Genome-based Biotechnologies in Aquaculture
# Table of contents

Abbreviations and acronyms iii
Abstract v

1. INTRODUCTION 1
   1.1 History of genomic research 1
   1.2 Development of genomics as a new branch of science 1
   1.3 Science demands the development of sequencing technologies 2
   1.4 Aquaculture genomics, a historical review 3

2. TRADITIONAL GENETIC BIOTECHNOLOGIES FOR AQUACULTURE 4
   2.1 Selective breeding 4
   2.2 Polyploidy 5
   2.3 Gynogenesis 6
   2.4 Androgenesis 7
   2.5 Sex reversal 7
   2.6 Gene transfer 7

3. DNA MARKER TECHNOLOGIES 9
   3.1 History of DNA marker technologies 9
   3.2 Genomic variations as the basis of polymorphism 10
   3.3 Allozyme markers 10
   3.4 Restriction fragment length polymorphism markers 11
   3.5 Mitochondrial DNA markers 11
   3.6 DNA barcoding 12
   3.7 RAPD markers 13
   3.8 Amplified fragment length polymorphism markers 14
   3.9 Microsatellite markers 15
   3.10 SNP markers 17
   3.11 Restriction site-associated DNA sequencing (RAD-) markers 19

4. GENOME MAPPING TECHNOLOGIES 20
   4.1 Genetic linkage mapping of aquaculture genomes 20
   4.2 Physical mapping of aquaculture genomes 21
   4.3 Radiation hybrid mapping 23
   4.4 Optical mapping 24
   4.5 Integration of physical and linkage maps 24

5. GENOME SEQUENCING TECHNOLOGIES 24
   5.1 First-generation DNA sequencers 24
   5.2 Second-generation (the next generation) sequencers 25
   5.3 Illumina sequencers 25
   5.4 Third-generation DNA sequencers 26
   5.5 Application of next-generation sequencing 26
   5.6 Genome sequencing in aquaculture species 26
   5.7 Genome annotation of aquaculture species 28
6. TRANSCRIPTOME ANALYSIS OF AQUACULTURE SPECIES 28
   6.1 Development of EST resources of aquaculture species 29
   6.2 RNA-seq technologies 31
   6.3 Analysis of differentially expressed genes using RNA-seq 32
   6.4 Analysis of co-expressed genes using RNA-seq data sets 33
   6.5 Gene ontology, enrichment analysis and pathway analysis 33
   6.6 Analysis of allele-specific expression 33

7. UNDERSTANDING THE NON-CODING PORTIONS OF THE GENOMES 34
   7.1 Long non-coding RNAs 35
   7.2 MicroRNAs and their target genes 35
   7.3 ENCODE project and FAANG project 35

8. GENETIC ANALYSIS TECHNOLOGIES 36
   8.1 Traits important for aquaculture 36
   8.2 Quantitative trait locus mapping in aquaculture species 36
   8.3 Quantitative trait locus analysis of disease resistance 38
   8.4 Genome-wide association studies of performance traits 39

9. GENOME-BASED GENETIC SELECTION TECHNOLOGIES 39
   9.1 Marker-assisted selection 39
   9.2 Sex markers and their applications 40
   9.3 Genomic selection 40

10. GENOME EDITING TECHNOLOGIES 41
    10.1 Zinc finger nuclease and history of genome editing technologies 41
    10.2 TALEN 41
    10.3 CRISPR/Cas9 42
    10.4 Comparison of TALEN and CRISPR/Cas9 42

11. CHALLENGES FOR THE APPLICATIONS OF GENOME-BASED TECHNOLOGIES IN AQUACULTURE 43
    11.1 Discoupling of genomics with breeding programmes 43
    11.2 Bioinformatics challenges 44
    11.3 Computational limitations 44
    11.4 Funding challenges for aquaculture workers 44
    11.5 Unbalanced research advances in the world 45
    11.6 Challenges of working with small scale farmers 45
    11.7 The push for a fast return on investment 45
    11.8 Ethical and legislative issues 45
    11.9 Public perception 47
    11.10 Technology transfer 47

12. CONCLUDING REMARKS 48

References 50
### Abbreviations and acronyms

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFLP</td>
<td>amplified fragment length polymorphism</td>
</tr>
<tr>
<td>ASE</td>
<td>allele-specific expression</td>
</tr>
<tr>
<td>BAC</td>
<td>bacterial artificial chromosome</td>
</tr>
<tr>
<td>BOLD</td>
<td>Barcode of Life Data System</td>
</tr>
<tr>
<td>Cas9</td>
<td>CRISPR-associated protein-9 nuclease</td>
</tr>
<tr>
<td>cDNA</td>
<td>complementary DNA</td>
</tr>
<tr>
<td>CNV</td>
<td>copy number variation</td>
</tr>
<tr>
<td>COI</td>
<td>cytochrome oxidase I</td>
</tr>
<tr>
<td>CRISPR</td>
<td>clustered regularly interspaced short palindromic repeats</td>
</tr>
<tr>
<td>CRISPR/Cas</td>
<td>clustered regularly interspaced short palindromic repeats-CRISPR associated</td>
</tr>
<tr>
<td>DSB</td>
<td>double-strand break</td>
</tr>
<tr>
<td>EST</td>
<td>expressed sequence tag</td>
</tr>
<tr>
<td>FDA</td>
<td>Food and Drug Administration</td>
</tr>
<tr>
<td>F1</td>
<td>first filial generation</td>
</tr>
<tr>
<td>F2</td>
<td>second hybrid generation</td>
</tr>
<tr>
<td>FPC</td>
<td>fingerprinted contigs</td>
</tr>
<tr>
<td>GMO</td>
<td>genetically modified organism</td>
</tr>
<tr>
<td>GO</td>
<td>gene ontology</td>
</tr>
<tr>
<td>GWAS</td>
<td>genome-wide association study</td>
</tr>
<tr>
<td>HPC</td>
<td>high-performance computer clusters</td>
</tr>
<tr>
<td>IPR</td>
<td>intellectual property rights</td>
</tr>
<tr>
<td>kb</td>
<td>kilobase</td>
</tr>
<tr>
<td>IncRNA</td>
<td>long non-coding RNA</td>
</tr>
<tr>
<td>INDELS</td>
<td>insertions or deletions</td>
</tr>
<tr>
<td>IPN</td>
<td>infectious pancreatic necrosis</td>
</tr>
<tr>
<td>LG</td>
<td>linkage group</td>
</tr>
<tr>
<td>MAS</td>
<td>marker-assisted selection or marker-aided selection</td>
</tr>
<tr>
<td>miRNA</td>
<td>microRNA</td>
</tr>
<tr>
<td>mRNA</td>
<td>messenger RNA</td>
</tr>
<tr>
<td>mtDNA</td>
<td>mitochondrial DNA</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Full Form</td>
</tr>
<tr>
<td>--------------</td>
<td>-----------</td>
</tr>
<tr>
<td>NCBI</td>
<td>National Center for Biotechnology Information</td>
</tr>
<tr>
<td>NHEJ</td>
<td>non-homologous end joining</td>
</tr>
<tr>
<td>ORF</td>
<td>open reading frame</td>
</tr>
<tr>
<td>QTL</td>
<td>quantitative trait locus</td>
</tr>
<tr>
<td>PCR</td>
<td>polymerase chain reaction</td>
</tr>
<tr>
<td>RAD-seq</td>
<td>restriction site-associated DNA sequencing</td>
</tr>
<tr>
<td>RAPD</td>
<td>random amplified polymorphic DNA</td>
</tr>
<tr>
<td>RFLP</td>
<td>restriction fragment length polymorphism</td>
</tr>
<tr>
<td>RNA</td>
<td>ribonucleic acid</td>
</tr>
<tr>
<td>RNA-seq</td>
<td>ribonucleic acid sequencing</td>
</tr>
<tr>
<td>RH</td>
<td>radiation hybrid</td>
</tr>
<tr>
<td>sgRNA</td>
<td>single guide RNA</td>
</tr>
<tr>
<td>siRNA</td>
<td>short interfering RNA</td>
</tr>
<tr>
<td>SNP</td>
<td>single nucleotide polymorphism</td>
</tr>
<tr>
<td>TALEN</td>
<td>transcription activator-like effector nucleases</td>
</tr>
<tr>
<td>ZFN</td>
<td>zinc finger nuclease</td>
</tr>
</tbody>
</table>
Abstract

Genome technologies have been developed to study genome structure, organization, expression and function, and to select and modify genomes of interest to increase benefits to humans. Of these genome technologies, DNA marker technologies have been intensely used to map the genome to understand genome structure and organization. These DNA marker technologies include restriction fragment length polymorphism markers; mitochondrial DNA markers; DNA barcoding; random amplified polymorphic DNA markers; amplified fragment length polymorphism markers; microsatellite markers; single nucleotide polymorphism (SNP) markers; and restriction site-associated DNA sequencing markers (SNP markers per se). Although these marker systems have been used at various levels for various purposes, the microsatellite markers and SNP markers are currently the most important. Various genome mapping technologies have been developed, including both genetic mapping and physical mapping methods. Genetic mapping is based on recombination during meiosis, while physical mapping is based on fingerprints of DNA segments. Although several variations of physical mapping methods are available, such as radiation hybrid mapping and optical mapping, the most popular physical mapping method is the bacterial artificial chromosome based fingerprinting.

The most dramatic event in the genomic sciences is the invention of the next-generation sequencing technologies. The second- and third-generation sequencing technologies literally revolutionized the way science is conducted. These technologies now allow sequencing of the whole genome de novo, or mass sequencing of genomes of populations. Extension of their application allows characterization of the transcriptomes and the non-coding portions of the genome and the functions.

By coupling genome mapping technologies with aquaculture trait evaluations, quantitative trait locus (QTL) mapping allows the identification of genes underlining the performance and production traits. Following the mapping of QTLs, marker-assisted selection or genomic selection can be conducted. Genomes can be edited or modified almost any way now as designed by scientists. Therefore, technologies are sufficiently mature to make some really large contributions for improving aquaculture traits. However, there are a number of challenges, which include bioinformatics challenges, lack of resources in some parts of the world, difficulties in working with individual farmers, and ethical and legislative challenges, that must be overcome in order to have broad applications of genome technologies.
1. INTRODUCTION

1.1 History of genome research
The discussion of sequencing the entire human genome started approximately in the mid-1980s. Such a task was daunting at the time because of the limitations of sequencing technologies. The most efficient automated DNA sequencer was able to run 96 samples at a time, and approximately 500-800 base pairs of nucleotides could be read per sample. The sequencing reactions themselves were not fully automated and were expensive. With this technology, it was estimated that the project could take at least 15 years to complete. With many planning activities involving government and scientific communities with various workshops, the project was officially launched in 1990 with a single goal of sequencing and assembling the entire genome of a single human individual containing three billion base pairs of DNA. The project started in the United States, but became an international collaborative project involving the participation of universities and research centres in China, France, Germany, Japan, the United Kingdom of Great Britain and Northern Ireland, and the United States of America.

It should be noted that a parallel project was conducted by a private corporation, the Celera Corporation, which formally launched in 1998. It is widely believed that this parallel run of the project from the private sector significantly increased competition, allowing speeding up of the project as well as enhancement of the sequence quality.

1.2 Development of genomics as a new branch of science
Into the mid-phase of the 13-year human genome project, it was becoming evident that it would not be possible to put the human genome together with just DNA sequences generated with sequencing technologies. This was because the vast majority of the three billion base pairs of human DNA are repetitive in nature. These repetitive elements of the genome are different, but highly similar sequences. Computer software would not lend the capability to assemble all such sequences correctly into their proper place. Therefore, the demands on the development of additional technologies and research to meet this challenge were tremendously high. A new branch of science, genomics, was born. The scope of genomics is to apply the techniques of genetics and molecular biology for the understanding of genome structure, organization, expression and functions. Starting from its beginnings, genomics was meant to be a branch of science that uses principles and methods of multidisciplines, including those of biochemistry, molecular biology, genetics, statistics and informatics, to name a few. It is clear, as detailed below, that genomics is now the centre of life sciences. Just as the “molecularization” of biology in the past 40 years, now many scientific disciplines have been or are being "genomicized". This is largely because genomics has led to major changes on how science is conducted. Rather than on the traditional process of proposing a hypothesis based on existing knowledge, testing the hypothesis experimentally, and accepting or rejecting the hypothesis as the route of scientific research, genomics and its related large data sets have allowed the change from hypothesis-driven research to data-driven research. All the data can be suddenly created, and it is the researchers’ role to figure out the simple facts from the complex data.

As progress of the human genome project was being made, it became obvious that sequences themselves can be assembled by extension of the overlapping DNA sequences into longer contiguous sequences (contigs). However, the size of the contigs is limited by the presence of the repetitive elements. Thus, the human genome sequences could be assembled into many - hun-
dreds of thousands - small contigs. Although the sequences within the contigs were known, the relationship or the positions of the contigs in the genome could not be determined by sequence assembly. This demanded the development of methods to place and order the contigs onto the chromosomes and eventually to the genome. This was the start of the development of genome mapping technologies, including genetic linkage mapping, physical mapping, radiation hybrid mapping, and optical mapping. Within the area of physical mapping, several technologies were developed, with the most popular being radiation hybrid mapping and large-insert libraries such as BAC (bacterial artificial chromosome)-based fingerprinting for mapping the physical genomic DNA. Similar to the situation of the human genome project, genome projects of aquaculture species require similar technological support to assemble billions of base pairs of DNA together correctly. These mapping technologies will be discussed in detail below.

1.3 Science demands the development of sequencing technologies

The sequencing difficulties and the involved costs have demanded rapid development of sequencing technologies. Principles for many sequencing technologies were known even at the beginning of the human genome project, but the practicalities were not possible because of the limitations of manufacture limitations of minimizing the reactions allowing many reactions to happen within a single machine simultaneously. Two lines of technological advances had the largest push on sequencing technologies. One was the advances in polymerase chain reaction, and the other one the advances in nano technologies.

Genomic sciences have made drastic advances in the past ten years, largely because of the application of next-generation sequencing technologies. It is not just the high throughput that has revolutionized the way science is conducted, but the rapidly reducing cost for sequencing has also made the technologies applicable to all aspects of molecular biological research, as well as to all organisms including aquaculture and fisheries species. Twenty some years ago, Dr Francis Collins, currently the director of the National Institutes of Health, had a vision of achieving the sequencing of one genome with US$1 000, and that is almost possible now. From the billion-dollar project of the human genome project, to the genome projects of agriculture animals with a budget of ten million dollars or so, down to a million dollars just a few years ago, to the current cost level of just tens of thousands of dollars for a de novo sequencing project, the potential for research using genomic approaches is limitless. Today, commercial services are available worldwide for genome sequencing projects, whether they are new sequencing projects for a species or resequencing projects for many individuals. The key issue is to achieve a balanced output of quality and quantity with minimal costs.

Rapid technological advances provide huge opportunities to apply modern genomics to enhance aquaculture production and performance traits. However, the world is facing a number of new challenges, especially in the area of bioinformatics. This challenge may be paramount for aquaculture researchers and educators. Although aquaculture students are well educated in aquaculture, they may have no background in computer science or lack the knowledge of bioinformatics analysis of large data sets. The large data sets in tera-scales themselves pose great computational challenges. Therefore, new ways of thinking are required in terms of education and training of the next generation of scientists. For instance, a few laboratories in the world may be sufficient for the production of data, but an infinite number of laboratories may be required for data analysis or bioinformatics data mining to link data with biology.
1.4 Aquaculture genomics, a historical review

Aquaculture is a relatively new but increasingly growing sector of agriculture. It is very important not only for economic interests, but is also of social and cultural significance. In many Asian countries, serving seafood at the dinner table is seen as highly prestigious. Seafood accounts for 20 percent of animal protein sources for the world population. The seafood industry involves a total of 144 million tonnes annual production with 44 million fishers and fish farmers and 2.1 million vessels, contributing US$166 billion to the world economy and over US$25 billion of international trade annually. Currently, aquaculture accounts for over 40 percent of seafood consumed in the world, and China is the only country where aquaculture produces more than 50 percent of consumed seafood (FAO, 2016).

