Chapter 1
Genomic Variations and Marker Technologies for Genome-based Selection

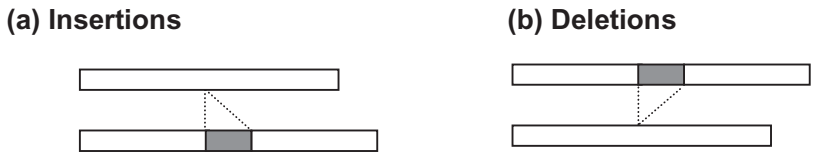
Zhanjiang (John) Liu

Genetic Variations at the Genomic Level

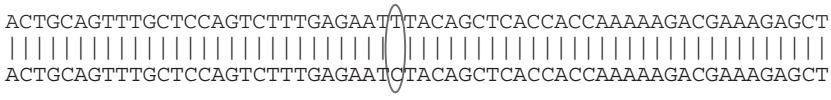
The genome compositions of each individual of the same species are similar but different at the level of DNA sequences and its encoding capacity (sometimes in terms of what genes are transcribed, but perhaps more often in terms of how much the gene products are made), and thereby have different transcriptional activities, encoding similar but different proteins, or encoding same or similar proteins in different quantities, leading to different biological characteristics and performance. Upon comparison of the genomes of individuals within a population with their reference genome sequence of the species, several general types of genetic variations can be found (Figure 1.1): (1) a deletion due to the loss of one or more of bases; (2) insertion due to gain of one or more bases; (3) base substitution at various positions; (4) inversion of a DNA segment in its orientation; (5) rearrangements of multiple DNA segments within a both small and larger scope of the genome; and (6) copy number variation (CNV) due to insertions, deletions, and duplication or multiplication of a DNA segment(s). A deletion mutation and an insertion mutation can be viewed as the same phenomenon depending on what is used as the reference. Deletions/insertions in random genomic locations probably do not have much impact to its biological activities except when deletion/insertion happens within a gene or within its regulatory elements. Insertion/deletion of single base or two bases within a protein coding sequence would cause frameshift of the protein being encoded, thus leading to the completely different amino acid sequences downstream of the mutation. However, deletion/insertion of three bases or multiple of three bases (e.g., 6 base pair [bp], 9 bp) within a protein coding sequence would cause a deletion of one amino acid or multiple amino acids depending on the extent of the deletion/insertion. In the first case of deletion/insertion of one or two bases into a protein coding sequences, the biological impact could be highly significant. Such mutations could cause total loss of functions of the protein. In the later case, deletion/insertion of three or multiple of three bases would lead to a protein missing one or a few amino acids but the upstream and downstream amino acid sequences should still be the same. In this case, the protein function may or may not be altered depending

Next Generation Sequencing and Whole Genome Selection in Aquaculture Edited by Zhanjiang (John) Liu
 ISBN: 978-0-813-80637-2

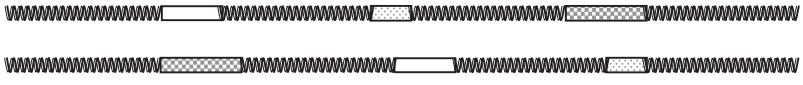
1. Indels



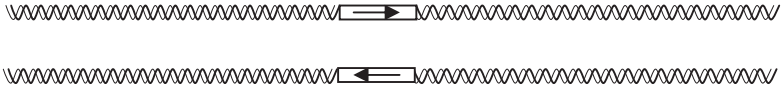
2. Base substitutions/single nucleotide polymorphisms



