122–134.

Hedgecock D. 2005. Crossbreeding Pacific oyster for high yield: Project termination report. Technical report, Western Regional Aquaculture Center.

Hedgecock D, V Chow, and RS Waples. 1992. Effective population numbers of shellfish broodstocks estimated from temporal variance in allelic frequencies. Aquaculutre, 108, pp. 215–232.

Hedgecock D and JP Davis. 2000. Improving Pacific oyster broodstock through crossbreeding. J Shellfish Res, 19, pp. 614–615.

Hedgecock D and JP Davis. 2006. Heterosis and crossbreeding in the Pacific oyster, *Crassotrea gigas.* Aquaculture, submitted.

Hedgecock D, P Grupe, and ML Voigt. 2006a. Mapping genes affecting shell color and shape in the Pacific oyster *Crassostrea gigas.* In: National Shellfisheries Association, Programs and Abstracts, volume 98.

Hedgecock D, G Li, S Hubert, K Bucklin, and V Ribes. 2004a. Widespread null alleles and poor cross-species amplification of microsatellite DNA loci cloned from the Pacific oyster, *Crassostrea gigas.* J Shellfish Res, 23, pp. 379–385.

Hedgecock D, G Li, and ML Voigt. 2004b. Mapping heterosis QTL in the Pacific oyster *Crassostrea gigas.* In: Plant and Animal Genome Meetings XII. Plant and Animal Genome Meetings.

Hedgecock D, JZ Lin, S DeCola, CD Haudenschild, E Meyer, DT Manahan, and B Bowen. 2006b. Transcriptomic analysis of growth heterosis in larval Pacific oysters (*Crassotrea gigas*), Proc Natl Acad Sci USA, 104, pp. 2313–2318.

Hedgecock D, DJ McGoldrick, and BL Bayne. 1995. Hybrid vigor in Pacific oysters: an experimental approach using crosses among inbred lines. Aquaculture, 137, pp. 285–298.

Hedgecock D, DJ McGoldrick, DT Manahan, J Vavra, N Appelmans, and BL Bayne. 1996. Quantitative and molecular genetic analyses of heterosis in bivalve molluscs. J Exp Mar Biol Ecol, 203, pp. 49–59.

Hedgecock D and F Sly. 1990. Genetic drift and effective population sizes of hatchery-propagated stocks of the Pacific oyster, *Crassostrea gigas.* Aquaculture, 88, pp. 21–38.

Houle D. 1989. Allozyme-associated heterosis in *Drosophila melanogaster.* Genetics, 123, pp. 789–801.

Hu YP, RA Lutz, and RC Vrijenhoek. 1993. Overdominance in early-life stages of an American oyster strain. J Heredity, 84, pp. 254–258.

Hubert S and D Hedgecock. 2004. Linkage maps of microsatellite DNA markers for the Pacific oyster *Crassostrea gigas.* Genetics, 168, pp. 351–362.

Huttunen S and C Schötterer. 2002. Isolation and characterization of microsatellites in *Drosophila virilis* and their cross species amplification in members of the *D. virilis* group. Mol Ecol Notes, 2, pp. 593–597.

Jenny MJ, R Chapman, A Mancia, YA Chen, DJ Mckillen, H Trent, JS Almeida, P Lang, JM Escoubas, E Bachere, ZJ Liu, PS Gross, C Cunningham, PEM Cupit, X Guo, A Tanguy, S De Guise, and GW Warr. 2006. New resource for oyster genomics: development and characterization of a cDNA microarray for *Crassostrea virginica* and *C. gigas.* Mar Biotechnol, in press.

Jenny MJ, AH Ringwood, ER Lacy, AJ Lewitus, J Kempton, P Gross, G Warr, and R Chapman. 2002. Potential indicators of stress response identified by expressed sequence tag analysis of hemocytes and embryos from the American oyster, *Crassostrea virginica.* Mar Biotechnol, 4, pp. 81–93.

Jongeneel CV, C Iseli, BJ Stevenson, GJ Riggins, AM Lal, RA Harris, MJ O'Hare, AM Neville, AJG Simpson, and RL Strausberg. 2003. Comprehensive sampling of gene expression in human cell lines with Massively Parallel Signature Sequencing. Proc Natl Acad Sci USA, 100, pp. 4702–4705.

Koehn RK, WJ Diehl, and TM Scott. 1988. The differential contribution by individual enzymes of glycolysis and protein catobilism to the relationship between heterozygosity and growth rate in the Coot clam, *Mulinia lateralis.* Genetics, 118, pp. 121–130.

Koehn RK and PM Gaffney. 1984. Genetic heterozygosity and growth rate in *Mytilus edulis.* Mar Biol, 82, pp. 1–7.

Lannan JE. 1980. Broodstock management of *Crassostrea gigas:* IV. inbreeding and larval survival. Aquaculture, 21, pp. 352–356.

Launey S and D Hedgecock. 2001. High genetic load in the Pacific oyster *Crassostrea gigas.* Genetics, 159, pp. 255–265.