Aquaculture genomics officially started in the 1990s, although related genome research was conducted in the 1980s. The start of aquaculture genomics was marked by the first Aquaculture Genomics Workshop, held in 1997 in Dartmouth, Massachusetts, United States of America. This workshop targeted a group of six species for genome research in the United States: salmonids, catfish, tilapia, striped bass, oysters and shrimps.

As any other agricultural sector, sustained production requires research on basic biology, including growth, nutrition, reproduction, physiology, and genetics and genomics. One interesting observation is that all these research fields are being unified through the use of genomic technologies or, that is, they are being “genomicized”.

The availability of a draft whole genome sequence significantly enhances genome research and applications of genome-based technologies to improve agricultural production and quality. As such, whole genome sequences are available or nearly available for major livestock species, including cattle, swine, chicken and horse. Whole genome sequencing used to be nearly “utopian” because of its involved huge costs. Take the human genome as an example; it took the major genome centres over a decade to complete the draft genome sequencing of the first human individual, from 1986 to 2000. The direct cost involved in the human genome sequencing was at the level of hundreds of millions of dollars. Along with the sequencing effort, related human genome research had cost much more. Human genome related research probably spent in excess of US$6 billion–US$8 billion. Thereafter, the sequencing of the cattle genome cost over US$60 million, much less expensive than the human genome sequencing, but yet is still at a cost level unthinkable for many aquaculture species. It was because of such huge costs whole genome sequencing used to be regarded as a milestone of a lifetime achievement.

The daunting task of whole genome sequencing has now become possible for many species, including for many aquaculture species. With the second- and third-generation sequencing technologies, the cost of sequencing a genome with a size of one billion base pairs has reduced to manageable levels, usually about US$100 000. With such a major reduction in costs, the “sequencing rush” is on the way for many species.

The rapid advances in sequencing technologies made science road maps and plans almost meaningless. In 2008, the United States Department of Agriculture (USDA) spent much energy developing the Blueprint for USDA Efforts in Agricultural Animal Genomics 2008–2017. Among many aquaculture species, the blueprint calls for the production of a draft genome sequence for catfish with only 6X genome coverage and a much lower coverage for several other aquaculture species,
including rainbow trout, tilapia and shrimps. As short as two years later, the drastic reduction in sequencing costs made it possible to sequence the whole genome of many aquaculture species.

Genomes of at least two dozens of aquaculture species have been sequenced, or are now being sequenced. Of the six aquatic species groups included in the United States Animal Genome NRSP-8 Program, a whole genome has been sequenced for tilapia, rainbow trout, Atlantic salmon, catfish, striped bass, and oysters and shrimps. While it is extremely exciting for aquaculture geneticists to have the whole genome sequenced for many of the important aquaculture species, it is pivotally important to address some of the most important issues related to whole genome sequencing. These include the issues related to effective assembly of the whole genome sequences and the issues related to the effective uses of the whole genome sequences. The usefulness of the whole genome sequence relies on the efficient sequence assembly and adequate sequence annotation, which in turn depends on the availability of a number of genome resources, including transcriptome sequencing and assembly and characterization of the non-coding part of the genome.

This document provides the basic concept, the descriptions of the technologies, and their application or potential application to aquaculture species. It begins with a brief description of traditional genetic biotechnologies for aquaculture, followed with DNA marker technologies, genome mapping technologies, genome sequencing technologies, transcriptome analysis of aquaculture species, understanding the non-coding portions of the genomes, genetic analysis technologies, genome-based genetic selection technologies, and genome editing technologies. At the end, the document discusses some challenges for the applications of genome-based technologies in aquaculture.

2. TRADITIONAL GENETIC BIOTECHNOLOGIES FOR AQUACULTURE

2.1 Selective breeding

The goal of selective breeding is to improve production and performance traits in farmed animals via the selection of individuals for desirable phenotypic characteristics. For aquaculture species, these traits may include faster growth, feed conversion efficiency, disease resistance, processing yields, low oxygen tolerance, stress tolerance, robustness, morphology and sexual maturation, among other traits of interest.

Aquaculture breeding programmes have a short history. Selective breeding programmes for strain enhancement began in the late 1960s (Gjedrem and Baranski, 2010) with the development of Atlantic salmon breeding programmes in Norway. Domestication of many aquaculture species is still a recent event. Therefore, a number of techniques are effective, including strain selection, cross breeding, hybridization, and within-strain selection. However, the main goal of genetic improvement programmes for production traits in modern breeding schemes is within-strain selection. This is typically achieved within a well-managed, commercial programme of family and pedigree tracking combined with extensive trait measurements on selection candidates or their relatives (Gutierrez and Houston, 2016). Selective breeding programmes exist for various aquaculture species, such as Atlantic salmon, rainbow trout, tilapia, common carp, grass carp, silver carp, crucian carp, rohu carp, yellowtail, sea bream, channel catfish, European seabass,
Selective breeding has proven to be very effective in enhancing the traits of agricultural plants and animals, including aquaculture species, through the use of genetics principles. For instance, the genetic gain has been greater than 12 percent per generation for growth rate and for disease resistance when challenge tests are applied (Gjedrem and Robinson, 2014). The main reasons for the large genetic gains observed for aquatic species are their relatively high fertility and the natural existence of broad genetic variation for economically important traits, both of which allow a very high selection intensity to be applied. However, the genetic improvement of species grown in aquaculture is far behind that achieved for animals and plants. Less than 10 percent of aquaculture production is based on genetically improved stocks (Gjedrem and Robinson, 2014). Reviews and books are available that cover selective breeding in great detail (e.g. Hulata 1995, 2001; Stickney, 1994; Gjedrem and Robinson, 2014; Gjedrem and Baranski, 2010; Gjedrem, Robinson and Rye, 2012), and therefore this subject area is not the focus of this review.

Advances in sequencing technology and genomics have significantly improved the tools available for the genetic improvement of livestock. In particular, the development of genetic markers and linkage maps has permitted great advances in the quantitative analyses of commercially important traits. Some of the genome-based technologies in relation to selection will be covered in various sections, such as the analysis of quantitative trait locus, marker-assisted selection and whole genome selection.

2.2 Polyploidy

Polyploidy is lethal in mammals and birds (Chourrout et al., 1986), but has led to the development of many productive plant varieties such as domesticated wheat. Triploid fish are viable (Thorgaard, Jazwin and Stier, 1981; Wolters, Libey and Chrisman, 1981; Chourrout, 1984; Cassani and Caton, 1986) and are usually sterile, while tetraploid fish are usually viable and fertile (Dunham, 2011). Through chromosome manipulations, aquaculture organisms can be obtained with various levels of chromosome sets. A regular diploid organism has two sets of chromosomes. The polyploid state refers to individuals with extra sets of chromosomes (Dunham, 2011). Triploid fish have three sets of chromosomes, and tetraploid fish have four sets of chromosomes. In addition to full sets of chromosomal compositions, aneuploids have at least a diploid set of chromosomes with one or more additional chromosomes, but not a full complement to the set (Dunham, 2011).

Triploids are organisms with three sets of homologous chromosomes. They are found spontaneously in both wild and cultured populations and can be induced in many commercial species of fish and shellfish. Triploidy is induced by allowing normal fertilization and then forcing retention of the second polar body (Chourrout, 1980, 1984; Lou and Purdom, 1984). The second polar body is retained by applying temperature (hot shocks or cold shocks), hydrostatic pressure, anaesthetics or chemical shocks shortly after fertilization (Dunham, 2011). Additionally, in some cases such as in rainbow trout, triploidy can be induced by applying high pH and high calcium to either the sperm or eggs (Ueda, Sato and Kobayashi, 1988). Many reviews and books are available covering the technical aspects of polyploidy, and therefore this paper will not cover this extensively.
The performance of triploid fish varies. Triploidy can affect growth, feed conversion efficiency, disease resistance and other traits. For growth, triploid fish can grow faster, at the similar rate, or slower. However, even for those that grow faster, this advantage is not obvious until sexual maturity. It is apparent in sterile fish that energy otherwise expended on development of gonads and gametes is converted into growth.

In addition to direct induction of triploid fish and shellfish such as oysters, the production of tetraploid fish has been a way for mass production of triploid fish (Chourrout et al., 1986). First-generation tetraploids can be produced by hydrostatic pressure treatment before the first cleavage and raised until the adult stage. With rainbow trout, survival and growth of tetraploids is severely depressed when compared with the diploid control (Chourrout et al., 1986). However, when the tetraploid male fish were mated with normal diploid female fish to produce triploid fish, the progenies were consistently normal with high survival rates. The progenies were found to be almost all triploids by karyology, which failed to detect a significant rate of aneuploidies. Therefore, the use of tetraploid as a means of producing triploid fish is a viable option, although the fertilizing ability of tetraploid males was always low, 0 to 97 percent of the control, with an average of 40 percent compared with the diploid controls (Chourrout et al., 1986).

A review by Piferrer et al. (2009) covers all the details for triploid production and the performance of triploid fish and shellfish as compared with the diploid counterparts. Interested readers are referred to this review article.

2.3 Gynogenesis

Gynogenesis is a form of all-female inheritance. In fish species, ultraviolet (UV) irradiation has been used to inactivate the sperms, and such UV-inactivated sperms are used to trigger gynogenetic development without contributing the paternal genome to the progeny. Practically, sperms from a closely related, but different species, are used to reduce the possibility of real fertilization in case the sperms are not completely inactivated (Arai, 2001; Suwa, Arai and Suzuki, 1994). The developing embryo is initially haploid, but the diploidy is recovered by inhibiting either the second polar body extrusion or the first cleavage. The gynogen produced by inhibiting the second polar body extrusion is called meiotic gynogen, while the gynogen produced by inhibiting the first cleavage is called mitotic gynogen. Mitotic gynogens are completely homozygous, while meiotic gynogens are partially heterozygous. The gene-centromere recombination makes the distal part of the chromosome from the centromere more heterozygous. Meitic gynogens are easier to produce than the mitotic gynogens because the survival rate of mitotic gynogens is very low.

One of the practical goals of gynogenesis is the production of genetically identical populations, i.e. the clonal lines (Arai, 2001). The clonal lines may be important for breeding programmes, and they can certainly be important for genome research because they provide a homozygous template for genome sequencing, such as the doubled haploid of channel catfish (Waldbieser, Bosworth and Quinioiu, 2010) that was used as the sequencing template for whole genome sequencing (Liu et al., 2016). Although clonal lines have been produced with aquaculture species such as ayu (Taniguchi, et al., 1996), amago salmon (Kobayashi et al., 1994) and hirame (Yamamoto, 1999), their large-scale aquaculture has not been realistic. The major purpose of gynogen production has been for research.
2.4 Androgenesis
Androgenesis refers to all-paternal inheritance. Androgens can be produced by irradiating eggs and then doubling the paternal genome. Androgens are more difficult to produce than gynogens (Scheerer et al., 1986), presumably because of the extremely low survival rate of irradiated eggs. Diploidy can be recovered only by blockage of first cell cleavage (Dunham, 2011).

Like gynogenesis, androgenesis can be used to produce clonal populations or monosex populations for the purpose of breeding programmes or to elucidate sex-determining mechanisms. If the male is heterogametic, XX and YY androgens will result in equal proportion. If the male is the homogametic, then the androgens will be 100 percent ZZ and all male. YY individuals are viable. As a matter of fact, the production of YY fish through androgenesis followed by regular mating with a normal XX female is a major way of producing all male populations in fish (Dunham, 2011; Parsons and Thorgaard, 1985). In many species, males grow faster, and use of monosex populations is of interest to aquaculturists.

2.5 Sex reversal
Sex bimorphism for growth is very common with aquaculture fish species. In some cases, males grow faster, while in other cases females grow faster. Apparently, the generation of a monosex population for faster growth is of interest to aquaculturists. In addition to chromosomal manipulations such as polyploidy, gynogenesis and androgenesis, monosex populations can also be created by hormonal treatment. Although genotypic sex is established at the time of fertilization, the phenotypic sex is not determined until later in development. For instance, channel catfish phenotypic sex is determined at around 19 days after fertilization. Among various approaches for sex reversal, hormonal sex reversal is the most used approach for aquaculture.

The phenotypic sex can be altered by administration of estrogens or androgens during the critical period of sex determination. Several androgens, most of them derivatives of testosterone, have been used to produce monosex male populations (Yamazaki, 1983; Dunham, 1990). For instance, 17-methyltestosterone (Dunham, 1990) is widely used for sex reversal in fish. Several estrogenic compounds have been used to produce monosex female populations, of which 3-estradiol is the most commonly used hormone for feminization (Yamazaki, 1983; Dunham, 1990). The hormonal treatment can be applied by bath soaking (Donaldson and Hunter, 1982; Yamazaki, 1983), in feed (Shelton, Rodriguez-Guerrero and Lopez-Macias, 1981), or through implants (Boney et al., 1984), depending on the developmental and culture characteristics of the species (Dunham, 2011). Readers interested in the technical details for the production of monosex population through sex reversal are referred to Dunham (2011).

2.6 Gene transfer
Gene transfer is a process of transferring one or a few foreign gene(s) into an organism. However, the foreign gene can be from other organisms or from the organism itself. The concept is relatively straightforward: if the functions of a gene are well known, then the gene can be transferred into the organism to deliver the functions. For instance, the growth hormone gene was well studied and known of its functions to be promoting growth, then the transfer of the growth hormone into an organism should promote growth. This simple concept was demonstrated when transgenic mice with the growth hormone gene grew much larger, up to 2.5 times larger than the nontransgenic controls (Palmiter et al., 1982).
The first successful gene transfer in fish was demonstrated in 1985 when Zhu, He and Chen transferred the human growth hormone gene in goldfish. Since then, transgenic fish have been produced with various aquaculture species, including rainbow trout (Chourrout et al., 1986), channel catfish (Dunham et al., 1987), Nile tilapia (Brem et al., 1988) and northern pike (Gross et al., 1992).

A number of techniques were developed for transferring the genes of interest into fish, including microinjection (Zhu et al., 1985) and electroporation (Inoue et al., 1990; Powers et al., 1992). Either method works well. However, transgenic technologies suffer from several major lines of shortcomings: (i) the doses of gene transfer cannot be controlled. Most often, multiple copies, often in the form of tandem head to tail arrays are integrated into the genome; (ii) the integration sites are random, and such sites can be within a functional gene; and (iii) the pleiotropic effect of genes cannot be controlled. The major concern of transgenic fish was the use of foreign DNA, including the promoters and the regulatory DNA sequences such as poly A signals. Liu et al. (1990) introduced the concept of “all-fish vectors” that contain all sequences from fish for transfer into fish.

Transgenic fish have generally performed the way scientists expected them to. For instance, significantly enhanced growth rates were observed with transgenic fish with growth hormone genes. This has been demonstrated in goldfish, channel catfish, northern pikes, Atlantic salmon, rainbow trout, tilapia and carp, among many other species (reviewed by Dunham, 2011), although failure to observe enhanced growth was also reported (Guyomard et al., 1989; Penman et al., 1991). The best example of growth enhancement was demonstrated with transgenic salmon.

In addition to the growth trait, gene transfer was used to improve several other traits, including cold tolerance (e.g. Fletcher and Davies, 1991) and disease resistance (e.g. Anderson, Mourich and Leong, 1996). For instance, transgenic expression of viral coat protein genes, or antisense of viral early phase genes, has been explored in shrimp (Ahanger et al., 2014). Attempts for improving disease resistance were made by expression transgenes of lytic peptides, and organisms containing these genes exhibited enhanced disease resistance (Dunham, 2011). Antimicrobial peptide genes were also used to enhance disease resistance. For instance, transgenic fish expressing cecropins were found to be more resistant against bacterial diseases (Dunham et al., 2002).