3. Rearrangements



4. Segmental inversions



5. Copy number variations

Figure 1.1 Types of genome variations. In principle, five types of genomic variations exist: 1. Indels that involve insertions or deletions of a segment, as indicated by the shaded boxes. 2. Base substitutions or single-nucleotide polymorphisms (SNPs) are simply differences of bases at a given DNA location. In the example, a T/C SNP is highlighted by the oval. 3. Rearrangements are genomic difference that resulted from the relocation of certain genomic segments of various sizes. Shown are three DNA segments that are present in both genomes, but they are located in different genome locations. Practically, such rearrangements can be intrachromosomal or extrachromosomal. 4. Segment inversions are changes of DNA segments in their orientation in the genome, as indicated by the change of the arrow direction. 5. Copy number variations are differences in copies of DNA segments (genes or just genomic segments) within genomes. In the example, one open box segment is in the first genome, but two open boxes are in the second genome; similarly, different numbers of segments exist between the two genomes as indicated by different sketched boxes.

on the position of the mutation and the amino acids involved. Serious biological impact can still result. For instance, in the case cystic fibrosis (CF), a 3-bp deletion at the amino acid position of 506 of the cystic fibrosis transmembrane regulator (CFTR) protein would lead to the most serious form of CF, even though the resulting protein losses just one amino acid.

Genome variations involve a wide range of segmental inversions or rearrangements. Very similar to the situation of deletions and insertions, such sequence changes could have huge biological impact depending on the location of the mutation and the genes or gene regulatory sequences involved in such mutations.

The most widespread genomic variation among individuals within a population is base substitution. Such base substitution along the DNA chain is defined as single-nucleotide polymorphisms (SNPs).

Inversion of a DNA segment in its orientation can be quite widespread in the genome, but this type of variation have not been well studied and probably will not be very useful for large-scale genomic studies.

CNV due to insertions, deletions, and duplication or multiplication of a DNA segment is widespread, and this type of genomic variation caught the attention of genome researchers just recently. CNV can involve large or small genome segments that are duplicated or multiplied in one genome while not in another. Such CNVs can involve genes or just genomic segments that do not harbor genes. Obviously, when genes are involved, the duplicated or multiplied genes can affect genome expression activities. The significance of CNV has caught much attention recently, and CNV could potentially be used for whole genome selection programs upon identification of correlation or causation of certain genome segments with performance traits. The importance of CNV in teleost fish is further signified by the fact that teleost fish had an additional round of genome duplication followed with random gene loss, thereby resulting in various CNV situations involving various genes. Because of this significance, CNV is included as an independent chapter in this book (Chapter 2).

A Review of DNA Marker Technologies

The entire task of DNA marker technologies is to provide the means to reveal DNA-level differences of genomes among individuals of the same species, as well as among various related taxa. Historically, these measurements relied on phenotypic or qualitative markers. 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 are inferential through controlled breeding and performance studies (Parker et al., 1998; Okumuş and Çiftci, 2003). Because these markers are polygenically inherited and have low heritability, 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 including protein markers and DNA markers were developed to overcome problems associated with phenotypic markers.

Allozyme Markers

Much before the discovery of DNA markers, allozyme markers were used to identify broodstocks in fish and other aquaculture species (Kucuktas and Liu, 2007). Allozymes are different allelic forms of the same enzymes encoded at the same locus (Hunter and Markert, 1957; Parker et al., 1998; May, 2003). Genetic variations detected in allozymes may be the result of point mutations, insertions, or deletions (indels). Allozymes have had a wide range of applications in fisheries and aquaculture including population analysis, mixed stock analysis, and hybrid identification (May, 2003). However, they are becoming a marker type of the past due to the limited number of loci that in turn prohibits genome-wide coverage for the analysis of complex traits (Kucuktas and Liu, 2007). In addition, mutation at the DNA level that causes a replacement of a similarly charged amino acid may not be detected by allozyme electrophoresis. Another drawback is that the most commonly used tissues in allozyme electrophoresis are the muscle, liver, eye, and heart, the collection of which is lethal.