Leitão A, C Thiriot-Quiévreux, P Boudry, and I Malheiro. 1999. A 'G' chromosome banding study of three cupped oyster species: *Crassostrea gigas, Crassostrea angulata* and *Crassostrea virginica* (Mollusca: Bivalvia). Genet Sel Evol, 31, pp. 519–527.

Li G, S Hubert, K Bucklin, V Ribes, and D Hedgecock. 2003. Characterization of 79 microsatellite DNA markers in the Pacific oyster *Crassostrea gigas.* Mol Ecol Notes, 3, pp. 228–232.

Li L and X Guo. 2004. AFLP-based genetic linkage maps of the Pacific oyster *Crassostrea gigas* Thunberg. Mar Biotechnol, 6, pp. 26–36.

Li WH, S Yi, and K Makova. 2002. Male-driven evolution. Curr Opin Genet Dev, 12, pp. 650–656.

Longwell A, S Stiles, and D Smith. 1967. Chromosome complement of the American oyster *Crassostrea virginica,* as seen in meiotic and cleaving eggs. Can J Genet Cytol, 9, pp. 845–856.

Mallet AL, E Zouros, KE Gartner-Kepkay, KR Freeman, and L Dickie. 1985. Larval viability and heterozygote deficiency in populations of marine bivalves: evidence from pair matings. Mar Biol, 87, pp. 165–172.

McAndrew BJ, RD Ward, and JA Beadmore. 1986. Growth rate and heterozygosity in plaice, *Pleuronectes platessa.* Heredity, 57, pp. 171–180.

McGoldrick DJ and D Hedgecock. 1997. Fixation, segregation and linkage of allozyme loci in inbred families of the pacific oyster *Crassostrea gigas* (Thunberg): implications for the causes of inbreeding depression. Genetics, 146, pp. 321–334.

McGoldrick D, D Hedgecock, L English, P Baoprasertkul, and R Ward. 2000. The transmission of microsatellite alleles in Australian and North American stocks of the Pacific oyster (*Crassostrea gigas*): selection and null alleles. J Shellfish Res, 19, pp. 779–788.

Meyer E. 2006. *Molecular Biological Analyses Of Nutrient Transporters And Growth Physiology In Marine Invertebrate Larvae.* PhD thesis, University of Southern California.

Moynihan EP and GAT Mahon. 1983. Quantitative karotype analysis in the mussel *Mytilus edulis* L. Aquaculture, 33, pp. 301–309.

National Research Council. 2004. *Non-native oysters in the Chesapeake bay. Committee on non-native oysters in the Chesapeake bay.* National Academies Press.

Newell R. 1988. Ecological changes in Chesapeake bay: are they the result of overharvesting the eastern oyster (*Crassostrea virginica*)? In: Lynch M and Krome E (eds) Understanding the estuary: advances in the Chesapeake Bay research. Chesapeake Research Consortium Publication 129 (CBP/TRS 24/88), pp. 536–546.

Noor M, R Kliman, and C Machado. 2001. Evolutionary history of microsatellites in the obscura group of *Drosophila.* Mol Biol Evol, 18, pp. 551–556.

Pace DA, AG Marsh, PK Leong, AJ Green, D Hedgecock, and DT Manahan. 2006. Physiological bases of genetically determined variation in growth of marine invertebrate larvae: a study of growth heterosis in the bivalve *Crassostrea gigas.* J Exp Mar Biol Ecol, 335, pp. 188–209.

Pauly D and J Maclean. 2003. *In a perfect ocean: the state of fisheries and ecosystems in the North Atlantic ocean.* Island Press, Washington.

Pépin L, Y Amigues, A Lepingle, J Berthier, A Bensaid, and D Vaiman. 1995. Sequence conservation of microsatellites between *Bos taurus* (cattle), *Capra hircus* (goat) and related species—examples of use in parentage testing and phylogeny analysis. Heredity, 74, pp. 53–61.

Pogson G and E Zouros. 1994. Allozyme and RFLP heterozygosities as correlates of growth rate in the scallop *Placopecten magellanicus:* a test of the associative overdominance hypothesis. Genetics, 137, pp. 221–231.

R Development Core Team. 2006. R: a language and environment for statistical computing. Published on the World Wide Web.

Ragone Calvo LM, GW Calvo, and EM Burreson. 2003. Dual disease resistance in a selectively bred eastern oyster, *Crassostrea virginica,* strain tested in Chesapeake Bay. Aquaculture, 220, pp. 69–87.

Reece K, W Ribeiro, P Gaffney, R Carnegie, and S Allen. 2004. Microsatellite marker development and analysis in the eastern oyster (*Crassostrea virginica*): confirmation of null alleles and non-mendelian segregation ratios. J Heredity, 95, pp. 346–352.

Reece K, W Ribeiro, C Morrison, and PM Gaffney. 2001. *Crassostrea virginica* microsatellite markers development, testing and preliminary linkage analysis. In: Plant and Animal Genome Meetings IX, http://www.intl-pag.org/pag/9/abstracts/W10_07.html.

Rico C, I Rico, and G Hewitt. 1996. 470 million years of conservation of microsatellite loci among fish species. P Roy Soc B-Biol Sci, 263, pp. 549–557.