In addition to the purpose of enhancing performance traits of aquaculture species, fish have been considered for the production of pharmaceuticals as biological factories (Dunham, 2011). Fish have potential advantages as bioreactors compared with mammals. These advantages include a short generation interval, low cost of maintenance of the animals, easy maintenance, large numbers of individuals, high-density culture, and mammalian viruses and prions that are not found in fish.

Public acceptance of transgenic fish has been relatively low because of two lines of concerns: (i) food safety concerns; and (ii) ecological safety concerns. The question on whether it is safe to consume transgenic fish has been one major question of consumers. As aquaculture species have aquatic living environments, tracking of transgenic aquatic animals is more difficult, and therefore the concerns over ecological safety have been serious. However, as reviews and books for gene transfer with aquaculture species are widely available, interested readers are referred to existing resources on the details of transgenic work with aquaculture species.
One commonality shared by all the traditional genetic technologies is that researchers are operating in a black box without the knowledge of the genome, how the genome works, and how the genetic changes made will affect the expression and functions of the genome. Modern genetic technologies depend on the knowledge of the genome, genome maps, genome composition, organization, expression and function. Therefore, in the next sections, a great level of detail will be provided to cover how genomes are being studied, biotechnologies that were developed for the understanding of the genomes, transcriptome, and gene networks and functions. Using genome information, genome-based technologies have been developed and are continuously being developed for genetic gains in aquaculture species.

3. DNA MARKER TECHNOLOGIES

3.1 History of DNA marker technologies

The development of molecular markers has been one of the major efforts in the first decade of genome research with aquacultured species (Liu and Cordes, 2004). In the early days of aquaculture genome research, most of the efforts were devoted to markers that were readily available without the availability of genome information or resources. This included the development of random amplified polymorphic DNA (RAPD) and amplified fragment length polymorphism (AFLP) markers (Bardakci and Skibinski, 1994; Naish et al., 1995; Liu et al., 1998; Liu et al., 1999; Bagley, Anderson and May, 2001). Although these efforts certainly contributed to the development of genome programmes at that time, these dominant markers are not very useful in terms of polymorphic DNA markers in the long term. Since 1997, great effort has been devoted to microsatellite development in aquaculture species (e.g. Nielsen, Crow and Founatin, 1999; Sakamoto et al., 2000; Rexroad et al., 2001; Nichols et al., 2003; Palti, Danzmann and Rexroad, 2003; Gilbey et al., 2004). Several approaches were adopted, including construction of microsatellite- enriched genomic libraries, targeted sequencing (Serapion, et al., 2004), and identification of microsatellites through data mining using genome resources such as expressed sequence tags (ESTs) or BAC-end sequences (Serapion et al., 2004; Wang et al., 2010; Xu and Guo, 2006; Li et al., 2007; Xu et al., 2006; Liu et al., 2009). Large numbers of useful microsatellites have been obtained through data mining of genome sequence surveys such as BAC-end sequences and EST data mining. However, it is apparent that such huge efforts can be avoided now with the application of next-generation sequencing technologies (see below).

The effort of marker development is shifted to the identification of single nucleotide polymorphisms (SNPs) over time (He et al., 2003; Liu et al., 2011; Sun et al., 2014; Wang et al., 2010; Van Bers et al., 2012; Xu et al., 2012; Kongchum et al., 2010; Rynänen and Primmer, 2006; Lorenz et al., 2010; Andreassen, Lunner and Høyheim, 2010; Gomez-Uchida et al., 2011; Harwood and Phillips; 2011; Hauser et al., 2011; Karlsson et al., 2011; Freamo et al., 2011; Yáñez et al., 2016). SNPs are now the markers of choice for genetic studies because they are the most abundant genetic variations widely distributed in the genome, and are generally bi-allelic polymorphisms that are amenable to automated genotyping. SNPs are efficient for genome-wide association studies (GWAS) because linkage disequilibrium can be detected with high- density SNPs when dealing with complex traits (Liu et al., 2014). Once again, the huge efforts devoted in SNP discoveries can now be easily achieved through next-generation sequencing.
3.2 Genomic variations as the basis of polymorphism

The genome compositions of each individual of fish or shellfish of the same species are similar but different at the level of DNA sequences. 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:

(i) a deletion due to the loss of one or more of DNA bases;
(ii) insertion due to gain of one or more DNA bases;
(iii) base substitution at various positions;
(iv) inversion of a DNA segment in its orientation;
(v) rearrangements of multiple DNA segments within the genome; and
(vi) copy number variation because of insertions, deletions, duplication or multiplication of a DNA segment(s) including genes.

The most widespread genomic variation among individuals within a population is base substitution. Such base substitution along the DNA chain is defined as SNPs. The second most widespread genomic variation among individuals within a population is probably deletions and insertions. A deletion mutation and an insertion mutation can be viewed as the same phenomenon depending on what is used as the reference, and therefore they are together referred to as “indels”. The molecular basis for microsatellite marker polymorphism is insertion or deletion because when the microsatellite repeats, expands or shrinks, the individuals differ in the number of repeats.

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

Copy number variation (CNV) owing to insertions, deletions and duplication or multiplication of a DNA segment is widespread, and this type of genomic variation has 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 copy number variations can involve genes or just genomic segments that do not harbour 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 programmes 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 have an additional round of genome duplication followed with random gene loss, thereby resulting in various CNV situations involving various genes.

3.3 Allozyme markers

Allozymes are protein products of genes that are encoded by a single gene locus (Kucuktas and Liu, 2007). Since they represent genes of known function, they are considered as Type I markers (Liu and Cordes, 2004). The term “isozyme” refers to multiple biochemical forms of an enzyme having identical substrate specificity (or 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 Markert, 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 catalyse the same reaction. However, the two terms are usually
used interchangeably. 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 past 30 years, a large amount of allelic frequency data has been collected from many fish species for management purposes. Although the use of allozyme data in aquaculture appears to be limited compared with population studies in fisheries, the aquaculture industry has long used this information for its development because aquaculture and fisheries cannot 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). However, because the number of allozymes are limited (approximately 100), the broad application of allozyme markers for genomic and genetic research is limited.

3.4 Restriction fragment length polymorphism markers

Restriction fragment length polymorphism (RFLP) markers (Botstein et al., 1980) were the very first generation of markers. RFLP was the most popular approach for analysis of genetic variation during the entire 1980s. As indicated by its name, RFLP is based on DNA fragment length differences after digesting genomic DNA with one or more restriction enzymes. In its original format, genomic DNA is digested by one or more restriction enzymes, separated on an agarose gel, and then followed with a procedure called Southern blot (Southern, 1975) to visualize the fragment length differences (Liu, Liu and Zhang, 2007). After sequencing and polymerase chain reaction (PCR) technologies, RFLP are nowadays analysed using PCR amplification, or PCR amplification followed by restriction analysis depending on if the size is the same.

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, nor deletions or insertions involving just a few base pairs 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 at that time. Although sequence information is widely available now with many aquaculture species, the limited use of RFLP makes it sort of obsolete.

3.5 Mitochondrial DNA markers

The mitochondrial genome evolves more rapidly than the nuclear genome. The rapid evolution of 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 (Liu, 2007a).

Mitochondrial DNA is maternally inherited for the most part, but there are reports of paternal leakage during fertilization (Birky, Fuerst and Maruyama, 1989). Because of the high levels of polymorphism and the ease of mitochondrial DNA analysis, mtDNA has been widely used as markers in aquaculture and fisheries settings. The non-Mendelian inheritance, however, greatly limits the applications of mtDNA for genome research. Nevertheless, as an identification tool
often used in aquaculture, mtDNA can be used as a supplemental tool for aquaculture genomic research.

Mitochondrial markers can be analysed using PCR or using restriction fragment length polymorphism (see RFLP markers), 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 (Okumuş and Çiftci, 2003; 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 the Xba I restriction site, causing an RFLP polymorphism that could easily be detected with 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, Monje and Juinio-Meñez, 2002). In this case, electrophoresis of PCR products in the region should provide information on mtDNA haplotypes (Billington, 2003).

Analysis of mtDNA polymorphism has become a useful genetic tool for studies of genetic divergence within and among populations (Avise, 1995). Because mtDNA shows considerable variations among individuals, it is regarded as an effective marker for population structure and geographic variations. Distinct mtDNA lineages are detected in many freshwater fishes in different parts of their species ranges. Since only half of the population (assuming 1:1 sex ratio) pass on their mtDNA to their offspring, the effective population size for mtDNA is smaller than that of nuclear DNA (Harrison, 1989).

Mitochondrial DNA markers have been used extensively to analyse genetic variations in several different aquaculture species, including striped bass (Wirgin and Maceda, 1991; Garber and Sullivan, 2006); channel catfish (Waldbieser, Bilodeau and Nonneman, 2003); walleye (Merker and Woodroff, 1996); salmonids (Nielsen, Hansen and Mensberg, 1998; Crespi and Fulton, 2004); red snapper (Pruett, Saillant and Gold, 2005); and bluegill (Chapman, 1989). Data analysis in mtDNA studies includes determining the number of mtDNA haplotypes, and calculating the haplotype frequencies and nucleotide diversity.

There are two major drawbacks of the mtDNA markers. One is the non-Mendelian inheritance of mtDNA, and the other is the proportion of the total genomic variation one can observe with mtDNA alone. Additionally, mtDNA markers are subject to the 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 the unparallel rate of heterogeneity at the same region (Liu and Cordes, 2004) all can place limitations to the validity of using mtDNA for genetic studies.

3.6 DNA barcoding

DNA barcoding involves the amplification and sequencing of a short universal molecular tag of approximately 650 base pairs from the 5’ region of the mitochondrial cytochrome oxidase I (COI) gene (Hebert et al., 2003; Tavares and Baker, 2008). DNA barcoding using COI has been widely employed in various biological fields with proven ability to differentiate closely related species in studies ranging from forensic sciences (Dawney et al., 2007) to molecular systematics (Hardman,
Importantly, a community-based effort to develop extensive DNA barcode libraries, most notably the Barcode of Life Data System (BOLD, www.boldsystems.org/index.php/databases), has led to the adoption of DNA barcoding technology as the gold standard for species identification and has greatly expanded the power of the technique. The BOLD system provides detailed information of COI-sequenced species, including the origin and current location of voucher specimens (Ratnasingham and Hebert, 2007). Out of almost 30 000 fish species estimated in the world, barcodes for more than 10 000 fish species are currently recorded in the BOLD system. These COI barcodes are gathered from several sources, including the Fish Barcode of Life Initiative (FISHBOL, 2010) and the Marine Barcode of Life (MarBOL, www.marinebarcoding.org). However, for many species, BOLD barcodes are gleaned from uncurated GenBank records and require additional validation before use.

DNA barcoding is most useful for species identification. It is simple and highly reliable. For aquaculture purposes, the most important application of DNA barcoding is for the protection of consumer interests against mislabeling of seafood. Mislabeling of global seafood products is a growing concern, particularly in Western nations that import a large percentage of fish and shellfish products. As seafood choices expand, consumers are faced with identifying products often prepared and processed in ways that remove differentiating anatomical and morphological features. Seafood exporters/importers, processors, wholesalers and restaurateurs may knowingly or unwittingly substitute one species for another as food travels through the supply chain from pond to plate (Wong et al., 2011). Seafood fraud may involve the labeling of a lower-market-value species as a higher-market-value species to realize a larger profit. In the case of international trade, seafood may be mislabeled to avoid protective tariffs, or transshipped or commingled with products from a third country to avoid import duties and/or Food and Drug Administration import alerts. Cases of such mislabeling are not uncommon in the media and scientific literature (Wong and Hanner, 2008; Hsieh, 1998; Marko et al., 2004; Miller and Mariani, 2010; Wong et al., 2011).

In spite of the value of DNA barcodes for species identification, it is not very useful for genetic and genomic studies because it is maternally inherited as a part of mitochondria.

3.7 RAPD markers
Random amplified polymorphic DNA (RAPD) is a PCR-based multi-locus 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 in length). Because the primers are short and relatively low annealing temperatures (often 36–40º C) are used, the likelihood of amplifying multiple products is great, 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 (Liu et al., 1998b, 1999b; Liu, 2007b). 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 insertions or deletions (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 values for RAPDs fall below those for microsatellites and SNPs, and RAPDs may not be as informative as AFLPs 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.

The most important applications of RAPD are for species identification, hybrid identification, strain differentiation and, to a much lesser extent, for genetic analysis such as mapping. RAPD markers were widely used in the 1990s, but over time the marker system is no longer popular because it is a dominant type of markers and dominant markers are not informative and cannot be compared between experiments or among laboratories (Liu, 2007b).

3.8 Amplified fragment length polymorphism markers

Amplified fragment length polymorphism (AFLP) is based on the selective amplification of a subset of genomic restriction fragments using PCR (Liu, 2007a). 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 subsets of amplified fragments are then analysed by denaturing polyacrylamide gel electrophoresis to generate the fingerprints.

AFLP analysis can be viewed as an advanced form of RFLP. Therefore, the molecular basis for RFLP and AFLP is similar. First, any deletions and/or insertions between the two restriction enzymes, e.g. 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 1/8 of Mse I sites are detected by AFLP since only the Eco RI primer is labelled and AFLP is designed to analyse 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 the 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 one billion base pairs), a complete AFLP scan would also examine over 2 million bases immediately adjacent to the restriction sites.

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.

AFLP has been widely used in aquaculture, such as the analysis of population structures, migration, hybrid identification, strain identification, parentage identification, genetic resource analysis, genetic diversity, reproduction contribution, and endangered species protection (Seki et al., 1999; Jorde, Palm and Ryman, 1999; Sun et al., 1999; Cardoso et al., 2000; Chong et al., 2000; Kai, Nakayama and Nakabo, 2002; Mickett et al., 2003; Whitehead et al., 2003; Mock et al., 2004; Campbell and Bernatchez, 2004; Simmons et al., 2006).

AFLP has also been widely used in genetic linkage analysis in the early days of aquaculture genomics (Liu et al., 1998a, 1999a; 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; Felip et al., 2005), and analysis of parental genetic contribution involving interspecific hybridization (Young et al., 2001) and meiogynogenesis (Felip et al., 2000).

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 markers also inflate 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. The dominant nature of AFLP fundamentally limits its broad applications for genetic analysis. In some cases, AFLP can be used as a rapid screening tool, and useful markers can then be converted to SCAR (sequence characterized amplified region) markers. However, genome scale applications of SCAR markers are unlikely. Now with the advances of next-generation sequencing, the fundamental principles of AFLP have been adopted in restriction site-associated DNA sequencing technologies (see below).

3.9 Microsatellite markers

Microsatellites are simple sequence repeats of 1–6 base pairs. They are highly abundant in various eukaryotic genomes, including all aquaculture species studied to date. In most fish genomes, the frequency of detecting microsatellites is approximately one microsatellite per 2–10 kilobases (kb) of DNA.

Dinucleotide repeats are the most abundant forms of microsatellites, of which the most abundant forms are AC and AG repeats. CG 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 tetra-nucleotide repeats, relatively A/T-rich repeat types are generally more abundant than G/C-rich repeat types. Microsatellites longer than tetranucleotide repeats (penta- and hexanucleotides) are much less abundant and therefore are less important as molecular markers (Toth, Gaspari and Jurka, 2000).

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 non-gene sequences (Toth, Gaspari and Jurka, 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 dis-
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.

The disadvantages of microsatellites as markers include the requirement for existing molecular genetic information, a large amount of up-front work for microsatellite development, and the tedious and labour-intensive nature of microsatellite primer design, testing and optimization.
of PCR conditions However, with the advances of next-generation sequencing technologies, the microsatellite discovery part of work has been minimized. Today, numerous microsatellites can be identified by data mining from genome sequences generated by next-generation sequencing.

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 and aquaculture research, including studies of genome mapping, parentage, kinships and genetic structure of stocks. 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, Moritz and Moore, 1995; Rico, Rico and Hewitt, 1996; Leclerc, Wirth and Bernatchez, 2000; Cairney, Taggart and Hoyheim, 2000). As a result, microsatellites can also be used for comparative genome analysis. If microsatellites can be tagged to gene sequences, their potential for use in comparative mapping is greatly enhanced (Kucuktas et al., 2009; Ninwichian et al., 2012b).