Restriction Fragment Length Polymorphism (RFLP)

Two specific technological advances, the discovery and application of restriction enzymes in 1973 and the development of DNA hybridization techniques in 1975, set the foundation for the development of the first type of DNA markers, RFLP (for a recent review, see Liu, 2007a). Restriction endonucleases cut DNA wherever their recognition sequences are encountered. Therefore, changes in the DNA sequence due to insertions/deletions (indels), base substitutions, or rearrangements involving the restriction sites can result in the gain, loss, or relocation of a restriction site. Digestion of DNA with restriction enzymes results in fragments whose number and size can vary among individuals, populations, and species. Two approaches are widely used for RFLP analysis. The first involves the use of Southern blot hybridization (Southern, 1975), while the second involves the use of PCR. Traditionally, fragments were separated using Southern blot analysis, in which genomic DNA is digested, subjected to electrophoresis through an agarose gel, transferred to a solid support such as a piece of nylon membrane, and visualized by hybridization to specific probes. Most recent analysis replaces the tedious Southern blot analysis with techniques based on 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 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.

The major strength of RFLP markers is that they are codominant markers; that is, both alleles in an individual are observed in the analysis. The major disadvantage of RFLP is the relatively low level of polymorphism. In addition, either sequence information (for PCR analysis) or a molecular probe (for Southern blot analysis) is required, making it difficult and time-consuming to develop markers in species

lacking known molecular information. Due to these disadvantages, the application of RFLP markers in aquaculture and fisheries has been, and will be, limited.

Mitochondrial Markers

Mitochondrial genome evolves more rapidly than the nuclear genome. The rapid evolution of the mitochondrial DNA (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. The analysis of mitochondrial markers is mostly RFLP analysis, or direct sequence analysis (Liu and Cordes, 2004). Due to the high levels of polymorphism and the ease of mtDNA analysis, mtDNA has been widely used as markers in aquaculture and fisheries settings. However, mtDNA is maternally inherited in most cases, and this non-Mendelian inheritance greatly limits the applications of mtDNA for genome research. In addition, most aquaculture-related traits are controlled by nuclear genes. For most aquaculture finfish species, their nuclear genome is at the level of a billion base pairs, while their mitochondrial genomes are usually tens of thousands of times smaller than the nuclear genome. Clearly, in spite of their usefulness for the identification of aquaculture stocks, mtDNA markers will not be tremendously useful for aquaculture genome research and genetic improvement programs in aquaculture. However, some recent studies suggested that mtDNA could influence performance traits such as growth (Steele et al., 2008).

Microsatellite Markers

When the Human Genome Project was launched in the mid-1980s, the capacity and capabilities of available DNA marker technologies seriously limited genome research. Such severe limits put pressure to develop more efficient marker systems for analysis of complex traits and genome organizations. At the end of 1980s, the simple sequence repeats (SSRs) or microsatellites were discovered; and they have since been used as one of the most preferred marker types because of their high levels of polymorphism, abundance, roughly even genome distribution, codominant inheritance, and small locus size that facilitate PCR-based genotyping (Tautz, 1989).

Microsatellites can be viewed as special cases of insertions or deletions. An addition of a dinucleotide microsatellite repeat can be viewed as an insertion of 2 bp into the genome. They are perhaps the most abundant type of insertions and deletions.

Microsatellites are SSRs of 1–6 bp. They 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 genome in terms of the involved base pairs, depending on the compactness of the genomes. Generally speaking, more compact genomes tend to contain smaller proportion of repeats including SSRs, but this generality is not always true. For example, the highly compact genome of Japanese pufferfish contains 1.29% of microsatellites, but its closely related *Tetraodon nigroviridis* genome contains 3.21% of microsatellites (Crollius et al., 2000).

During a genomic sequencing survey of channel catfish, microsatellites were found to represent 2.58% of the catfish genome (Xu et al., 2006; Liu et al., 2009). In fugu, one microsatellite was found for every 1.87 kb of DNA. For comparison, in the human genome, one microsatellite was found for 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% are present as trinucleotide repeats; and 13.5% as tetranucleotide repeats. 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 is 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. Microsatellites longer than tetranucleotide repeats (penta- and hexanucluotides) are much less abundant, and are 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.