Sato M, K Kawamata, N Zaslavskaya, A Nakamura, T Ohta, T Nishikiori, V Brykov, and K Nagashima. 2005. Development of microsatellite markers for Japanese scallop (*Mizuhopecten yessoensis*) and their application to a population genetic study. Mar Biotechnol, 7, pp. 713–728.

Schlötterer C, B Amos, and D Tautz. 1991. Conservation of polymorphic simple sequence loci in Cetacean species. Nature, 354, pp. 63–65.

Sekino M, M Hamaguchi, F Aranishi, and K Okoshi. 2003. Development of novel microsatellite DNA markers from the Pacific oyster *Crassostrea gigas.* Mar Biotechnol, 5, pp. 227–233.

Swanson-Wagner RA, Y Jia, R Decook, LA Borsuk, D Nettleton, and PS Schnable. 2006. All possible modes of gene action are observed in a global comparison of gene expression in a maize F1 hybrid and its inbred parents. Proc Nat Acad Sci USA, 103, pp. 6805–6810.

Tanguy A, X Guo, and SE Ford. 2004. Discovery of genes expressed in response to *Perkinsus marinus* challenge in eastern (*Crassostrea virginica*) and Pacific (*C. gigas*) oysters. Gene, 38, pp. 121–131.

Thiriot-Quiéveux C and N Ayraud. 1982. Les caryotypes de quelques espéces de bivalves et de gastéropodes marins. Mar Biol, 70, pp. 165–172.

Vadopalas B and P Bentzen. 2000. Isolation and characterization of di- and tetranucleotide microsatellite loci in Geoduck clams, *Panopea abrupta.* Mol Ecol, 9, pp. 1435–1436.

Wada K. 1975. Electrophoretic variants of leucine aminopeptidase of the Japanese pearl oyster *Pinctada fucata* (Gould). Bull Natl Pearl Res Lab Jap, 19, pp. 2125–2156.

Wang Y and X Guo. 2004. Chromosomal rearrangement in Pectinidae revealed by rRNA loci and implications for bivalve evolution. Biol Bull, 207, pp. 247–256.

Wang Y, Z Xu, and X Guo. 2001. A centromeric satellite sequence in the Pacific oyster (*Crassostrea gigas* Thunberg) identified by fluorescence in situ hybridization. Mar Biotechnol, 3, pp. 486–492.

Wetz MS, AJ Lewitus, E Koepfler, and K Hayes. 2002. Impact of the eastern oyster *Crassostrea virginica* on microbial community structure in a salt marsh estuary. Aquat Microb Ecol, 28, pp. 87–97.

Williams G. 1975. *Sex and evolution.* Princeton University Press ID, Princeton, NJ.

Xu Z, X Guo, P Gaffney, and JC Pierce. 2001. Chromosomal location of the major ribosomal RNA genes in *Crassostrea virginica* and *Crassostrea gigas.* Veliger, 44, pp. 79–83.

Yu Z and X Guo. 2006. Identification and mapping of disease-resistance QTLs in the eastern oyster, *Crassostrea virginica* Gmelin. Aquaculture, 254, pp. 160–170.

Zouros E, A Ball, C Saavedra, and K Freeman. 1994. An unusual type of mitochondrial DNA inheritance in the blue mussel *Mytilus.* Proc Natl Acad Sci USA, 91, pp. 7463–7467.

Zouros E and DW Foltz. 1987. The use of allelic isozyme variation for the study of heterosis. Isozymes: Curr Topics Biol Medical Res, 13, pp. 1–59.

Zouros E and G Pogson. 1994. Heterozygosity, heterosis, and adaptation. In: Beaumont A, Ed. Genetics and Evolution of Aquatic Organisms. London, Chapman and Hall.

Zouros E, M Romero-Dorey, and AL Mallet. 1988. Heterozygosity and growth in marine bivalves: further data and possible explanations. Evolution, 42, pp. 1332–1341.