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: (i) despite being very abundant, the development of hundreds of thousands or millions of microsatellite markers is practically almost impossible; (ii) automation has not been possible for microsatellite genotyping; multiplexing has been limited to about a dozen of loci, at the most; and (iii) 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 single nucleotide polymorphism, including its rapid identification and automation in genotyping, makes SNP the far more preferred marker system for genome studies.

3.10 SNP markers

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 identify them and to genotype them rapidly in large numbers of samples was not possible until several major technological advances in the late 1990s, and especially after the adoption of next-generation sequencing after 2006.

The differences between alleles of SNPs are so 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. SNP discovery depends on sequencing. SNPs are able to be identified with high confidence more recently because of much deeper sequenc-
ing coverage provided by the next-generation sequencing technologies in comparison with the Sanger sequencing. The large-scale sequence data enabled efficient and effective identification of SNPs from genomes of various organisms. Using Illumina-based ribonucleic acid sequencing technology, large numbers of SNP markers have been identified from catfish (Liu et al., 2011), yielding large sets of gene-associated SNPs within channel catfish, blue catfish, and between channel catfish and blue catfish. Moreover, whole genome resequencing of four major channel catfish aquaculture populations and one wild population was conducted to identify a total of 8.4 million putative SNPs (Sun et al., 2014). On average, there is one SNP for every 93 base pairs in the catfish genome (Liu et al., 2016). The abundant and high-quality SNPs were devoted to the construction of high-density SNP arrays (Liu et al., 2014), enabling large-scale genotyping of genetic markers for GWAS (Geng et al., 2015), high-density linkage mapping (Li et al., 2015), fine QTL mapping, haplotype analysis, and whole genome-based selection. Similarly, genome-wide SNPs have been identified from common carp (Xu et al., 2012), rainbow trout (Palti et al., 2015a), pearl oyster (Jones et al., 2013), and Atlantic salmon (Yáñez et al., 2016).

High throughput and efficient SNP genotyping depends on the development of SNP arrays. Two types of SNP array technologies dominate the SNP marker. One is the Illumina BeadArray technology and the other is the Affymetrix SNP technology.

**The Illumina BeadArray technology:** Illumina offers two viable options for aquaculture researchers interested in SNP genotyping – the GoldenGate assay and the iSelect HD Custom BeadChip. The GoldenGate assays rely on allele-specific primer extension for SNP calling. In the GoldenGate assay, DNA samples are first bound to paramagnetic particles. Three oligonucleotides are designed for each SNP locus: two allele-specific oligos and a locus-specific oligo that hybridizes several bases downstream from the SNP site and which contains a bead-specific address. Following hybridization between genomic DNA and assay oligonucleotides, the template-primer complex is extended with DNA polymerase. Only when extension happens the allele-specific primer is brought in close proximity with the locus-specific primer for ligation. The ligation joins the appropriate allele-specific product (genotype) with the locus-specific primer (address) to form a full-length product that serves as a template for PCR using Cy3- and Cy5-labelled allele-specific primers. The single-stranded, dye-labelled DNAs are hybridized to their complement bead type contained on a BeadChip through their locus-specific primer address, fluorescent signal captured and SNP called.

The Illumina iSelect BeadChip uses a related technique, single-base extension (SBE) for SNP calling. In this approach, a two-step allele detection strategy is employed. Amplified, fragmented genomic DNA is first hybridized to bead-bound 50-mer oligos, providing locus specificity. Then SBE is carried out, allowing for the incorporation of a fluorescently labelled deoxyxynucleotide for assay readout and SNP calling.

**The Affymetrix Axiom genotyping technology:** Although the Illumina BeadArray technology is very efficient, it has limitations with the densities and the number of SNPs that can be analysed with a reasonable cost. This limitation is overcome by Affymetrix Axiom technology that allows a huge number of SNPs to be analysed, up to millions of SNPs.

A number of SNP arrays have been developed for aquaculture species (Table 1). Because of the high densities used, most of these were constructed using the Affymetrix technology (Table 1). For instance, a catfish 250K SNP array using Affymetrix Axiom genotyping technology has
recently been developed (Liu, et al., 2014), which has been used for genome-wide association studies, high-density linkage mapping, fine-scale QTL mapping, haplotype analysis, and whole genome-based selection. More recently, a 690K SNP array is being constructed for catfish.

In addition to SNP arrays, various other methods are available for SNP genotyping. These include Sequenom’s MassARRAY, Beckman Coulter’s SNPstream platform, the Applied Biosystems OpenArray system and Fluidigm’s Dynamic Array (EP1/BioMark system), among many other methods (Liu et al., 2011).

### TABLE 1
Development of high-density SNP arrays in aquaculture species

<table>
<thead>
<tr>
<th>Species</th>
<th>SNP array technology</th>
<th>SNP array density</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atlantic salmon</td>
<td>Illumina iSelect technology</td>
<td>15K</td>
<td>Gidskehaug et al., 2011</td>
</tr>
<tr>
<td>Atlantic salmon</td>
<td>Affymetrix Axiom technology</td>
<td>286K</td>
<td>Houston et al., 2014</td>
</tr>
<tr>
<td>Catfish</td>
<td>Affymetrix Axiom technology</td>
<td>250K</td>
<td>Liu et al., 2014</td>
</tr>
<tr>
<td>Catfish</td>
<td>Affymetrix Axiom technology</td>
<td>690K</td>
<td>Authors’ unpublished data</td>
</tr>
<tr>
<td>Common carp</td>
<td>Affymetrix Axiom technology</td>
<td>250K</td>
<td>Xu et al., 2014a</td>
</tr>
<tr>
<td>Rainbow trout</td>
<td>Affymetrix Axiom technology</td>
<td>57K</td>
<td>Palti et al., 2015a</td>
</tr>
</tbody>
</table>

### 3.11 Restriction site-associated DNA sequencing (RAD-seq) markers

RAD sequencing, or simply RAD-seq, refers to a method called restriction site-associated DNA sequencing that can identify and score thousands of genetic markers randomly distributed across the target genome from a group of individuals using next-generation sequencing technology. RAD-seq works by first fragmenting the target genome using a restriction enzyme. After digestion, a series of molecular processing steps transform the DNA into a fragment library suitable for sequencing on a next-generation sequencing platform. Sequence data are then analysed to identify and score genetic variations in the samples or population of interest. The variations identified mostly are SNP markers in nature. RAD-seq simultaneously identify and genotype SNPs in the samples. RAD-seq is widely used for a variety of molecular genetic studies, including identification of genetic variants (SNPs), phylogenetic analysis, germplasm assessment, analysis of population structure, linkage and QTL mapping, and GWAS analysis.

RAD-seq has been broadly used in aquaculture species. With aquaculture species, RAD-seq has been used for mapping QTLs (Houston et al., 2012; Gagnaire et al., 2013; Liu et al., 2015; Palti et al., 2015b; for a recent review, see Yue et al., 2014). Palaiokostas et al. (2013) used RAD-seq to map the sex-linked markers in Atlantic halibut. RAD-seq markers were used for linkage mapping in several aquaculture species, including Atlantic salmon (Gonen et al., 2014), sea urchins (Zhou et al., 2015), pearl oysters (Li and He, 2014) and Japanese flounder (Shao et al., 2015), and also for conservation genetic studies (Ogden et al., 2013; Huete-Pérez and Quezada, 2013).

Given the power of RAD-seq markers, its future perspective is good. However, for many applications involving over 100 individuals, analysis of a common set of RAD-seq markers is required. In most practical settings, the common set of RAD-seq markers can be limited in numbers, often 5 000–15 000 SNPs. Therefore, for analysis requiring a higher level of marker densities, RAD-seq
will not be able to deliver the required marker numbers. In addition, the cost of next-generation sequencing can add up quickly when the number of samples are increased.

4. GENOME MAPPING TECHNOLOGIES

The genomes of aquaculture fish vary from several hundreds of millions of base pairs to several billion base pairs. It is very difficult to study such large genomes without first “breaking” them into smaller pieces, and then sorting out their relationships, which is the task of genome mapping. There are two distinctive types of mapping methods: genetic linkage mapping and physical mapping, which produce genetic linkage maps and physical maps, respectively. While both maps are a collection of genetic markers and gene loci, genetic maps’ distances are based on the genetic linkage information and recombination rate between markers, while physical maps use actual physical distances of DNA, usually measured in the number of base pairs.

4.1 Genetic linkage mapping of aquaculture genomes

Researchers begin a genetic map by collecting samples of blood or tissue from family members. DNA is then isolated from the samples and analysed for association of marker patterns. When markers are inherited mostly together, they are “linked” physically on the same chromosome. When they are always linked together, they are located at the same genetic locus. Their tightness of linkage depends on the distance between the markers. The farther away the two markers, the more likely a recombination can happen between them. Based on the frequency of their recombination, the genetic distances are assigned. The unit of genetic distance is defined as centiMorgan (cM). Each cM equals 1 percent of the recombinants.

The first steps of building a genetic map are the development of genetic markers and a mapping population. Since the closer the two markers are on the chromosome, the more likely they are to be passed on to the next generation together; therefore, the “co-segregation” patterns of all markers can be used to reconstruct their order. With this in mind, the genotypes of each genetic marker are recorded for both parents and in each individual in the progenies. The quality of the genetic maps is largely dependent upon these two factors: the number of genetic markers on the map and the size of the mapping population. The two factors are interlinked, as a larger mapping population could increase the “resolution” of the map and prevent the map being “saturated” with stacked markers.

In genetic mapping, any sequence feature that can be faithfully distinguished from the two parents can be used as a genetic marker. Genes, in this regard, are represented by “traits” that can be faithfully distinguished between two parents. Their linkage with other genetic markers are calculated the same way as if they are common markers, and the actual gene loci are then bracketed in a region between the two nearest neighbouring markers. The entire process is then repeated by looking at more markers.

All markers that are more or less co-segregated belong to the same linkage group (LG). The number of linkage groups is equal to the number of chromosomes. Thus, a genetic linkage map is composed of polymorphic markers that are assigned to LGs, and their distances within the LG are defined by the recombination fraction among the markers.
A number of population types are suitable for genetic linkage mapping. Most often, second hybrid generation (F2) or higher generation intercrosses are good because in F2 populations, markers, as well as the traits, are segregating. In a similar fashion, backcross progenies can be used for genetic linkage mapping as well. For loci that are heterozygous in the parents, they would be segregating in first filial generation (F1) already, and therefore F1 populations have also been used for genetic linkage mapping.

All types of polymorphic markers can be used for genetic linkage mapping. However, in order to make a linkage map that has a high density of markers, thousands or tens of thousands of markers are needed. In this regard, microsatellites and SNP markers are most appropriate because they are the most abundant types of markers in the genomes. Genetic linkage maps have been constructed in many aquaculture species, and some examples are shown in Table 2, although the marker density and resolution differ greatly. Once again, the marker density provides the level of genome coverage, while the resolution is defined by the number of samples used in the genetic analysis.

<table>
<thead>
<tr>
<th>Species</th>
<th>Number and type of markers</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asian seabass</td>
<td>790 microsatellites and single nucleotide polymorphisms (SNPs)</td>
<td>Wang et al., 2011</td>
</tr>
<tr>
<td>Atlantic salmon</td>
<td>5,650 SNPs</td>
<td>Lien et al., 2011</td>
</tr>
<tr>
<td>Brown trout</td>
<td>288 microsatellites and 13 allozymes</td>
<td>Gharbi et al., 2006</td>
</tr>
<tr>
<td>Catfish</td>
<td>54,342 SNPs</td>
<td>Li et al., 2015</td>
</tr>
<tr>
<td>Common carp</td>
<td>732 microsatellites</td>
<td>Zhang et al., 2013</td>
</tr>
<tr>
<td>Eastern oyster</td>
<td>282 amplified fragment length polymorphisms</td>
<td>Yu and Ximing, 2003</td>
</tr>
<tr>
<td>European seabass</td>
<td>Microsatellites</td>
<td>Chistiakov et al., 2005</td>
</tr>
<tr>
<td>Grass carp</td>
<td>279 microsatellites and SNPs</td>
<td>Xia et al., 2010a</td>
</tr>
<tr>
<td>Japanese flounder</td>
<td>1,375 microsatellites</td>
<td>Castaño-Sánchez et al., 2010</td>
</tr>
<tr>
<td>Pacific oyster</td>
<td>1,166 SNPs and microsatellites</td>
<td>Hedgecock et al., 2015</td>
</tr>
<tr>
<td>Rainbow trout</td>
<td>2,226 microsatellites and SNPs</td>
<td>Guyomard et al., 2012</td>
</tr>
<tr>
<td>Scallop</td>
<td>169 microsatellites</td>
<td>Li et al., 2012b</td>
</tr>
<tr>
<td>Sea bream</td>
<td>321 microsatellites, expressed sequence tags and SNP markers</td>
<td>Tsigenopoulos et al., 2014</td>
</tr>
<tr>
<td>Shrimp</td>
<td>3,999 SNPs</td>
<td>Baranski et al., 2014</td>
</tr>
<tr>
<td>Tilapia</td>
<td>525 microsatellites</td>
<td>Lee et al., 2005</td>
</tr>
<tr>
<td>Yellowtail</td>
<td>217 microsatellites</td>
<td>Ohara et al., 2005</td>
</tr>
</tbody>
</table>

4.2 Physical mapping of aquaculture genomes

Since actual base-pair distances are generally hard or impossible to directly measure, physical maps are actually constructed by first shattering the genome into smaller pieces. By characterizing each single piece and assembling them back together, the overlapping path or “tiling path” of these small fragments would allow researchers to infer physical distances between genomic features. The fragmentation of the genome can be achieved by restriction enzyme cutting or by physically shattering the genome by processes like sonication. Once cut, the DNA fragments are
separated by electrophoresis. The resulting pattern of DNA migration (i.e. its genetic fingerprint) is used to identify what stretch of DNA is in the clone. By analysing the fingerprints, contigs are assembled by automated (FingerPrintedContigs), or by manual means (Pathfinders) into overlapping DNA stretches.

Physical mapping starts with the cloning of large genomic DNA segments into cloning vectors such as bacterial artificial chromosome (BAC) vectors. The genomic segments are approximately 150–200 kb long. Thus, a genome of one billion base pairs would require 5 000–7 000 BAC clones. It takes an average of 6–10X genome coverage to make sure that the whole genome is covered at least once everywhere. These segments are generated by random partial restriction digestions, and therefore analysis of multiple clones (6–10 times) from the same genomic location would allow the determination of their positional relationship. This is usually done by restriction fingerprinting.

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). Such an order is established by placing overlapping restriction patterns one to another to extend to the whole genome. In practical situations, however, contigs would break upon the gaps that are present in the BAC library.

Fingerprinting of BAC clones is conducted by fluorescence-based fingerprinting methods. In general, methods that produce a greater number of bands can detect overlaps more efficiently than methods that produce fewer bands. Labelling the fragments with different colours increases the information content. The procedure of Ding et al. (2001) gave the largest number of informative fragments followed by the SnaPshot labelling method (Luo et al., 2003).

After fingerprinting, 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, FingerPrintedContigs (FPC) (Soderlund et al., 2000). 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 of clones into the same contigs.

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 of different restriction enzymes.