Microsatellites are 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 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 frameshift mutations in the translated sequences (Metzgar et al., 2000; Li and Guo, 2004). Because the majority of microsatellites exist in the form of dinucleotide repeats, any mutation by expansion or shrinking would cause frameshift of the protein encoding open frames if they reside within the coding region. This also explains why the majority of microsatellites residing within coding regions have been found to be trinucleotide repeats, although the presence of dinucleotide repeats and their mutations within the coding regions do occur.

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, although 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 one another by as little 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 quite a

simple sequence as well, prohibiting the design of PCR primers for the amplification of microsatellite loci within a small size limit.

Microsatellites are highly polymorphic as a result of their hypermutability, and thereby 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. Microsatellite mutation rates have been reported as high as 10^{-2} per generation (Weber and Wong, 1993; Crawford and Cuthbertson, 1996; Ellegren, 2000), which is several orders of magnitude greater than that of nonrepetitive DNA (10^{-9}; Li, 1997). In several fish species, the mutation rates of microsatellites were reported to be at the level of 10^{-3} per locus per generation: 1.3×10^{-3} in common carp (Zhang et al., 2008), 2×10^{-3} in pipefish (Jones et al., 1999), $3.9–8.5 \times 10^{-3}$ in salmon (Steinberg et al., 2002), and 2×10^{-3} in dollar sunfish (MacKiewicz et al., 2002).

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 are 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 that the patterns of microsatellite genotypes fit the genetic model under application.

The disadvantage of microsatellites as markers include the requirement for existing molecular genetic information, a large amount of up-front work for microsatellite development, and 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. Technically, the simplest way to identify and characterize a large number of microsatellites is through the construction of microsatellite-enriched small-insert genomic libraries (Ostrander 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. In spite of the simplicity in the construction of microsatellite-enriched libraries, and thereby the identification and characterization of microsatellite markers, for a large genome project, the real need of direct microsatellite marker development may not be the wisest approach. Recent progress in sequencing technologies with the next generation of sequencers will allow large numbers of genomic sequence tags to be generated that would include numerous microsatellites. Microsatellites can be identified and sequenced directly from genome sequence surveys such as bacterial artificial chromosome (BAC)-end sequencing (Xu et al., 2006; Somridhivej et al., 2008; Liu et al., 2009), and from expressed sequence tag (EST) analysis from which many microsatellites can be developed into type I markers (Liu et al., 1999; Serapion et al., 2004). 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 insufficient for the design of PCR primers. While introns are not a problem for

microsatellites derived from BAC-end sequencing, sequencing reactions often terminate immediately after the microsatellite repeats, which also makes flanking sequences insufficient for the design of PCR primers.

Microsatellites have been an extremely popular marker type in a wide variety of genetic investigations. Over the past decade, microsatellite markers have been used extensively in fisheries research including studies of genome mapping, parentage, kinships, and stock structure. The major application of microsatellite markers is for the construction of genetic linkage and quantitative trait locus (QTL) maps. This is because of the high polymorphic rate of microsatellite markers. When a resource family is produced, the male and female fish parents are likely to be 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 the use of microsatellite across species borders is sometimes possible if the flanking sequences are conserved (Fitzsimmons et al., 1995; Rico et al., 1996; Cairney et al., 2000; Leclerc et al., 2000). As a result, microsatellites can be also used for comparative genome analysis. If microsatellites can be tagged to gene sequences, their potential for use in comparative mapping is greatly enhanced.

In spite of the popularity and great utilization of microsatellites, several major limitations of microsatellites restrict them to rise to the top of all marker systems:

1. In spite of being very abundant, development of hundreds of thousands or millions of microsatellite markers is practically almost impossible.
2. Automation has not been possible for microsatellite genotyping. Multiplexing has been limited to about a dozen of loci, at the most.
3. For the most part, microsatellites can be just associated with traits, but are not usually the causes of the phenotypic variations.