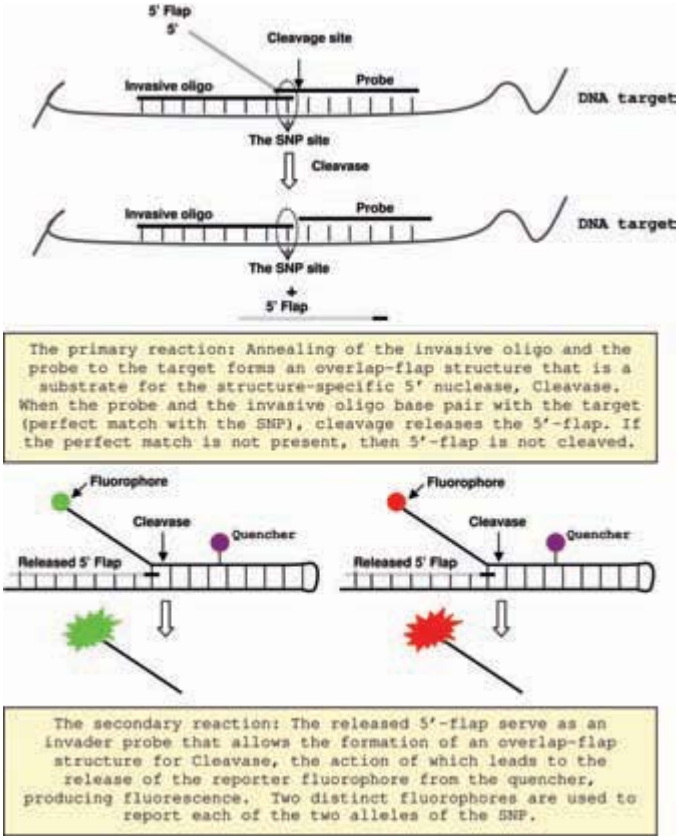


Figure 6.5. A schematic presentation of the invader assay of SNPs.

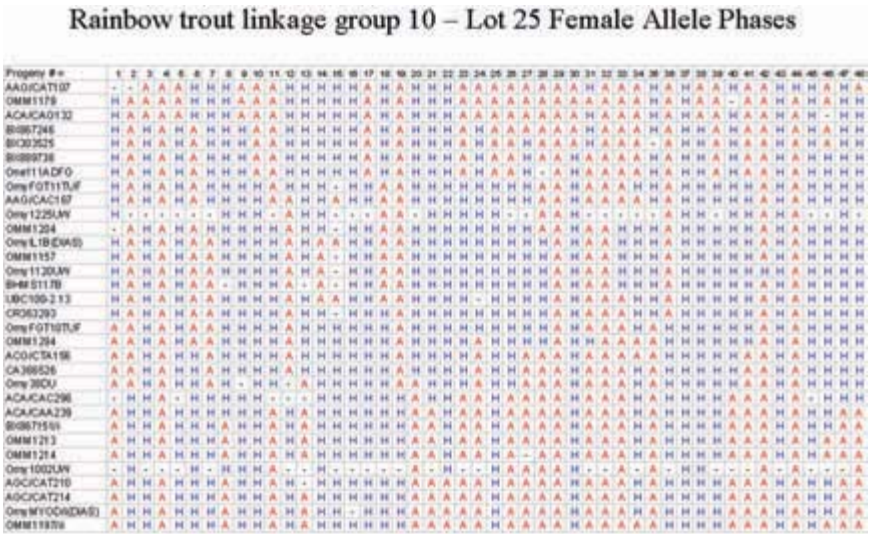


Figure 10.1. A phase map of linkage group 10 from the female genetic map in rainbow trout obtained from one of the mapping panels used by Danzmann et al. (2005).

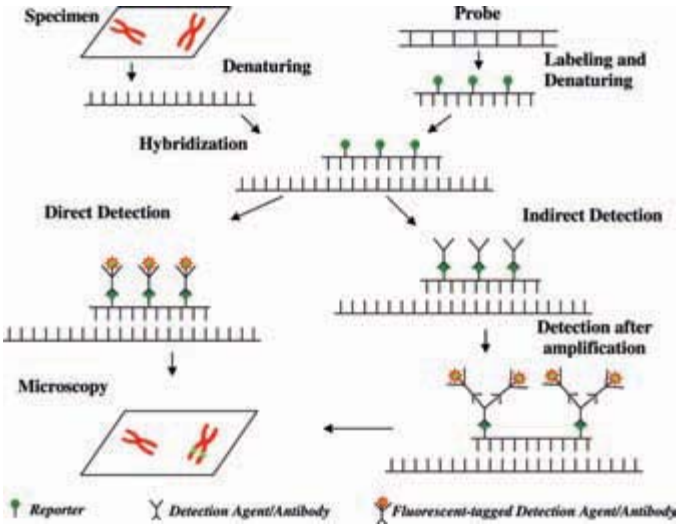


Figure 17.1. Schematic presentation of basic steps in fluorescence in situ hybridization.