Physical maps were constructed for a limited number of aquaculture fish and shellfish species (Table 3). Recent advances in the generation of long sequencing reads may actually reduce the demands of physical mapping. However, physical maps can provide an independent validation for the whole genome reference sequences.
TABLE 3.
Examples of physical maps constructed from aquaculture species

<table>
<thead>
<tr>
<th>Species with physical maps</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atlantic salmon</td>
<td>Ng et al., 2005</td>
</tr>
<tr>
<td>Tilapia</td>
<td>Katagiri et al., 2005</td>
</tr>
<tr>
<td>Channel catfish</td>
<td>Xu et al., 2007</td>
</tr>
<tr>
<td>Rainbow trout</td>
<td>Palti et al., 2009</td>
</tr>
<tr>
<td>Common carp</td>
<td>Xu et al., 2011</td>
</tr>
<tr>
<td>Asian seabass</td>
<td>Xia et al., 2010b</td>
</tr>
<tr>
<td>Scallop</td>
<td>Zhang et al., 2011a</td>
</tr>
</tbody>
</table>

4.3 Radiation hybrid mapping

Irradiation and fusion gene transfer technology has been used for a long time. In the early 1990s, Cox et al. 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). They were able to establish a linear order of DNA markers by observing their co-retention in hybrid cell lines. Radiation hybrid mapping strategies are based on the concept that markers that are close together on chromosomes will frequently be co-retained in the same hybrids – the probability that irradiation will induce a chromosome break between two markers increases as the physical distance between the two markers increases. To provide adequate statistical support for mapping marker retention frequencies, the percentage of times a marker is scored positive in a radiation hybrid (RH) panel is critical. RH mapping is calculated based on the co-retention of markers in fragments across the hybrid panel cell lines. The estimated frequency of breakage between two markers is \( \theta \), which ranges from 0 to 1, and is analogous to recombination frequencies \( r \) used in genetic mapping. A \( \theta \) value of 0 means two markers are always co-retained; a value of 1 means they are co-retained 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 the use of centiMorgans in linkage mapping. Hence, the observation of chromosome breaks between two markers in RH mapping is analogous to observing recombination between two markers in genetic mapping. To simplify this procedure, Walter et al. (1994) reported the development of whole genome radiation hybrid panels. The 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.

Very different from linkage mapping where polymorphic markers must be used, RH mapping does not require polymorphic markers, but just known DNA sequences of the species for PCR amplification. It actually prefers the use of gene markers because gene sequences are more likely to be conserved and more likely to be unique in the genome. Because RH mapping defines linkage as co-retention of markers in the same cell after irradiation and fusion, the use of gene markers reduces the complexities in PCR amplification.

Radiation hybrid mapping has been used for various mammalian species, but less so for aquaculture species. With aquatic and aquaculture species, it has been only used for just a few species, such as zebrafish, European seabass and gilthead seabream (Dahm et al., 2006; Senger et al., 2006; Sarropoulou et al., 2007; Guyon et al., 2010).
4.4 Optical mapping

Optical mapping is a physical mapping method for constructing high-resolution restriction maps of a whole genome from single, fluorescently stained molecules of DNA (Schwartz et al., 1993). In principle, a single DNA molecule can be digested by restriction endonuclease, stained with fluorescence dye, and then the fluorescence is captured optically. The physical lengths of the restriction fragments are recorded. Such ordered, whole genome coverage of restriction fingerprints is then referred to as optical maps. The advantage of this mapping technology is the total void of cloning or gel electrophoresis.

Although the principles of optical mapping have been established for more than two decades, it has not been widely used until recently. This is largely due to recent technological advances in nanotechnology and the ability to optically capture the fluorescence from a single molecule of DNA. Optical mapping is now mostly used to validate the whole genome reference sequence assembly (Dong et al., 2013; Reslewic et al., 2005; Zhou et al., 2007). Optical mapping has not been used in aquaculture species yet, according to the author’s knowledge.

4.5 Integration of physical and linkage maps

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.

One way to integrate physical maps with genetic linkage maps is to genetically map polymorphic markers anchored to known physical map contigs. Such markers typically can be developed from BAC end sequencing. This approach was used for map integration of a number of aquaculture species such as catfish and carp (Li et al., 2015; Zhao et al., 2013).

5. GENOME SEQUENCING TECHNOLOGIES

5.1 First-generation DNA sequencers

The double helix DNA structure was revealed in 1953. It took the next 24 years, until 1977, for the world to develop technology to sequence DNA. Although the Maxam-Gilbert chemical method was invented at the same time and the group shared the Nobel Prize with Frederick Sanger’s group, the method became rapidly obsolete because of its use of toxic chemicals in the sequencing reactions. With the development of Sanger’s “chain termination”, or the dideoxy sequencing technique (Sanger, Nicklen and Coulson, 1977), it marks the theoretical maturation of sequencing technologies. The chain-termination technique makes use of chemical analogues of the deoxyribonucleotides, the dideoxyribonucleotides that will terminate the synthesizing DNA chain upon their addition. Sanger sequencing became the most common technology used to sequence DNA for many years, until early 2000. Automated sequencers, most typical of the ABI sequencers based on Sanger sequencing, were developed in the early 1980s and they became the first generation of DNA sequencers.
The first-generation DNA sequencers, such as ABI 3700 or 3730, can sequence 96 samples per run with a read length of 500–800 base pairs per sample. Thus, the output per run was approximately 60,000 base pairs. The cost was very high in today’s standards.

5.2 Second-generation (the next generation) sequencers

The pressure to produce more, faster and cheaper DNA sequences that provide basic information pivotal for scientific research and personalized medicine has been the force for the development of DNA sequencing technologies. Starting with the capability to sequence the first base of DNA (Maxam and Gilbert, 1977; Sanger, Nicklen and Coulson, 1977), DNA sequencing technologies have had revolutionary advances reflected in their progress, from manual sequencing to automation, from sequencing a single template to mass parallel sequencing of billions of reads, and from costing about US$15 per base to less than just a few pennies per million base pairs. A recent review by Heather and Chain (2016) provides a good coverage of the history of DNA sequencing technologies.

The so-called next-generation sequencing started with the launch of the Roche 454 Genome Sequencer FLX System in 2005, followed by the Solexa (now Illumina) sequencing platform, commercialized in 2006. The GS FLX System based on sequencing-by-synthesis (pyrosequencing) technology was developed by 454 Life Sciences as the first next-generation sequencing platform available on the market (Margulies et al., 2005).

Although several aquaculture genomes were initially sequenced using the 454 technology, such as Atlantic salmon, Atlantic cod, rainbow trout, crucian carp, scallops and catfish (Vera et al., 2008; Salem et al., 2010; Hou et al., 2011; Star et al., 2011; Liao et al., 2013), some of the fundamental issues associated with the pyrosequencing technology made the 454 system more vulnerable to sequencing errors. In addition, its high but much lower output compared with several other platforms made it obsolete rapidly.

5.3 Illumina sequencers

The Solexa sequencing platform was commercialized in 2006. The principle is on the basis of sequencing-by-synthesis chemistry. Its efficiency and capabilities are powered by massive parallel sequencing of hundreds of millions of templates simultaneously.

In 2008, Illumina introduced an upgrade, the Genome Analyzer II. It offered a powerful combination of the cBot and Paired-End module. cBot is a revolutionary automated system that creates clonal clusters from single molecule DNA templates, preparing them for sequencing by synthesis on the Genome Analyzer. For Genome Analyzer II, the run time was highly decreased and the output per paired-end run can reach 45–50 Gb. Compared with Sanger sequencing, the Illumina system is able to produce more data at a reduced time and cost. In 2010, Illumina launched its HiSeq series of sequencers, whose capacity was enhanced to over 400 million reads per lane, or over 3 billion reads per run. Currently, Illumina carries a series of platforms of MiniSeq system (maximum output 7.5 Gb), MiSeq (maximum output 15 Gb), NextSeq series (maximum output 120 Gb), HiSeq series (maximum output 1,500 Gb), and HiSeq X series (maximum output 1,800 Gb). Clearly, the drastic increase of output allowed almost a proportional drop in costs.

In addition to the Illumina system, several other systems offer similar output and sequencing quality. These include the Applied Biosystems SOLiD System, which is based on a sequencing-by-li-
gation technology. The advantage of this platform is its very high sequencing accuracy. However, it appears that Illumina technology is the dominant technology in today’s marketplace among all second-generation DNA sequencers.

5.4 Third-generation DNA sequencers

There is no clear definition and clear borderline between the second- and the third-generation sequencers. However, it is generally accepted that the third-generation sequencing technologies are marked by single molecule sequencing (SMS) and real-time sequencing (Heather and Chain, 2016).

The first SMS technology was developed in the early 2000s (Braslavsky et al., 2003; Harris et al., 2008). Helicos was the first company that worked on third-generation sequencing, but it closed down in 2012. Currently, the most widely used third-generation technology is the single molecule real-time platform from Pacific Biosciences (van Dijk et al., 2014). With the new P6-C4 chemistry, the PacBio sequencers can generate 50 000–100 000 reads per flow cell. This sequencing platform provides long reads of up to 40 kb, with an average of 10–15 kb. Therefore, each flow cell can generate 500 million to 1 billion bases. The long reads provide tremendous advantage for de novo genome sequencing.

One major disadvantage of the PacBio sequencing is its high error rate of almost 10 percent. However, the errors occur randomly along the bases of DNA. High genome coverage would allow generation of consensus sequences. For instance, if the sequences have a 10X coverage, the error to occur at any given base becomes $1 \times 10^{-10}$. Recently, PacBio has produced base correction software, allowing generation of error-corrected consensus. In some cases, Illumina sequences can be used to correct the sequence errors in PacBio sequences.

In addition to the PacBio sequencing, there are several other third-generation sequencing platforms. Of these, the most particularly promising is the Nanopore DNA sequencing. However, at this time, these platforms are less frequently used in the marketplace.

5.5 Application of next-generation sequencing

Next-generation sequencing technologies are applied in a variety of areas. Some areas of applications include the following: (i) de novo genome sequencing, whole-genome resequencing; (ii) marker development for the identification of microsatellites or SNP markers; (iii) transcriptome sequencing for the analysis of genome level expression profiling and identification of differentially expressed genes or co-induced genes; (iv) large-scale analysis of epigenetic regulation, such as DNA methylation, by deep sequencing of bisulfite-treated DNA; and (v) genome-wide mapping of DNA-protein interactions by deep sequencing of DNA fragments pulled down by chromatin immunoprecipitation.

5.6 Genome sequencing in aquaculture species

With the advances of the sequencing technologies, rapid progress has been made and is being made with whole genome sequencing with aquaculture species. Genome sequencing projects have been initiated with at least 30 fish and shellfish species. These projects included the sequencing of the genomes for several species of carps such as the common carp (Xu et al., 2014b), grass carp (Wang et al., 2015) and rohu carp (personal communication, unpublished),
and the sequencing of the tilapia genome, the catfish genome, the seabass genome, shark genome, cavefish genome, platyfish genome, sole genome, yellow croaker genome, rainbow trout genome, and the Atlantic salmon genome, among many others (Table 4). Quite a few shellfish genomes have also been sequenced. These include the Pacific oyster (Zhang et al., 2012), eastern oyster, abalone, shrimp and scallops.

Most of the aquaculture species genomes have been sequenced using the Illumina technology, and in some cases supplemented with third-generation sequencing technologies such as PacBio sequencing. The assembly qualities vary. Four parameters are often used to assess the quality of the genome assemblies: the completeness, the accuracy, the contiguity and the connectivity (Table 4).

<table>
<thead>
<tr>
<th>Species</th>
<th>Contig N50 (Kb)</th>
<th>Scaffold N50 (Mb)</th>
<th>% on chromosome</th>
<th>Sequencing platform</th>
<th>Total size (Mb)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stickleback</td>
<td>83.2</td>
<td>10.8</td>
<td>86.9</td>
<td>9.0X Sanger</td>
<td>463</td>
<td>Jones et al., 2012</td>
</tr>
<tr>
<td>Catfish</td>
<td>77.2</td>
<td>7.73</td>
<td>97.2</td>
<td>Illumina, PacBio</td>
<td>783</td>
<td>Liu et al., 2016</td>
</tr>
<tr>
<td>Grass carp</td>
<td>40.8</td>
<td>6.46</td>
<td>64.0</td>
<td>132X Illumina</td>
<td>900.5</td>
<td>Wang et al., 2015</td>
</tr>
<tr>
<td>Seabass</td>
<td>53.2</td>
<td>5.09</td>
<td>86.0</td>
<td>30X</td>
<td>675.4</td>
<td>Tine et al., 2014</td>
</tr>
<tr>
<td>Shark</td>
<td>46.6</td>
<td>4.5</td>
<td>-</td>
<td>454, Sanger</td>
<td>937</td>
<td>Venkatesh et al., 2014</td>
</tr>
<tr>
<td>Turbot</td>
<td>31.2</td>
<td>4.3</td>
<td>-</td>
<td>Illumina</td>
<td>544</td>
<td>Figueras et al., 2016</td>
</tr>
<tr>
<td>Tilapia</td>
<td>29.3</td>
<td>2.80</td>
<td>70.9</td>
<td>269X Illumina</td>
<td>1 010</td>
<td>Brawand et al., 2014</td>
</tr>
<tr>
<td>Cavefish</td>
<td>14.7</td>
<td>1.78</td>
<td>-</td>
<td>94X Illumina</td>
<td>964</td>
<td>McGaugh et al., 2014</td>
</tr>
<tr>
<td>Zebrafish</td>
<td>25.0</td>
<td>1.55</td>
<td>96.5</td>
<td>7.5X Sanger, Illumina</td>
<td>1 410</td>
<td>Howe et al., 2013</td>
</tr>
<tr>
<td>Medaka</td>
<td>9.8</td>
<td>1.41</td>
<td>89.7</td>
<td>10.6X Sanger</td>
<td>700.4</td>
<td>Kasahara et al., 2007</td>
</tr>
<tr>
<td>Platyfish</td>
<td>22.0</td>
<td>1.1</td>
<td>90.2</td>
<td>454, Illumina</td>
<td>669</td>
<td>Schartl et al., 2013</td>
</tr>
<tr>
<td>Common carp</td>
<td>68.4</td>
<td>1.0</td>
<td>51.8</td>
<td>454, Illumina, SOLiD</td>
<td>1 690</td>
<td>Xu et al., 2014b</td>
</tr>
<tr>
<td>Tetraodon</td>
<td>16.0</td>
<td>0.98</td>
<td>64.6</td>
<td>8.3X Sanger</td>
<td>342.4</td>
<td>Jaillon et al., 2004</td>
</tr>
<tr>
<td>Coelacanth</td>
<td>12.7</td>
<td>0.92</td>
<td>-</td>
<td>Illumina</td>
<td>2 860</td>
<td>Amemiya et al., 2013</td>
</tr>
<tr>
<td>Sole</td>
<td>26.5</td>
<td>0.87</td>
<td>93.3</td>
<td>Illumina</td>
<td>477</td>
<td>Chen et al., 2014</td>
</tr>
<tr>
<td>Cod</td>
<td>7.1</td>
<td>0.69</td>
<td>44.1</td>
<td>454</td>
<td>753</td>
<td>Star et al., 2011</td>
</tr>
<tr>
<td>Yellow croacker</td>
<td>25.7</td>
<td>0.50</td>
<td>-</td>
<td>76X Illumina</td>
<td>644</td>
<td>Wu et al., 2014</td>
</tr>
<tr>
<td>Rainbow trout</td>
<td>7.7</td>
<td>0.38</td>
<td>54.0</td>
<td>70X Illumina</td>
<td>1 900</td>
<td>Berthelot et al., 2014</td>
</tr>
<tr>
<td>Lamprey</td>
<td>-</td>
<td>0.17</td>
<td>-</td>
<td>Sanger</td>
<td>816</td>
<td>Smith et al., 2013</td>
</tr>
<tr>
<td>Fugu</td>
<td>16.5</td>
<td>0.05-0.1</td>
<td>-</td>
<td>5.6X Sanger</td>
<td>332.5</td>
<td>Aparicio et al., 2002</td>
</tr>
</tbody>
</table>
In addition, the proportion of the genome anchored to the chromosomes is also another important parameter. Completeness refers to the percentage of the genome actually sequenced; the accuracy refers to the correctness of the assembly. Because of the large genome sizes and the complexities of the genome sequences, errors may exist in some genome assemblies. Some of the genome assemblies are of poor quality, including many mistakes, and a good example is the Pacific oyster's genome assembly, which was reported to include a high percentage of mistakes with its scaffolds (Hedgecock et al., 2015). The contiguity refers to the median length of the contiguous sequences as expressed as N50 length of the contigs. Connectivity refers to the distribution of scaffold sizes as expressed by N50 scaffold size (Table 4).