On top of these limitations of microsatellites, recent advances in molecular markers will have a major impact on the choice of DNA markers. In particular, the rapid progress in SNP including its rapid identification and automation in genotyping make SNP the far more preferred marker system for genome studies.

Random Amplified Polymorphic DNA (RAPD) Markers

At the beginning of the 1990s, efforts were also devoted to develop multiloci, PCR-based fingerprinting techniques. Such efforts resulted in the development of two marker types that were highly popular for a while: RAPD (Welsh and McClelland, 1990; Williams et al., 1990) and amplified fragment length polymorphism (AFLP; Vos et al., 1995).

RAPD is a multilocus DNA fingerprinting technique using PCR to randomly amplify anonymous segments of nuclear DNA with a single short PCR primer (8–10 bp in length) (for a recent review, see Liu, 2007b). Because the primers are short, relatively low annealing temperatures (often 36–40°C) must be used. 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. 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 insertions or deletions (indels) in the regions between the two close primer binding sites. The potential power for detection of polymorphism is reasonably high as compared with RFLP, but much lower than microsatellites; 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 biallelic locus (presence or absence of an amplified product), polymorphic information content (PIC) values for RAPDs fall below those for microsatellites. The major advantages of RAPD markers are their applicability to all species regardless of known genetic, molecular, or sequence information, relatively high level of polymorphic rates, simple procedure, and a minimal requirement for both equipment and technical skills.

RAPD has been widely used in genetic analysis of aquaculture species, but its further application in genome studies is limited by its lack of high reproducibility and reliability. In addition, RAPD is inherited as dominant markers, and transfer of information with dominant markers among laboratories and across species is difficult.

AFLP Markers

Alternatives of RAPD that overcome the major problems such as its low reproducibility were actively sought in the early part of the 1990s. AFLP (Vos et al., 1995) was the outcome of such efforts. AFLP is based on the selective amplification of a subset of genomic restriction fragments using PCR (for a recent review, see Liu, 2007c). Genomic DNA is digested with restriction enzymes, and double-stranded DNA adaptors with known sequences are ligated to the ends of the DNA fragments to generate primer binding sites for 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 such 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.

AFLP analysis is an advanced form of RFLP. Therefore, the molecular basis for RFLP and AFLP are similar. First, any deletions and/or insertions between the two restriction enzymes, for example, between *Eco*RI and *Mse* I that are most often used in AFLP analysis, will cause shifts of 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 roughly 1 of 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 the 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 for a typical fish genome with 1 billion bp), a complete AFLP scan would also examine over 2 million bases immediately adjacent to the restriction sites.

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. For each pair of restriction enzyme used in the analysis, for example, *Eco*RI and *Mse* I, a total of approximately 500,000 *Eco*RI-*Mse* I fragments would exist for a genome with a size of 1×10^9 bp. Theoretically, 4096 primer combinations compose a complete genome-wide scan of the fragment length polymorphism using the two restriction enzymes if three bases are used for selective amplification. As hundreds of restriction endonucleases are commercially available, the total power of AFLP for analysis of genetic variation can not be exhausted. However, it is probably never necessary to perform such exhaustive analysis. Since over 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 takes no more than 5–10 primer combinations.

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. 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 numbers of 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 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; and 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.

AFLP has been widely used in aquaculture such as analysis of population structures, migration, hybrid identification, strain identification, parentage identification, genetic resource analysis, genetic diversity, reproduction contribution, and endangered species protection (Jorde et al., 1999; Seki 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; Campbell and Bernatchez, 2004; Mock et al., 2004; Simmons et al., 2006).