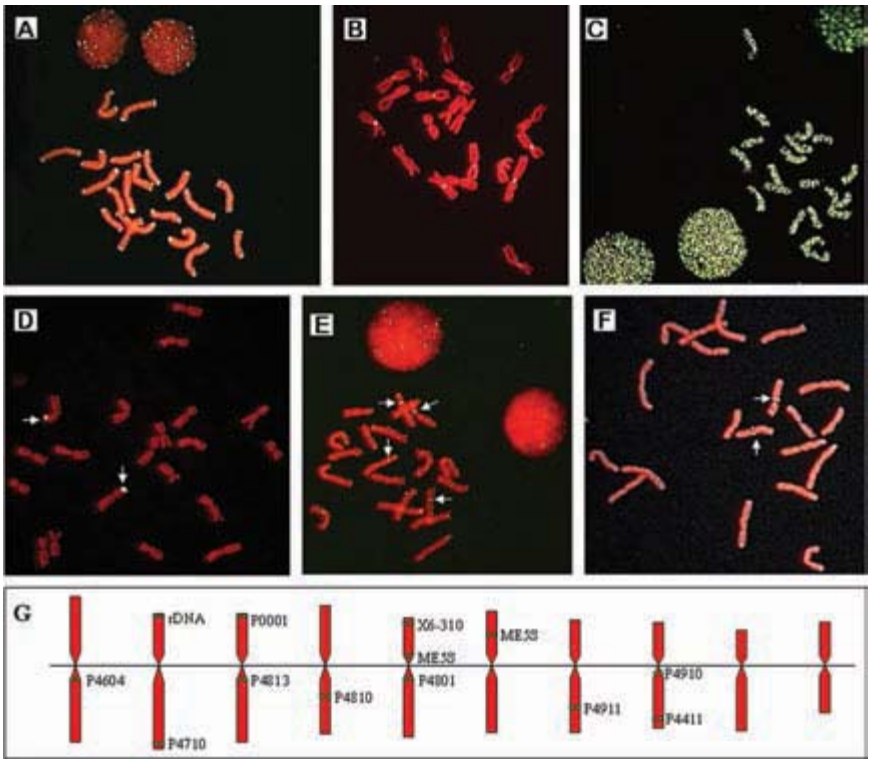


Figure 17.2. Fluorescence in situ hybridization with various probes in the eastern oyster. A, the vertebrate telomeric sequence; B, a centromeric element; C, a small interspersed repetitive element; D, the major rRNA genes (18S-5.8S-28S); E, the 5S rRNA gene; F, a P1 clone; and G, a preliminary cytogenetic map.

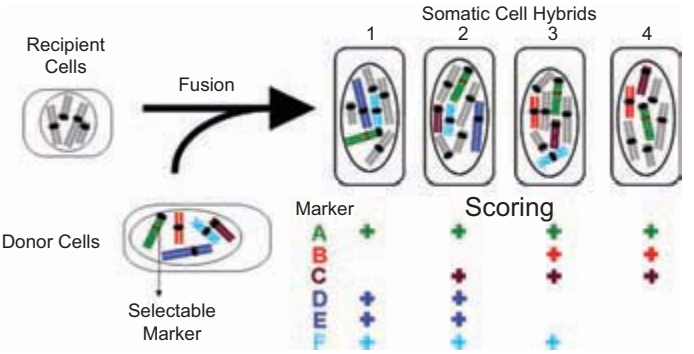


Figure 18.1. A schematic representing a fusion event between a donor cell line and a recipient cell line in the formation of a Somatic Cell Hybrid panel.

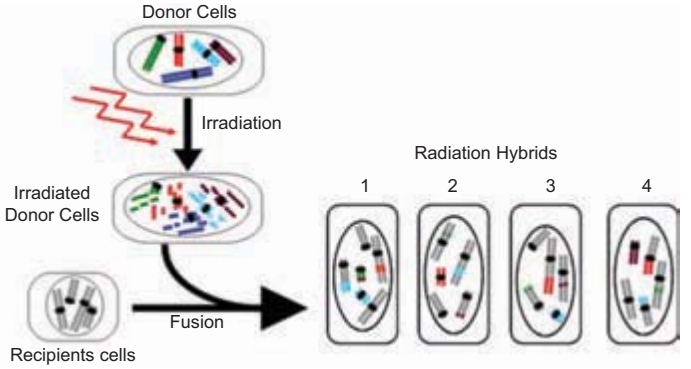


Figure 18.2. A schematic representing irradiation of donor cells and fusion to recipient cells to form a radiation hybrid cell panel. Donor cell fragments are retained in the hybrid cell lines as translocations, insertions, or partial chromosomes retaining replication machinery.

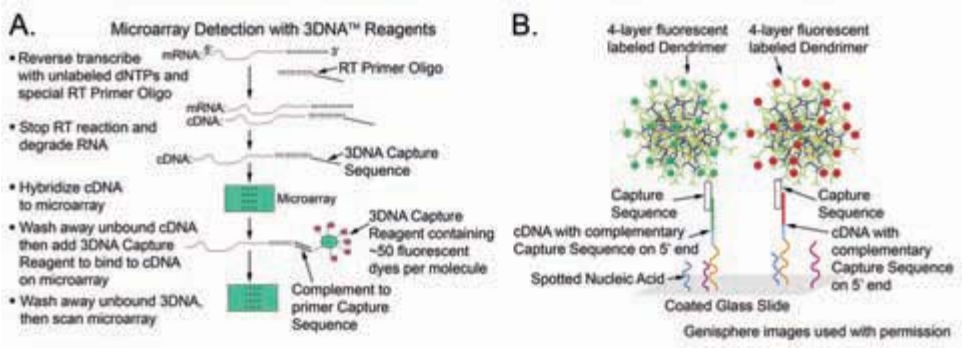


Figure 22.1. Steps involved in cDNA target synthesis and microarray hybridizations using Genisphere (Hatfield, PA) Expression Array Detection Kits. B. An example of a dual color nucleic acid microarray hybridization using Genisphere 3DNA reagents and methods. Genisphere figures are used with permission.