5.7 Genome annotation of aquaculture species

Annotation is the process by which pertinent information about raw DNA sequences is added to the genome databases. This involves describing different regions of the code and identifying which regions can be called genes. Part of genome annotation deals with the discovery and identification of protein coding genes in the genome. Strictly speaking, genome annotation covers the interpretation of the genome sequence, including but not limited to, protein encoding genes, gene structure and organization such as exons, introns, promoter sequences, regulatory elements such as enhancers and silencers, polyadenylation signals, transcriptional factor binding sites, epigenetic information, and polymorphism information such as the positions of SNPs. However, for aquaculture species, identification of genes, especially protein coding genes, is the first step for genome annotation.

A genome can be annotated by computer software predictions or by evidence-based approaches. Many software packages are available, but two of them are the most popular. The first is called Fgenesh, and the other is called Augustus. These software packages predict genes based on sequence similarities and the presence of open reading frames for proteins. Evidence-based annotation is based on the identification of transcripts detected from the cells of the organism. In the past, expressed sequence tag (EST) analysis was the major source of the transcript evidence. Recently, the application of ribonucleic acid sequencing (RNA-seq) has allowed rapid progress for transcriptome analysis, as detailed below.

6. TRANSCRIPTOME ANALYSIS OF AQUACULTURE SPECIES

Transcriptome refers to the complete composition of RNAs of an organism. Years ago, EST analysis was the major approach for transcriptome sequencing. Recently, RNA-seq using next-generation sequencing has allowed the most rapid progress.

6.1 Development of EST resources of aquaculture species

Expressed sequence tags are single-pass sequences of random complementary DNA (cDNA) clones from cDNA libraries. They are traditionally generated using Sanger sequencing and therefore the resultant sequences are approximately 500 to 800 base pairs in length. Several years ago, because sequencing was relatively cheap, large numbers of ESTs can now be generated at a reasonably low cost from either the 5’ or 3’ end of a cDNA clone to get an insight into transcrip-
tionally active regions. ESTs were used as a primary resource for human gene discovery (Adams et al., 1991). Thereafter, there has been an exponential growth in the generation and accumulation of EST data in public databases for various organisms, with approximately 74 million ESTs now available in these databases (http://www.ncbi.nlm.nih.gov/dbEST; March 2016 release number 130101, all species).

EST analysis is an effective genomic approach for rapid identification of expressed genes, and has been widely used in genome-wide gene expression studies in various tissues, developmental stages and under different environmental conditions (Franco et al., 1995). In addition, the availability of cDNA sequences has accelerated further molecular characterization of genes of interest and provided sequence information for microarray construction and genome annotation (Rise et al., 2004).

Gene expression analysis plays an important role in identifying differentially expressed genes under different environmental conditions and gene expression regulation, shedding light on gene functions. EST analysis has been demonstrated effective for detection of differential expression and regulation of certain genes. Without normalization or subtraction in library construction, the number of the sequenced ESTs for a given gene reflected the abundance of the gene expression at the corresponding scenario (e.g. environmental conditions, developmental stages, treatments).

Direct EST sequencing is inefficient in discovery of rarely expressed genes. To solve this problem, the method to construct normalized cDNA libraries was developed (Bonaldo, Lennon and Soares, 1996; Soares et al., 1994). The basic principle is using hybridization to reduce redundant genes and increase the representation of rarely expressed genes. However, with the adoption of RNA-seq technology, as described below, it is unlikely that additional EST analysis will be conducted because of the low efficiency as compared with RNA-seq.

Initial annotation of ESTs can be conducted by simple sequence similarity comparisons. Further annotation analysis can be carried out after obtaining the consensus sequences (putative unigenes), such as determination of gene identity based on homology search, open reading frame identification, gene ontology annotation and gene-enrichment analysis (e.g. Nakaya et al., 2007).

In order to assign gene identity to contigs and singletons, homology search is widely used. Such an approach is especially helpful for newly studied species. BLAST is the most widely used program to obtain high throughput EST analysis and annotation results. The BLAST package provides different flavours of algorithms for sequence similarity searching. BLASTX is used to search against protein database by translated consensus EST sequences, while BLASTN is used to search against nucleotide sequence databases. The National Center for Biotechnology Information (NCBI), ENSEMBL and Swiss-Prot are three important databases for BLAST search. For instance, the Swiss-Prot database has a fully manually curated and annotated unigene database, Uniprot, which can be used for identifying putative function for unigene by BALSTX. The NCBI provides a dbEST database that can be used to search novel transcripts by BLASTN. The dbEST is a main EST resource database, including ESTs for over 200 aquaculture species. The ENSEMBL database can provide chromosome location information of genes, which is a useful tool for comparative genome analysis. However, the BLAST sequence similarity comparison provides only sequence
homology information, and one cannot purely rely on BLAST for gene identification. Detailed phylogenetic analysis and/or orthology analysis is needed to determine the identities of genes.

For a greater level of annotation, the open reading frame (ORF) is identified to determine the full or portion of coding region in the unigene. The unigene with a full ORF usually represents a full length cDNA. There are some useful tools for ORF detection. For example, the ESTScan (Iseli, Jongeneel and Bucher, 1999) can extract coding regions from low-quality ESTs and correct frame shift errors. OrfPredictor (Min et al., 2005) is another program for identification of protein-coding sequences from ESTs through predicting the most probable coding regions from all six translation frames.

Gene ontology (GO) annotation can provide descriptions of gene products behaving in a cellular context. Gene functions are placed into three categories: biological processes, cellular components and molecular functions. Consensus sequences can be linked to GO terms and assigned a possible function by Blast2GO (Conesa et al., 2005).

Gene ontology enrichment analysis is to cluster most relevant GO terms associated with certain biological pathways. The GOEAST (Zheng and Wang, 2008), Ontologizer (Bauer et al., 2008), GeneTrail (Backes et al., 2007), and DAVID functional annotation tool (Huang, Sherman and Lempicki, 2009) are useful tools for these analyses.

EST analysis is an efficient approach for gene discovery and gene identification. For instance, between 2001 and 2007, catfish ESTs increased from 10 000 to 44 000 and the putative genes number increased from 5 905 to 25 000 (Li et al., 2007). In the Pacific oyster (Crassostrea gigas), 40 845 high-quality ESTs represented 29 745 unique transcribed sequences (Fleury et al., 2009); in gilthead sea bream (Sparus auratus), 30 000 ESTs represented 18 196 putative unigenes (Louro et al., 2010). Currently, there are over 180 aquaculture species having more than 100 ESTs in dbEST, with approximately a dozen species having more than 10 000 ESTs (Table 5).

EST analysis can provide comparisons of gene expression profiling in different tissues and conditions. For instance, in a recent study with rainbow trout (Oncorhynchus mykiss), Kondo et al. (2011) sequenced over 30 000 ESTs from rainbow trout adipose tissue. These ESTs were used to search adipokine-related genes. The result showed that none of them encoded adipokine and PPARγ gene, which play important roles in mammalian adipocytes. Further qRT-PCR results confirmed EST analysis results, that is, rainbow trout adiponectin transcripts were weakly detected in adipose tissue but strongly detected in muscle, suggesting the difference of energy metabolism between fish and mammals (Kondo et al., 2011). Chini et al. (2008) constructed normalized cDNA libraries from liver, ovary and testis in bluefin tuna (Thunnus thynnus), identifying several sequences with known functions in other organisms, but not previously described in this species. Also, sequences were described being expressed in one, two or more tissue libraries. Similarly, Zou et al. (2011) constructed normalized cDNA libraries from testis, ovary and mixed organs of mud crab (Scylla paramamosain). Through EST analysis, sex-specific transcripts were identified.
### TABLE 5:
Some examples of aquatic species with major expressed sequence tag (EST) resources (>10,000) using zebrafish (1,488,275) as a reference

<table>
<thead>
<tr>
<th>Species</th>
<th>Number of ESTs</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Danio rerio</em> (zebrafish)</td>
<td>1,488,275</td>
</tr>
<tr>
<td><em>Ciona intestinalis</em></td>
<td>1,205,674</td>
</tr>
<tr>
<td><em>Xenopus laevis</em> (African clawed frog)</td>
<td>677,911</td>
</tr>
<tr>
<td><em>Oryzias latipes</em> (Japanese medaka)</td>
<td>666,891</td>
</tr>
<tr>
<td><em>Salmo salar</em> (Atlantic salmon)</td>
<td>498,245</td>
</tr>
<tr>
<td><em>Ictalurus punctatus</em> (channel catfish)</td>
<td>354,516</td>
</tr>
<tr>
<td><em>Oncorhynchus mykiss</em> (rainbow trout)</td>
<td>287,564</td>
</tr>
<tr>
<td><em>Crassostrea gigas</em> (Pacific oyster)</td>
<td>206,388</td>
</tr>
<tr>
<td><em>Litopenaeus vannamei</em> (white shrimp)</td>
<td>161,248</td>
</tr>
<tr>
<td><em>Ictalurus furcatus</em></td>
<td>139,475</td>
</tr>
<tr>
<td><em>Oreochromis niloticus</em> (Nile tilapia)</td>
<td>120,991</td>
</tr>
<tr>
<td><em>Petromyzon marinus</em> (sea lamprey)</td>
<td>120,731</td>
</tr>
<tr>
<td><em>Sparus aurata</em></td>
<td>79,216</td>
</tr>
</tbody>
</table>

EST resources provide sequence information for microarray development. For instance, in a recent study, Booman et al. (2011) developed a large-scale oligonucleotide microarray platform containing 20,000 features (20K), which was used to study immune response of the Atlantic cod spleen with stimulation of formalin-killed atypical *Aeromonas salmonicida* (Booman et al., 2011). Similarly, oligo microarray for gilthead sea bream (*Sparus aurata*) was developed based on ESTs, and the microarray was used to identify 1,050 differentially expressed genes between two developmental stages (Ferraresso et al., 2008).

Although EST analysis has been important for transcriptome characterization, it is now becoming expensive, relative to several of the recently developed approaches, as detailed below. However, EST resources still have a great value to serve as reference for RNA-seq analysis. It has been found that ESTs are useful for high-quality reference-guided assembly of next-generation sequencer-generated short reads (Liu et al., 2011).

### 6.2 RNA-seq technologies

RNA-seq is a technology to sequence transcriptomes using next-generation sequencing technologies. It has been widely used for analysis of gene expression profiling and identification of differentially expressed genes. RNA-seq can be done with a number of sequencing platforms, including the Illumina sequencing platform, ABI Solid Sequencing, and less efficiently, the Life Science’s 454 sequencing. Of these, the Illumina HiSeq sequencers are the most popular because of their very high throughput and accuracy of sequencing reads.

RNA-seq analysis starts with RNA samples. Before RNA-seq, considerations need to be made to allow statistical analysis of the results with proper biological and technical replica. For instance, a control sample can be compared with the infected samples at various times after infection. In order to allow statistical analysis of the results, a minimal of three biological replications is required for each condition (treatment).
The biological issues of RNA-seq are actually extremely simple. The researchers are interested in which genes are expressed, how much they are expressed, and how the samples in different conditions compare. Before answering these questions, the immediate task is to assemble the short reads into reference transcriptome. In general, one of the two types of assembly methods can be used for the assembly of RNA-seq sequences, depending on the existing genome resources. If a reference genome sequence is available, reference-guided assembly methods can be used. In contrast, de novo RNA-seq assembly methods must be used in the absence of a reference genome sequence.

RNA-seq has been extensively used for the identification of gene-associated markers. In catfish, hundreds of thousands of gene-associated SNPs have been identified by deep sequencing of RNA from many individuals of both channel catfish and blue catfish, which will be used in the development of high-density catfish SNP chips for genome-wide association studies (Liu et al., 2011). In the study to understand the adaptive divergence between dwarf and normal lake whitefish species, 454 sequencing was used with the aim to generate a set of SNP markers; 89 SNPs showed pronounced allele frequency differences between sympatric normal and dwarf whitefish (Renaut, Nolte and Bernatchez, 2010).

RNA-seq data sets are being deposited to the NCBI’s Sequence Read Archive (SRA) at www.ncbi.nlm.nih.gov/sra. The RNA-seq data sets can be searched, retrieved and downloaded from the SRA database. For instance, searching of SRA using “catfish” as the keyword results in 109 records, as of March 2016. Researchers can download any of the records for additional analysis. Such data sets are most often used for meta-analysis.

### 6.3 Analysis of differentially expressed genes using RNA-seq

RNA-seq has many applications. However, the two popular analyses using RNA-seq are the identification of differentially expressed genes during development, under a specific physiological condition, or after certain treatments. The basic principle for the identification of differentially expressed genes is to calculate the ratio of counts of short reads aligned to each gene before and after treatment. For instance, if 50 RNA-seq reads align to gene X after infection while only 5 RNA-seq reads align to the same gene before infection, then the infection induced the expression of this gene is tenfold. This is to demonstrate the concept. In practice, however, detailed normalization and statistical analysis must be conducted.

With RNA-seq, the higher the expression level a gene has, the more reads are sequenced from this gene. However, the read count is correlated with the size of the gene exons. Therefore, the sizes of the gene must be normalized to compare their expression. Reads per kilobase per million (RPKM) has been widely used as a reasonable normalizer. RPKM is the number of reads per kb long of a transcript per 1 000 000 reads of RNA-seq sequences. 

\[
\text{RPKM} = 10^9 \times \frac{C}{NL}
\]

where C is the total number of reads mapped onto the gene; N is the total number of mapped reads; and L is the sum of the genes in base pairs. For instance, the RPKM of a 2 kb transcript with 3 000 alignments in a sample of 10 million of mapped reads is calculated as:

\[
\text{RPKM} = 10^9 \times 3 \, 000 \times 10 \, 000 \, 000 = 150
\]

Once all the read counts are normalized into RPKM, expression levels among genes and before and after treatment can be calculated.
6.4 Analysis of co-expressed genes using RNA-seq data sets

In addition to the identification of differentially expressed genes, RNA-seq can be used to determine correlated or coordinated expression. Such analysis can be quite insightful. For instance, a single treatment such as high temperature treatment may induce a common set of genes, and these genes may all have similar functions in dealing with the heat shocks.

The correlation of expression patterns among various genes, referred to as “co-expression”, can be revealed by network analysis. An increasing number of studies have demonstrated associated behaviour of genes with related biological functions (Carter et al., 2004; Rocke and Durbin, 2001). Given that most biological processes cannot be carried out by a single gene, analysis of co-expressed genes from RNA-seq data sets may be quite informative.

6.5 Gene ontology, enrichment analysis and pathway analysis

Gene ontology (GO) is a controlled vocabulary term to describe gene characteristics in terms of their localization and function. Transcriptome information can be analysed as to what genes are enriched after a certain treatment. This can be accomplished by sequentially analysis of GO, followed by enrichment analysis to determine which GO terms are enriched after the treatment, and then by pathway analysis to determine what gene pathways these enriched genes are involved in. Such analysis can provide functional insights into the induced or suppressed expression patterns.

Gene ontology was initially developed by researchers studying the genome of three model organisms: Drosophila melanogaster (fruit fly), Mus musculus (mouse) and Saccharomyces cerevisiae (yeast) in 1998 (Gene Ontology Consortium, 2010). Now, databases for many other model organisms have joined the Gene Ontology Consortium and made contributions to this project (Gene Ontology Consortium, 2015). The GO project provides three structured ontologies that describe gene products in terms of their biological processes, cellular components and molecular functions in a species independent manner. A number of software packages have been developed for GO analysis. The most popular program for GO analysis is the Blast2GO.