AFLP has also been widely used in genetic linkage analysis (Kocher et al., 1998; Liu et al., 1998, 1999; Griffiths and Orr, 1999; Agresti et al., 2000; Robison et al., 2001; Rogers et al., 2001; Li et al., 2003; Liu et al., 2003; Felip et al., 2005), and analysis of parental genetic contribution involving interspecific hybridization (Young et al., 2001) and meiogynogenesis (Felip et al., 2000). In a study of the black rockfish (*Sebastes inermis*), Kai et al. (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 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 et al. (2000) used AFLP for the analysis of five geographical populations of the 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.

In spite of its popularity, AFLP has two fundamental flaws that prohibit its wider applications in the future: the dominance inheritance and lack of information to link it to genome sequence information. In some cases, AFLP can be used as a rapid screening tool, and useful markers can then be converted to sequence-characterized amplified region (SCAR) markers. However, genome-scale applications of SCAR markers are unlikely.

SNP

SNP describes polymorphisms caused by point mutations that give rise to different alleles containing alternative bases at a given nucleotide position within a locus (for a recent review, see Liu, 2007d). Such sequence differences due to base substitutions have been well characterized since the beginning of DNA sequencing in 1977, but genotyping SNPs for large numbers of samples was not possible until several major technological advances in the late 1990s. SNPs are again becoming a focal point of molecular markers since they are the most abundant polymorphism in any organism, adaptable to automation, and reveal hidden polymorphism not detected with other markers and methods. SNP markers have been regarded by many as the markers of choice in the future.

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.

Trend of DNA Marker Technologies

DNA marker technologies become essential for aquaculture genetics research and the genetic improvement of aquaculture species. As a matter of fact, DNA markers, both the quality and quantity, have always been a limiting factor for in-depth genome research. Throughout the years, aquaculture geneticists have used various markers including allozyme markers, mitochondrial markers, RFLP markers, RAPD, AFLP, microsatellites, and SNPs. The overall trend, however, has been driven by (1) the need for large numbers of markers for high density coverage of the genomes and (2) the need for sequence-tagged markers for comparative genome analysis. Such demands have driven aquaculture genetic research away from using systems that do not offer a great number of markers such as RFLP and allozyme markers, and away from anonymous dominant markers such as RAPD and AFLP. Microsatellites, being codominant and sequence-tagged, have recently become very popular. However, with the draft genome sequence very soon becoming available for major aquaculture species, microsatellites are not without limitations. Their genotyping can be multiplexed, but the extent of multiplexing is limited. Automation of microsatellite genotyping is limited, thus prohibiting large-scale genome-wide applications. Mapping of thousands of microsatellites to the genome is a lot of work, and analysis using tens or hundreds of thousands of microsatellites would be a daunting task, if not technically impossible, for repeated analysis. This only leaves the SNP marker system to be viable. SNPs are the most abundant in genomes when compared with any other types of markers; SNPs are sequence-tagged and therefore would allow comparative mapping analysis; SNP genotyping is highly automated and therefore is adaptable to large-scale genome-wide analysis. Therefore, it is clear that SNP markers are the choice marker of the future. In spite of the current lack of draft whole genome sequences for many aquaculture species, it is anticipated that they will soon become available for major aquaculture species. In addition, the availability of next generation sequencing technologies makes it unnecessary to have the whole genome draft sequences in order to develop a large number of SNP markers.

Assessment of the Usefulness of Various Markers for Genome-based Selection

The following are the characteristics of the markers suitable for genome-wide applications and genome-based selection:

1. The markers should provide the genome coverage as desired for the traits, whether that is a robust use of huge number of markers across the entire genome, or a subset of the markers previously identified to be relevant for the traits.
2. The markers should provide a uniform coverage of the genome in terms of inter-marker distances.
3. The markers can be genotyped with automation, and whole genome analysis is possible with just one or a limited number of genotyping analysis.

SNPs are the only marker type that are most suitable for genome-based selection as they meet the marker number test: large numbers of SNPs should be available for almost any species; they meet the genome distribution and spacing test as SNPs are very abundant and appropriate SNPs can be selected for use in genome-based selection; they meet the test of automation as many genotyping platforms are available for SNPs.