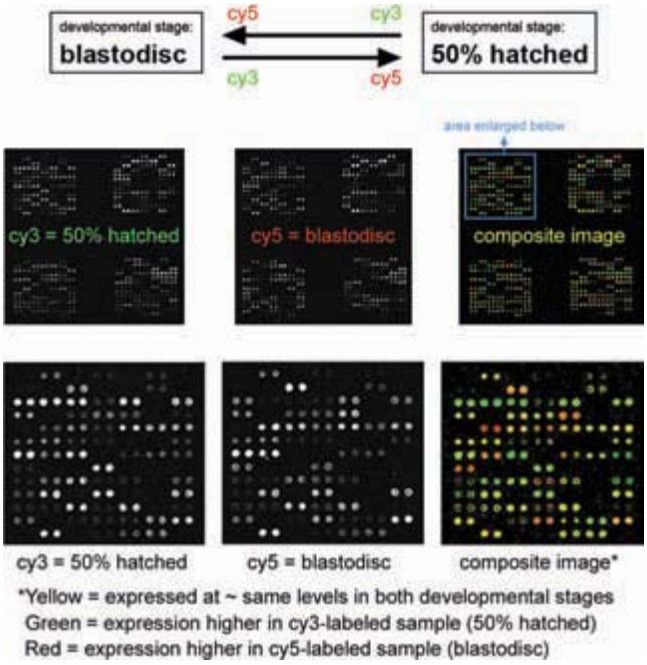


Figure 22.2. Design for a microarray experiment directly comparing global gene expression in two salmonid developmental stages. Arrows symbolize microarrays, with arrow bases showing Cy3-labeled samples and arrow heads showing Cy5-labeled samples.

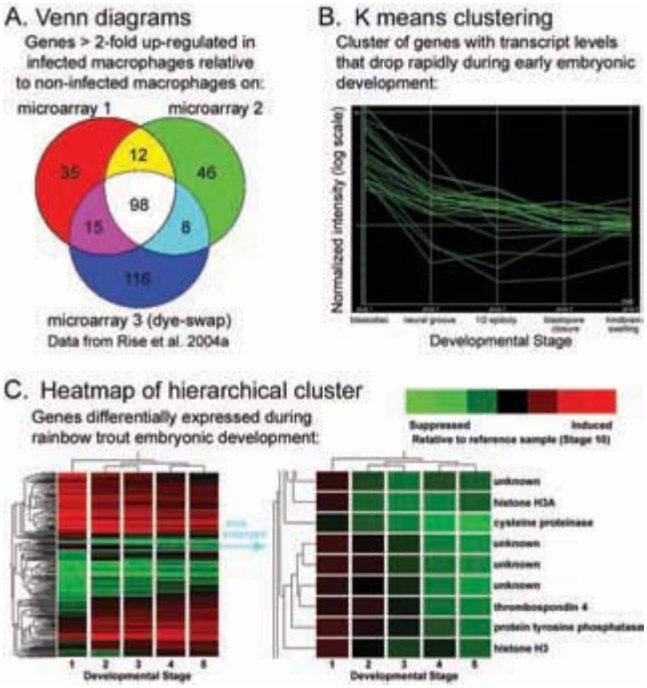


Figure 22.4. Examples of methods for analyzing microarray data.

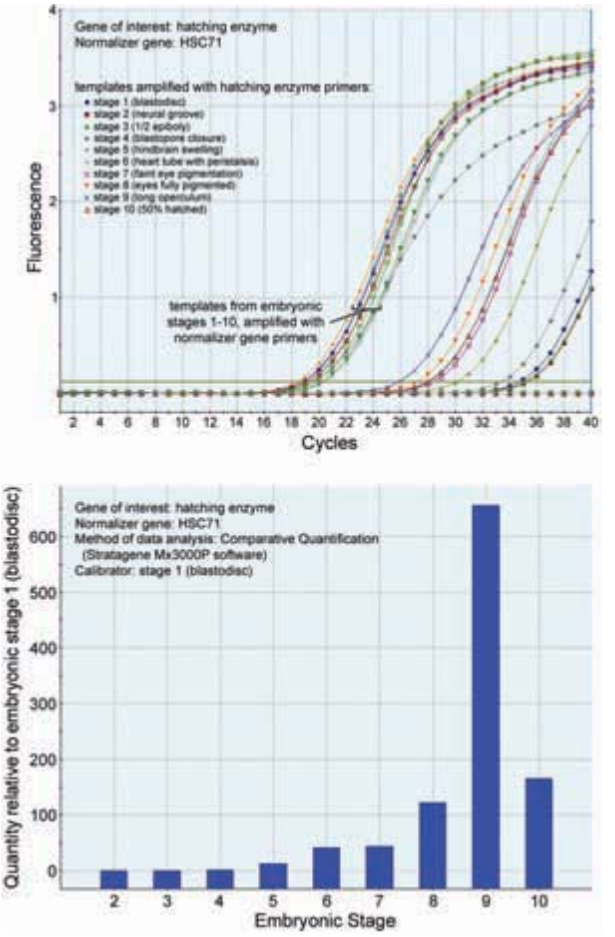


Figure 22.5. Amplification plots and chart showing developmental expression profile of a microarray-identified gene of interest (hatching enzyme) normalized to HSC71.