Gene enrichment analysis compares the input gene set with the reference to determine if it is enriched. Gene set enrichment analysis is a method to identify classes of genes or proteins that are over-represented in a set of genes or proteins (Subramanian et al., 2005). The method uses statistical approaches to identify significantly enriched or depleted groups of genes. The principal foundation of enrichment analysis is that the gene set should have a higher chance to be selected if its underlying biological process is abnormal under a given condition (Huang, Sherman and Lempicki, 2009). A number of software packages have been developed for gene enrichment analysis. Gene set enrichment analysis is the most popular approach for enrichment analysis. Following the enrichment analysis, gene pathway analysis can be conducted to determine what pathway is operating under the condition of analysis.

6.6 Analysis of allele-specific expression

A diploid organism has two sets of chromosomes and thereby two alleles at a given locus. Allele-specific expression (ASE) refers to the phenomenon that the two alleles are not equally expressed, up to the exclusive expression of only one of the two alleles. A number of recent studies have demonstrated that allele-specific gene expression is common (Lo et al., 2003; Yang et al., 2003; Chen et al., 2016). Interest in the existence of ASE in non-imprinted autosomal genes
has increased with awareness of the important role that variation in non-coding DNA sequences can play in determining phenotypic diversity (Knight, 2004).

A number of approaches have been used for the detection of ASE. In early studies, ASE was detected by a single-base extension of a primer adjacent to the variable single nucleotide polymorphism (Carrel and Willard, 2005; Cowles et al., 2002). Several recent studies applied a variety of technologies to scale up the tested genes (Guo et al., 2008), of which the array-based approach was the most widely used. A number of array-based ASE studies have been published in the past decade (e.g. Bjornsson et al., 2008; Daelemans et al., 2010). Owing to rapidly increasing throughput and decreasing costs, next-generation sequencing is rapidly replacing array-based technology for functional genomic assays (Rozowsky et al., 2011). In addition, the ability to resolve single-base differences, digital quantification, and comprehensive genome-wide coverage provides information on the abundance and the allelic biases in transcripts or regulatory DNA, which otherwise could not be achieved using hybridization-based arrays (Wood et al., 2015). RNA-seq technology using high-throughput sequencing platforms allows for relatively unbiased measurements of expression levels across the entire length of a transcript. This technology has several advantages, including the ability to detect transcription of unannotated exons and measure both overall and exon-specific expression levels and assay allele-specific expression (Pickrell et al., 2010). Notably, RNA-seq is the only technology that provides concurrent allelic and total expression data.

Measuring ASE is vital to better understanding global mechanisms of genetic variations. ASE analysis has been widely employed in mammals, insects and plant systems (Bell et al., 2013; Combes et al., 2015; Gregg et al., 2010; Serre et al., 2008; Shi et al., 2012; Springer and Stupar, 2007; Wittkopp, Haerum and Clark, 2008). In aquaculture species, despite many reports regarding expression of specific genes, very little is known about ASE (Murata, Oda and Mitani, 2012; Shen et al., 2012). Recently, Chen et al. (2016) reported that ASE is highly enriched with ribosomal protein genes. With more and more application of RNA-seq in aquaculture species, it is inevitable that it will be used for the analysis of ASE.

7. UNDERSTANDING THE NON-CODING PORTIONS OF THE GENOMES

During the first two decades of molecular biology research in the 1970s and 1980s, it was widely believed that only a small fraction, 1–5 percent of the genome, was transcribed. Part of the reason for this was due to technological limitations for the detection of transcripts expressed at low levels. However, this notion was challenged by the discovery of new classes of regulatory non-coding RNAs (ncRNAs). As such, the term “transcript” is now used in a broader sense, from the original usage to mostly refer to the protein-encoding messenger RNA (mRNA), now to various transcriptional products that cover almost the entire genome. Thus, the concept of pervasive transcription evolved to include various types of RNAs, in addition to the traditional mRNA, rRNA (ribosomal RNA) and tRNA (transfer RNA), now to also include various non-coding RNAs (Mercer, Dinger and Mattick, 2009). The proportion of such non-coding RNAs may vary among species, but are becoming more and more represented. In humans, for instance, the current research revealed that only one-fifth of the transcription across the human genome is associated with protein-coding genes (Kapranov et al., 2007), suggesting at least four times more non-coding than coding RNA sequences.
Various types of non-coding RNAs have been identified, including long non-coding RNAs (lncRNAs), microRNAs (miRNAs), short interfering RNAs (siRNAs), Piwi-interacting RNAs (piRNAs), and small nucleolar RNAs (snoRNAs) (Kapranov et al., 2007).

### 7.1 Long non-coding RNAs

Among the non-coding RNAs, lncRNAs are non-coding RNAs whose sizes are greater than 200 bases. Obviously, such classification is arbitrary, but is based on practical considerations, including the separation of RNAs in common experimental protocols.

The functions of non-coding RNAs are being unraveled and the discoveries are continuing. Among the non-coding RNAs, the functions for the small regulatory non-coding RNAs such as miRNA are probably the best studied (Bartel, 2009). The functions for lncRNAs are not well understood, but recent research has indicated that lncRNAs could be involved in a number of functions (Huarte and Rinn, 2010; Pauli et al., 2012), including: (i) interacting with and modulating the activity of the chromatin modifying machinery (Huarte et al., 2010; Nagano et al., 2008; Rinn et al., 2007; Tian, Sun and Lee, 2010); (ii) serving as the decoys in the sequestration of miRNAs (Poliseno et al., 2010), transcription factors (Hung et al., 2011), or other proteins (Tripathi et al., 2010); and (iii) serving as precursors for the generation of small RNAs (sRNAs) (Kapranov et al., 2007). In addition to the co-expression of lncRNAs and mRNAs, the co-localized expression of lncRNA and protein coding genes was also observed (Ponjavic et al., 2009), suggesting their cooperative actions and/or sharing of cis-regulatory elements in the transcription process. In many instances, the act of lncRNA transcription alone is sufficient to regulate the expression of nearby genes (e.g. Martens, Laprade and Winston, 2004; Petruk et al., 2006; Wilusz and Sharp, 2013), or distant genes through modification of chromatin complexes (e.g. Tsai et al., 2010), or binding to transcription elongation factors (Yang, Froberg and Lee, 2014).

Analysis of lncRNA in aquaculture species is very limited. A few studies are being conducted in rainbow trout and catfish (Al-Tobasei, Paneru and Salem, 2016; authors’ unpublished data), but the studies are still in the stage of infancy.

### 7.2 MicroRNAs and their target genes

Several distinct classes of small non-coding RNAs, including miRNA, siRNA, piRNA, and repeat-associated short interfering RNA (rasiRNA), have been identified. These molecules are typically ~18–40 nucleotides in length, and play profound roles in many cellular processes. The miRNAs, with a length of ~22 nucleotides, play critical roles in post-transcriptional regulation of gene expression.

A number of studies have been conducted for the analysis of microRNA with aquaculture species. These include analysis in tilapia (Yan et al., 2012a, 2012b), sea cucumber (Li et al., 2012), Atlantic cod (Johansen et al., 2011), and channel catfish (Barozai et al., 2012). However, functional analysis of their target genes in aquaculture species is rarely existent.

### 7.3 ENCODE project and FAANG project

After the assembly of the human genome sequence, in 2003, the National Human Genome Research Institute in the United States of America launched the ENCODE (Encyclopedia of DNA Elements) project. The objective of ENCODE was to obtain a complete list of the functional ele-
ments of the human genome, including the elements that act at the protein and RNA level, as well as the regulatory elements for transcription, post-transcriptional regulation, translation and replication. This type of project has been extended to annotate the animal genomes, including those of aquaculture species, and this project is called FAANG (Functional Annotation of ANimal Genomes, www.faang.org/plan).

8. GENETIC ANALYSIS TECHNOLOGIES

8.1 Traits important for aquaculture
The practical purpose of aquaculture genomics and genetics studies is to reveal the genetic basis of performance and production traits and use such information for genetic breeding programmes. With aquaculture species, domestication is a very recent event for many species. Therefore, phenotypic variations are huge among strains, lines and individuals within strains or lines. Although strain selection is an effective strategy, modern breeding programmes focus on selection within strains.

A number of traits are important for aquaculture production. These include growth rates, feed conversion efficiency, disease resistance, low oxygen tolerance, stress tolerance, processing yields, sexual maturation time, robustness, body conformation and reproductive traits, among many others. In catfish, phenotypic data sets have been produced for various traits that are important for aquaculture. Such data sets, along with genetic pedigrees, are important resources for genetic and QTL mapping as well as genome-wide association studies.

8.2 Quantitative trait locus mapping in aquaculture species
The fundamental goal of aquaculture genomics in the practical sense is to understand the genomic basis for performance and production traits. Because most aquaculture traits are complex traits that are likely controlled by multiple genes, QTL mapping is the core of applied aquaculture genomics. In recent years, great efforts have been made in this area and good progress has been made. QTL analyses have been conducted in several dozens aquaculture species. Some typical examples are summarized in Table 6. The studied traits include growth rate, disease resistance, sex maturation time, body conformation, fat content, response to stress, swimming abilities, salinity tolerance, muscle traits, osmoregulation capacities and smoltification, among other traits. Of these, the largest amount of efforts has been devoted to QTL mapping of growth traits and disease resistance.

<table>
<thead>
<tr>
<th>Species</th>
<th>Traits</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arctic charr</td>
<td>Body weight, condition factor and age of sexual maturation</td>
<td>Moghadam et al., 2007; Küttner et al., 2011</td>
</tr>
<tr>
<td>Arctic charr</td>
<td>Salinity tolerance</td>
<td>Norman et al., 2011</td>
</tr>
<tr>
<td>Asian seabass</td>
<td>Growth</td>
<td>Wang et al., 2006; 2011</td>
</tr>
<tr>
<td>Asian seabass</td>
<td>Resistance against viral nervous necrosis disease</td>
<td>Liu et al., 2015a</td>
</tr>
<tr>
<td>Species</td>
<td>Traits</td>
<td>Reference</td>
</tr>
<tr>
<td>----------------------</td>
<td>---------------------------------------------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Asian seabass</td>
<td>Omega-3 fatty acids contents</td>
<td>Xia et al., 2014</td>
</tr>
<tr>
<td>Atlantic salmon</td>
<td>Body weight and condition factor</td>
<td>Reid et al., 2005</td>
</tr>
<tr>
<td>Atlantic salmon</td>
<td>Adaptive traits</td>
<td>Boulding et al., 2008</td>
</tr>
<tr>
<td>Atlantic salmon</td>
<td>Growth</td>
<td>Baranski, Moen and Våge, 2010</td>
</tr>
<tr>
<td>Atlantic salmon</td>
<td>Resistance against infectious pancreatic necrosis virus</td>
<td>Gheyas et al., 2010; Houston et al., 2010; Houston et al., 2008; Moen et al., 2009</td>
</tr>
<tr>
<td>Atlantic salmon</td>
<td>Resistance against infectious salmon anaemia</td>
<td>Moen et al., 2007</td>
</tr>
<tr>
<td>Atlantic salmon</td>
<td>Flesh colour</td>
<td>Baranski, Moen and Våge, 2010</td>
</tr>
<tr>
<td>Atlantic salmon</td>
<td>Life history</td>
<td>Vasemägi et al., 2010</td>
</tr>
<tr>
<td>Atlantic salmon</td>
<td>Resistance to pancreas disease</td>
<td>Gonen et al., 2015</td>
</tr>
<tr>
<td>Atlantic salmon</td>
<td>Late sexual maturation</td>
<td>Gutierrez et al., 2014</td>
</tr>
<tr>
<td>Blacklip abalone</td>
<td>Growth</td>
<td>Baranski et al., 2008</td>
</tr>
<tr>
<td>Catfish</td>
<td>Columnaris disease resistance</td>
<td>Geng et al., 2015</td>
</tr>
<tr>
<td>Catfish</td>
<td>Head size and shape</td>
<td>Authors’ unpublished data</td>
</tr>
<tr>
<td>Clam</td>
<td>Clam metrix</td>
<td>Lu et al., 2013</td>
</tr>
<tr>
<td>Coho salmon</td>
<td>Hatch timing, weight, length and growth</td>
<td>McClelland and Naish, 2010</td>
</tr>
<tr>
<td>Common carp</td>
<td>Muscle fiber-related</td>
<td>Zhang et al., 2011b</td>
</tr>
<tr>
<td>Common carp</td>
<td>Growth rate</td>
<td>Boulton et al., 2011</td>
</tr>
<tr>
<td>Common carp</td>
<td>Swimming ability</td>
<td>Laghari et al., 2015</td>
</tr>
<tr>
<td>Eastern oyster</td>
<td>Disease resistance</td>
<td>Yu and Guo, 2006</td>
</tr>
<tr>
<td>European seabass</td>
<td>Growth</td>
<td>Louro et al., 2016</td>
</tr>
<tr>
<td>European seabass</td>
<td>Body weight, morphometric traits and stress response</td>
<td>Massault et al., 2010</td>
</tr>
<tr>
<td>Gilthead sea bream</td>
<td>Sex determination and body growth</td>
<td>Loukovitis et al., 2011</td>
</tr>
<tr>
<td>Gilthead sea bream</td>
<td>Resistance to fish pasteurellosis</td>
<td>Massault et al., 2011</td>
</tr>
<tr>
<td>Gilthead seabream</td>
<td>Skeletal deformities</td>
<td>Negrín-Báez et al., 2015</td>
</tr>
<tr>
<td>Japanese flounder</td>
<td>Vibrio anguillarum resistance</td>
<td>Wang et al., 2014</td>
</tr>
<tr>
<td>Kelp grouper</td>
<td>Growth</td>
<td>Kessuwan et al., 2016</td>
</tr>
<tr>
<td>Large yellow croaker</td>
<td>Growth</td>
<td>Ye et al., 2014</td>
</tr>
<tr>
<td>Pacific abalone</td>
<td>Growth-related</td>
<td>Liu et al., 2007</td>
</tr>
<tr>
<td>Pacific oyster</td>
<td>Growth</td>
<td>Guo et al., 2012</td>
</tr>
<tr>
<td>Pacific oyster</td>
<td>Resistance against summer mortality</td>
<td>Sauvage et al., 2010</td>
</tr>
<tr>
<td>Pacific oyster</td>
<td>Growth</td>
<td>Guo et al., 2012</td>
</tr>
<tr>
<td>Rainbow trout</td>
<td>Upper thermal tolerance</td>
<td>Jackson et al., 1998; Danzmann, Jackson and Ferguson, 1999; Perry et al., 2001; Perry et al., 2005</td>
</tr>
<tr>
<td>Rainbow trout</td>
<td>Life history</td>
<td>Leder, Danzmann and Ferguson, 2006</td>
</tr>
</tbody>
</table>
### Species | Traits | Reference
---|---|---
Rainbow trout | Spawning time | O’Malley et al., 2003; Colihueque et al., 2010
Rainbow trout | Osmoregulation capacities | Le Bras et al., 2011
Rainbow trout | Development rate | Robison et al., 2001; Easton et al., 2011
Rainbow trout | Whirling disease resistance | Baerwald et al., 2011
Rainbow trout | Growth | Wringe et al., 2010
Rainbow trout | Smoltification | Nichols et al., 2008
Rainbow trout | Bacterial cold water disease resistance | Vallejo et al., 2014a; Palti et al., 2015b
Rainbow trout | Flavobacterium psychrophilum resistance | Vallejo et al., 2014b
Rainbow trout | Osmoregulation capacity | Le Bras et al., 2011
Rainbow trout | Cortisol response to crowding | Liu et al., 2015b
Rainbow trout | Response to crowding stress | Rexroad et al., 2013
Tilapia | Sex determination | Cnaani et al., 2007; Shirak et al., 2006
Turbot | Growth | Sánchez-Molano et al., 2011
Turbot | Aeromonas resistance | Rodríguez-Ramilo et al., 2011
Turbot | Resistance against Philasterides | Rodríguez-Ramilo et al., 2013
Turbot | Resistance to viral haemorrhagic septicaemia | Rodríguez-Ramilo et al., 2014
Turbot | Sex differentiation | Viñas et al., 2012
Zhikong scallop | Size-related traits | Zhan et al., 2009

### 8.3 Quantitative trait locus analysis of disease resistance

Disease resistance is among the most important traits for aquaculture species. Therefore, great efforts have been made in mapping QTLs controlling disease resistance.