Acknowledgments

Research in my laboratory is supported by grants from the United States Department of Agriculture (USDA)'s Agriculture and Food Research Initiative Animal Genome and Genetic Mechanisms Program, USDA National Research Initiative (NRI) Basic Genome Reagents and Tools Program, Mississippi–Alabama Sea Grant Consortium, Alabama Department of Conservation, United States Agency for International Development, National Science Foundation, and US-Israel Binational Agricultural Research and Development Fund (BARD). The author would like to thank Dr. Huseyin Kucuktas for helping with drawings of the figures, and Dr. Hong Liu, Dr. Donghong Liu, Ms. Tingting Feng, and Ms. Hao Zhang for their assistance with the references.

References

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

Beckmann JS and Weber JL. 1992. Survey of human and rat microsatellites. *Genomics*, 12:627–631.

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

Campbell D and Bernatchez L. 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*, 21:945–956.

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

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

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

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

Edwards YJK, Elgar G, Clark MS, and Bishop MJ. 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*, 278:843–854.

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

Felip A, Martinez-Rodriguez G, Piferrer F, Carrillo M, and Zanuy S. 2000. AFLP analysis confirms exclusive maternal genomic contribution of meiogynogenetic sea bass (*Dicentrarchus labrax* L.). *Mar Biotechnol*, 2:301–306.

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

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

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

Hunter RL and Markert CL. 1957. Histochemical demonstration of enzymes separated by zone electrophoresis in starch gels. *Science*, 124:1294–1295.

Jones AG, Rosenqvist E, Berglund A, and Avise JC. 1999. Clustered microsatellite mutations in the pipefish *Syngnathus typhle*. *Genetics*, 152:1057–1063.

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

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

Karsi A, Cao D, Li P, Patterson A, Kocabas A, Feng J, Ju Z, Mickett KD, and Liu Z. 2002. Transcriptome analysis of channel catfish (*Ictalurus punctatus*): Initial analysis of gene expression and micro satellite-containing cDNAs in the skin. *Gene*, 285:157–168.

Kijas JMH, Fowler JCS, Garbett CA, and Thomas MR. 1994. Enrichment of microsatellites from the citrus genome using biotinylated oligonucleotide sequences bound to streptavidin-coated magnetic particles. *Biotechniques*, 16:656–662.

Kocher TD, Lee WJ, Sobolewska H, Penman D, and McAndrew B. 1998. A genetic linkage map of a cichlid fish, the tilapia (*Oreochromis niloticus*). *Genetics*, 148:1225–1232.

Kucuktas H and Liu Z. 2007. Allozyme and mitochondrial markers. In: *Aquaculture Genome Technologies*, edited by Z Liu. Blackwell Publishing, Ames, IA, pp. 73–85.

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

Leclerc D, Wirth T, and Bernatchez L. 2000. Isolation and characterization of microsatellite loci in the yellow perch (*Perca flavescens*), and cross-species amplification within the family Percidae. *Mol Ecol*, 9:995–997.

Li WH. 1997. Genome organization and evolution. In: *Molecular Evolution*, edited by WH Li. Sinauer Associates, Inc, Sunderland, MA.

Li L and Guo XM. 2004. AFLP-based genetic linkage maps of the Pacific oyster *Crassostrea gigas* Thunberg. *Mar Biotechnol*, 6:26–36.

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

Liu H, Jiang Y, Wang S, Ninwichian P, Somridhivej B, Xu P, Abernathy J, Kucuktas H, and Liu ZJ. 2009. Comparative analysis of catfish BAC end sequences with the zebrafish genome. *BMC Genomics*, 10:592.

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

Liu ZJ. 2007a. Marking the genome: Restriction fragment length polymorphism (RFLP). In: *Aquaculture Genome Technologies*, edited by ZJ Liu. Blackwell Publishing, Ames, IA, pp. 11–20.