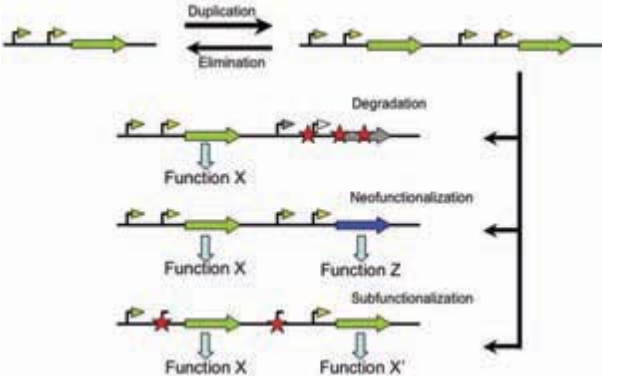


Figure 24.3. Gene duplication and loss. A gene is shown with two transcriptional elements.

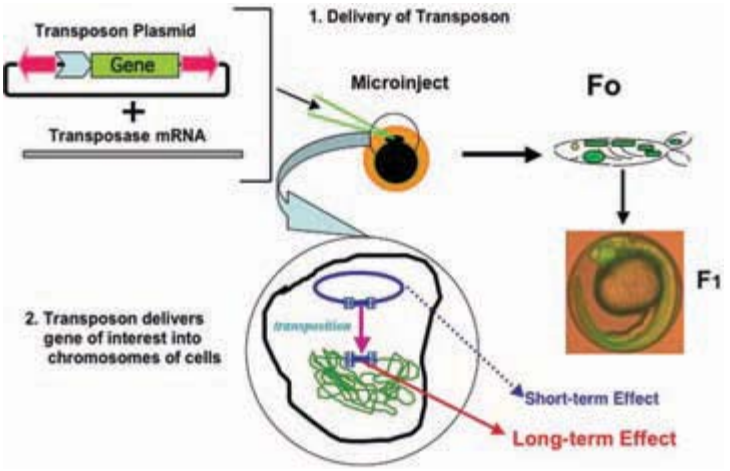


Figure 24.9. Transposon-mediated gene delivery in fish. Injection of Sleeping Beauty transposon vectors into a 1-cell fish embryo allows transposition of the transgene in a transposon from the plasmid into to a chromosome. The transposon (red inverted arrows with a promoter-gene inside, is carried on a plasmid. The transposase mRNA is co-injected and expresses the transposase enzyme that cuts the transposon out of the plasmid and integrates it into a chromosome. This leads to long-term expression of the reporter gene, which may interrupt and mutate a gene. Unintegrated plasmids generally express the transgene for only a short period of time. Expression of the transgene is mosaic in Fo animals but they pass on expression of the integrated gene in their offspring.

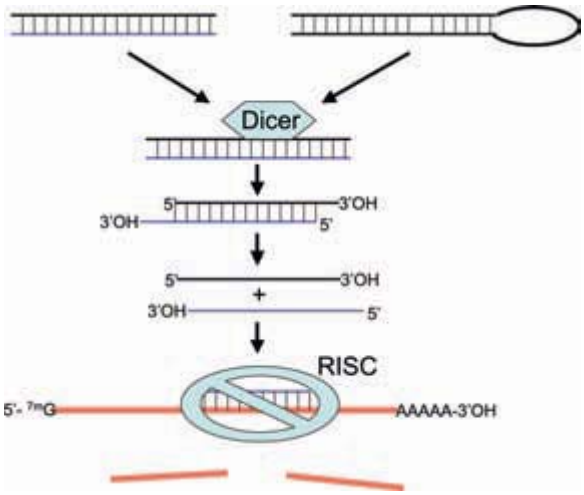


Figure 24.12. Production of RNAi to inhibit mRNA expression. Double-stranded RNA (dsRNA) from some source, top line, is recognized by the Dicer enzyme and cleaved into shorter, dsRNA segments of 21-23 bp with protruding 3'-ends. The strands of the short dsRNA then can denature and reanneal to a complementary sequence on an mRNA (red lines), which upon recognition by RISC will lead to mRNA degradation.

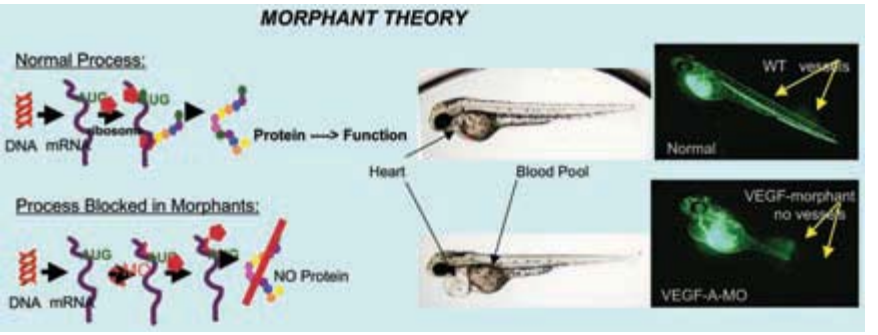


Figure 24.16. Mechanism of action of PMOs and an example of a PMO directed against VEGF in zebrafish. The top row of figures shows normal 48-hr zebrafish embryos with normal vascular that is highlighted by the green fluorescence in the blood vessels (right side). The heart is tucked under the head and the blood is distributed throughout the embryo in the vasculature. The bottom row of figures shows 48-hr VEGF Morphant embryos with a grossly enlarged heart and pooled blood (pinkish pool in the left image). The vasculature is not developed. The embryo is alive and can be subjected to further studies.