Liu ZJ. 2007b. Random amplified polymorphic DNA (RAPD). In: *Aquaculture Genome Technologies*, edited by ZJ Liu. Blackwell Publishing, Ames, IA, pp. 21–28.

Liu ZJ. 2007c. Amplified fragment length polymorphism (AFLP). In: *Aquaculture Genome Technologies*, edited by ZJ Liu. Blackwell Publishing, Ames, IA, pp. 29–42.

Liu ZJ. 2007d. Single nucleotide polymorphism (SNP). In: *Aquaculture Genome Technologies*, edited by ZJ Liu. Blackwell Publishing, Ames, IA, pp. 59–72.

Liu ZJ and Cordes JF. 2004. DNA marker technologies and their applications in aquaculture genetics (vol 238, pg 1, 2004). *Aquaculture*, 242:735–736.

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

Liu ZJ, Li P, Kocabas A, Karsi A, and Ju ZL. 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 Commun*, 289:317–324.

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

Lyall JEW, Brown GM, Furlong RA, Fergusonsmith MA, and Affara NA. 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*, 21:4641–4642.

MacKiewicz M, Fletcher DE, Wilkins SD, DeWoody JA, and Avise JC. 2002. A genetic assessment of parentage in a natural population of dollar sunfish (*Lepomis marginatus*) based on microsatellite markers. *Mol Ecol*, 11:1877–1883.

May B. 2003. Allozyme variation. In: *Population Genetics: Principles and Applications for Fisheries Scientists*, edited by EM Hallerman. American Fisheries Society, Bethesda, MD, pp. 23–36.

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

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

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

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

Okumuş Ý and Çiftci Y. 2003. Fish population genetics and molecular markers: II. Molecular markers and their applications in fisheries and aquaculture. *Turk J Fish Aquat Sci*, 3:51–79.

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

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

Rico C, Rico I, and Hewitt G. 1996. 470 million years of conservation of microsatellite loci among fish species. *Proc Biol Sci*, 263:549–557.

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

Rogers SM, Campbell D, Baird SJ, Danzmann RG, and Bernatchez L. 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*, 111:25–41.

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

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

Simmons M, Mickett K, Kucuktas H, Li P, Dunham R, and Liu ZJ. 2006. Comparison of domestic and wild channel catfish (*Ictalurus punctatus*) populations provides no evidence for genetic impact. *Aquaculture*, 252:133–146.

Smith MH and Chesser RK. 1981. Rationale for conserving genetic-variation of fish gene pools. *Ecol Bull*, 13–20.

Somridhivej B, Wang SL, Sha ZX, Liu H, Quilang J, Xu P, Li P, Hue ZL, and Liu ZJ. 2008. Characterization, polymorphism assessment, and database construction for microsatellites from BAC end sequences of channel catfish (*Ictalurus punctatus*): A resource for integration of linkage and physical maps. *Aquaculture*, 275:76–80.

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

Steele CA, Wheeler PA, and Thorgaard GH. 2008. Mitochondrial and maternal effects on growth in clonal rainbow. Plant and Animal Genome Conference XVI, San Diego, CA.

Steinberg EK, Lindner KR, Gallea J, Maxwell A, Meng J, and Allendorf FW. 2002. Rates and patterns of microsatellite mutations in pink salmon. *Mol Biol Evol*, 19:1198–1202.

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

Tautz D. 1989. Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Res*, 17:6463–6471.

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

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

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, et al. 1995. AFLP: A new technique for DNA fingerprinting. *Nucleic Acids Res*, 23:4407–4414.

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

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

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

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

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

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

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

Zhang Y, Liang L, Jiang P, Li D, Lu C, and Sun X. 2008. Genome evolution trend of common carp (*Cyprinus carpio* L.) as revealed by the analysis of microsatellite loci in a gynogentic family. *J Genet Genomics*, 35:97–103.