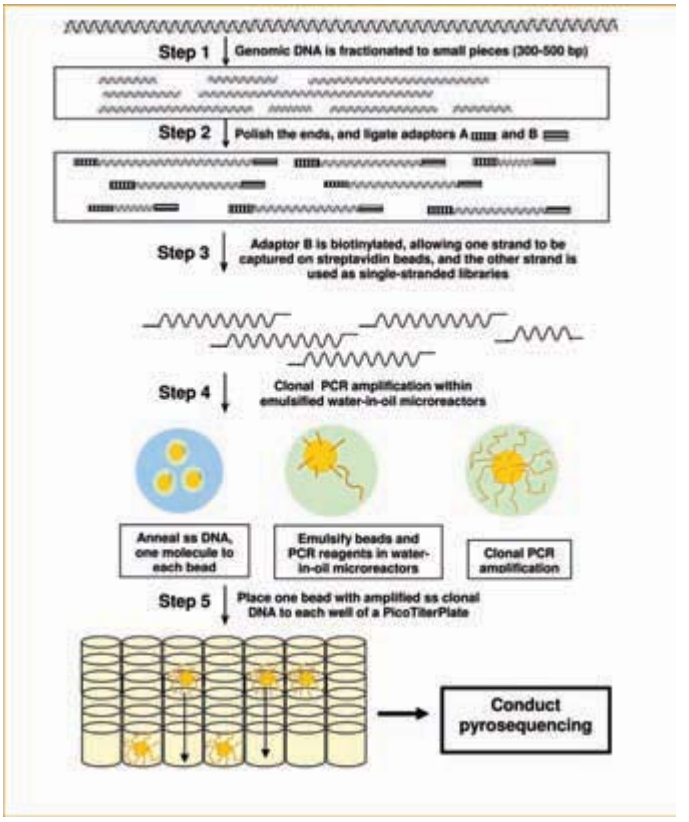


Figure 25.4. Schematic presentation of the principles of pyrosequencing.

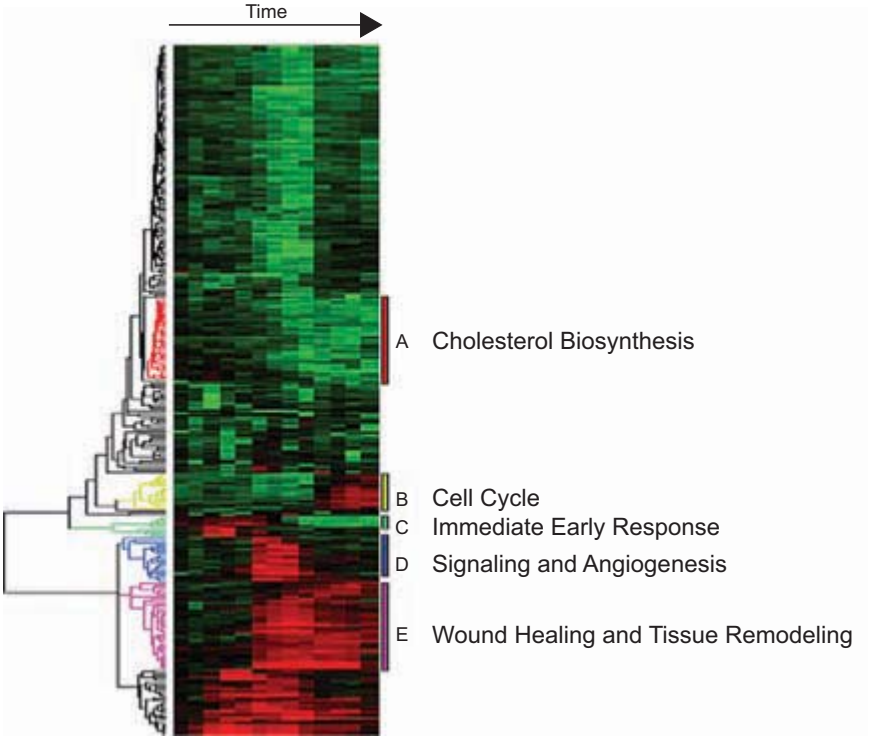


Figure 27.7. Hierarchical clustering of microarray data. Rows are genes. Columns are RNA samples at different time points. Values are the signals (expression levels), which are represented by the color spectrum. Green represents down-regulation while red represents up-regulation. The color bars beside the dendrogram show the clusters of genes which exhibit similar expression profiles (patterns). The bars are labeled with letters and description of possible biological processes involving the genes in the clusters. (Reprinted from Eisen et al., 1998).

Index

Bold page numbers refer to main entries.