The QTL effect for disease resistance varies. In some cases, major QTLs have been found that suggest a single gene or just a few genes may be operating, and in these situations, the mapped QTL is very useful for marker-assisted selection. For instance, several QTLs were mapped in Atlantic salmon for resistance against the infectious pancreatic necrosis (IPN) virus. One QTL on linkage group 21 can explain 25 percent of the observed within family variance in the overall data set, while the QTL on linkage groups (LGs) 26 and 19 were estimated to explain 18 and 9 percent of the variance, suggesting that these QTLs are major.

In other situations, multiple QTLs were mapped with each having a relatively small effect. Most disease resistance QTLs have a relatively small effect, suggesting many genes are involved in the resistance. In addition, it may also suggest that the phenotypic evaluation is difficult and the environment effect may be large such that the percentage of phenotypic variation explained by the QTL is small. For instance, in Asian seabass, viral nervous necrosis disease causes mass mortality in mariculture. Using 149 microsatellites, Liu et al. (2015a) successfully mapped five significant QTLs located in four LGs and eight suggestive QTLs in seven LGs; another five significant QTLs in three LGs and five suggestive QTLs in three LGs were detected for survival time. These
results suggested that viral nervous necrosis resistance in Asian seabass is controlled by many loci with small effects. The QTL with the largest effect accounted for only 2–4 percent of the phenotypic variations. In Atlantic salmon, a QTL was mapped to linkage group 8 that accounted for 6 percent of the phenotypic variation (Moen et al., 2007).

### 8.4 Genome-wide association studies of performance traits

Genome-wide association study (GWAS) is another method for mapping genes involved in performance traits. GWAS is different from QTL mapping in several aspects: (i) QTL analysis requires genetically structured families with pedigree information, while GWAS typically uses genetically unrelated individuals; and (ii) QTL mapping relies on the detection of genetic linkage, while GWAS depends on the detection linkage disequilibrium between the trait under study and the related markers.

GWAS has been extensively used for genetic analysis of genetic diseases in humans. It has been also used for aquaculture species. For instance, Geng et al. (2015) used GWAS to identify genes associated with disease resistance against columnaris disease in catfish (Geng et al., 2015). In another study, Tosh (2014) used GWAS to identify genes involved in harvest weight. Most recently, GWAS was used to identify associated markers with fillet yield in rainbow trout (Gonzalez-Pena, 2016). However, the application of GWAS in aquaculture species is still at the earliest stage.

### 9. GENOME-BASED GENETIC SELECTION TECHNOLOGIES

#### 9.1 Marker-assisted selection

Marker-assisted selection is also called marker-aided selection (MAS) and is a process whereby a selection decision is made based on the genotypes of DNA markers. MAS is especially useful for traits that are difficult to measure, lethal to measure, exhibit low heritability, and/or are expressed late in development. Its implementation requires information of DNA markers that are tightly linked to QTL for traits of interest based on QTL mapping or association studies (Lande and Thompson, 1990). Ideally, the DNA markers should be the causative mutation underlying the phenotypic variation. However, for practical purposes, it would not make a difference if the DNA marker is always linked with the trait of interest even if the marker variation is only correlated, but not the cause of the phenotypic difference. In order to implement MAS, QTLs need to be mapped and validated within the breeding populations. MAS has been applied mostly with plants and livestock animal species, but less so with aquaculture species, although a few good examples do exist for application in aquaculture (Ozaki et al., 2012).

The best example of MAS in aquaculture species is perhaps the situation of Japanese flounder. A microsatellite locus, Poli9-8TUF, was mapped to be near the major QTL for resistance to lymphocystis disease. Additional analysis indicated that the disease resistance was controlled by a single gene, and the resistance allele was dominant. Based on the marker linkage information, Fuji et al. (2007) developed a new population of Japanese flounder using MAS with the marker Poli9-8TUF. They selected a female homozygous for the favourable allele (B-favourable) and a male with a higher growth rate and good body shape, but without the resistant allele as parents. In the females, the marker Poli9-8TUF is tightly linked to the QTL for resistance to lymphocystis disease; therefore, a female was selected as the linkage disequilibrium-resistant parent. The
B-favourable allele was transmitted from the mother to the progeny. All the progeny are heterozygotes with the resistance allele, and the progeny was entirely resistant to lymphocystis disease, while the control group without B-favourable alleles showed incidences of 4.5 and 6.3 percent of mortality due to lymphocystis disease. These results clearly demonstrate that MAS is an efficient strategy for breeding. MAS lymphocystis disease-resistant flounder had a market penetration rate of 35 percent in Japan in 2012 (Ozaki et al., 2012).

Another good example of MAS is the selection of IPN resistance. In salmon, IPN is a major problem. One major QTL was mapped to linkage group 21, which accounts for 29 percent and 83 percent of the phenotypic and genetic variances, respectively. Three microsatellite markers were tightly linked to the QTL, and these markers have been used for the selection of IPN resistance (Moen et al., 2009).

It should be noted that MAS refers to marker-assisted selection, not marker selection. This means that markers are used to supplement the routine selective breeding programmes rather than replacing them. Although MAS is theoretically very sound and attractive, little is known about the economic benefits gained from MAS in aquaculture species, with the exception of the above cases where the phenotypes were controlled by a single gene rather than by many genes. Information of this nature is important because the additional genetic gains depend on the magnitude of the allelic effects, and thus the marginal increase should offset the costs of applying the technology (e.g. genotyping and labour costs).

9.2 Sex markers and their applications
In aquaculture settings, it is often important to know the sex information of fish and shellfish species. In addition to biological interest of sex determination and regulation, sex is of interest to aquaculturists because of sexual bimorphism. With most aquaculture species, one sex grows significantly faster than the other sex. For instance, females of soles, eels and many other species grow much faster than the males. In contrast, males grow much faster for tilapia and catfish, among many other species. In addition to growth rate, sex also affects body shape, colouration and carcass composition (Beardmore, 2001; Cnaani, 2009).

Sex-linked markers have been mapped for many aquaculture species. For instance, AFLP and microsatellite markers have been identified to be linked with sex in various aquaculture species, including common carp (Chen et al., 2009), tilapia (Lee, Penman and Kocher, 2003), catfish (Ninwichian et al., 2012a, 2012b), Zhikong scallops (Li et al., 2005), half-smooth tongue sole (Chen et al., 2007), white shrimps (Pérez et al., 2004), kuruma prawns (Li et al., 2003), and rainbow trout (Felip et al., 2005). These sex-linked markers have been useful for the identification of sex without phenotypic data.

9.3 Genome selection
Recent advances of genome analysis, including the availability of a large number of polymorphic markers, highly efficient genotyping platforms such as SNP arrays and the application of next-generation sequencing technologies, have allowed mapping of dense markers across the entire genome, which in turn enables an estimation of the genetic merit of every chromosome fragment contributing variation in a population with phenotypic observations. Not only can the merit of every chromosomal segment be estimated, but also all the traits of interest can be estimated simultaneously. Whole genome selection is based on such abilities of estimating the
value of every chromosome fragment contributing variation in a population with phenotypic observations (training); the results of training can be used to predict the merit of new animals that have genotypes but are not included in the training data set.

Genomic selection was first proposed by Meuwissen et al. (2001). Since then, it has gained a tremendous level of attention in the animal genetics community. Genomic selection is a form of marker-assisted selection in which genetic markers covering the whole genome are used so that all quantitative trait loci are in linkage disequilibrium with at least one marker. Simulation results and experimental results suggest that breeding values can be predicted with high accuracy using genetic markers alone, but more validation is required, especially in the samples of the population different from that in which the effect of the markers was estimated (Goddard and Hayes, 2007). Compared with MAS, genomic selection uses the estimated effect of many loci across the entire genome at once, not just the small number of linked loci as done with MAS, or individual locus as in GWAS.

Although genomic selection has been successfully used in dairy cow and beef cattle and other livestock species (Hayes et al., 2009), its use in aquaculture species has been slow (Ragavendran and Muir, 2011; Sonesson, 2011).

10. GENOME EDITING TECHNOLOGIES

10.1 Zinc finger nuclease and history of genome editing technologies

Genome editing refers to the ability to make specific changes at targeted genomic sites. The history for the development of genome editing technologies has been well summarized by Nemudryi et al. (2014). The zinc finger nuclease (ZFN) technology was developed in 1996. A zinc finger protein domain coupled with the FokI endonuclease domain was demonstrated to act as a site-specific nuclease cutting DNA at strictly defined sites in vitro (Kim, Cha and Chandrasegaran, 1996). This chimeric protein has a modular structure, with each zinc finger domain recognizing one nucleotide triplet. This method was used for genome editing of cultured cells of both plants and animals, including pluripotent stem cells (e.g. Bibikova et al., 2002; Townsend et al., 2009; Provasi et al., 2012). However, the ZFN-based technology has a number of disadvantages. It is complex to use, has a high cost, and the cleavage site is not accurate.

Active searches for better methods have been a major focus of research in the last several years. That has led to the development of new genome editing technologies, such as the transcription activator-like effector nucleases (TALENs) and the clustered regularly interspaced short palindromic repeats CRISPR-associated (CRISPR/Cas). These new genome editing technologies overcome the disadvantages of ZFN technology and have become very efficient for the modification of genomes through genome editing.

10.2 TALEN

Transcription activator-like effector nucleases (TALEN) are restriction enzymes that can be engineered to cut specific sequences of DNA. They are made by fusing a transcription activator-like effector DNA-binding domain to a DNA cleavage domain. Transcription activator-like effectors can be engineered to bind practically any desired DNA sequence, so when combined with the cleavage domain (a nuclease), DNA can be cut at specific locations (Boch et al., 2009). The restriction enzymes, when being introduced into cells, can be used for gene editing or genome
editing. Along with zinc finger nucleases and CRISPR-associated protein-9 nuclease (Cas9) proteins, TALEN is becoming a prominent tool in the field of genome editing.

Genome editing starts with efficient generation of a double-strand break (DSB) in the target DNA. DSBs are repaired either by homologous recombination, or, in the absence of a homologous repair template, via non-homologous end joining (NHEJ), which causes small insertions or deletions (INDELS) as the broken ends are ligated together. The creation of INDELS is exploited as a convenient method for a gene knockout. Both TALEN and CRISPR can edit DNA through either NHEJ or homologous recombination. TALEN is comprised of a pair of DNA binding proteins fused to the FokI nuclease, while CRISPR is a complex between the Cas9 nuclease and a target-specific single guide RNA (sgRNA).

TALEN and CRISPR/Cas9 endonucleases have provided ways to significantly improve genome editing efficiency. These endonucleases make a DSB at a predetermined DNA sequence and trigger natural DNA repair processes such as NHEJ or homologous recombination with a donor DNA template. Among these existing approaches, RNA-guided CRISPR/Cas9 is the most user-friendly and versatile system.

10.3 CRISPR/Cas9
CRISPR is abbreviated from clustered regularly interspaced short palindromic repeats, which are bacterial DNA containing short palindrome repeats and such repeats are regularly spaced. CRISPR-associated (Cas) proteins are nucleases that cut DNA. CRISPR is a mechanism of bacterial immunity against invading viruses or plasmids. Its mechanism of actions is similar to eukaryotic RNA interference pathways.

This system was first used as a genome editing system in 2012. Since then, CRISPR/Cas9 has been used for modification of genomes in many species, including zebrafish (Hwang et al., 2013; Jao, Wente and Chen, 2013). The CRISPR works by three components: CRISPR RNA (crRNA), which binds the target DNA and guides cleavage, and the trans-activating RNA (tracrRNA), which base pairs with the crRNA and enables the Cas9-crRNA complex to locate the targeted DNA and the Cas9 nuclease. In a typical situation, the Cas9 nuclease is transiently expressed using a promoter suitable to the cells for which genome editing is for, while the RNA components are co-transfected or injected with the expression construct of Cas9. The Cas9 is an enzyme that cuts DNA, and CRISPR is a collection of DNA sequences that tells Cas9 where to cut.

10.4 Comparison of TALEN and CRISPR/Cas9
TALEN and CRISPR differ in four aspects: (i) specificity; (ii) target selection; (iii) efficiency; and (iv) simplicity of construction. CRISPR achieves its specificity through the single guide RNA (sgRNA), which is an artificial fusion of two naturally occurring short RNAs (Jinek et al., 2012). The sgRNA directs the Cas9 nuclease to a 20 nucleotide target site on the chromosome, which must be immediately followed by an N-G-G trinucleotide known as the protospacer adjacent motif, or PAM. The sgRNA hybridizes with the strand opposite the PAM site, and Cas9 nuclease cuts the DNA. In this process, sgRNAs can tolerate up to five mismatches to guide mutagenesis to off-target sites (Fu et al., 2013).

Compared with CRISPR, TALEN has higher levels of specificity. A TALEN pair (each ∼18 base pair long, total 36 base pairs) must bind on opposite sides of the target site, separated by a “spacer”
ranging from 14 to 20 nucleotides. The target to match the 18 base pairs each on the opposite strand of the chromosome is expected to be unique, thus providing high specificity. In terms of target selection, both TALEN and CRISPR are quite flexible, but TALEN is more flexible. TALEN and CRISPR have a comparable efficiency, but a slightly higher efficiency (up to 70 percent) was reported for CRISPR. TALEN may be sensitive to cytosine methylation within CpG dinucleotides. Among all the features, CRISPR is simple to design and use compared with TALEN or ZFN technologies. Therefore, CRISPR is gaining its popularity in applications to various species systems. For each target site, all that is needed for CRISPR is to program a 20 nucleotide genomic target site into the overall sgRNA. Plasmid construction is straightforward and simple. For editing experiments, the sgRNA is co-expressed with the reusable Cas9 nuclease. In contrast, TALEN construction involves re-engineering a new protein for each target.

11. CHALLENGES FOR THE APPLICATIONS OF GENOME-BASED TECHNOLOGIES IN AQUACULTURE

As discussed above, genome-based technologies include DNA marker technologies, genome mapping technologies, sequencing technologies and genome editing technologies. To a certain extent, these technologies have been used in aquaculture species, but the potential for their applications is tremendous.

The most practically important genome-based technologies are probably genome editing technologies involving ZFN, TALEN or CRISPR/Cas9. Using such technologies allows precise knockout of genes, gene modification, or targeted gene insertion (Lauth et al., 2012). Such technologies have been used in cattle, pigs, rabbits and catfish, as well as in zebrafish (Doyon, 2008; Meng, 2008; Hauschild, 2011; Yu, 2011; Flisikowska, 2011; Dong, 2011).

With the rapid advances of genomic research and the development and applications of genome-based technologies, a number of challenges are facing the world. These include, but not limited to, the following.

11.1 Discoupling of genomics with breeding programmes

In the past two decades, huge progress has been made in the areas of genomic research and development of genome-based technologies. Applications of such technologies have made large strides with plants and livestock animal species; however, applications of genomic technologies in aquaculture species have been limited. Partly this was because of the uncoupling of genome research with breeding programmes. With the exception of Atlantic salmon and perhaps rainbow trout, where private corporations run major breeding programmes in which genomic information and genome-based technologies are more commonly used, the use of genomic information and genome-based technologies is very limited in other aquaculture species. A detailed examination of the situations indicated that lack of major breeding companies, public or private, may be a major bottleneck for breeding programmes of many aquaculture species, including various species such as common carp, grass carp, crucian carp, black carp, tilapia, catfish and various shellfish species. In shrimps, private or public institutions exist for breeding, but their efforts have been focused on the production of specific pathogen-free stocks rather than breeding for disease resistance using genomic information or genome-based technologies. Therefore, resources must be allocated from the government and other sources to support the breeding
programmes of aquaculture species. This is not only an issue of aquaculture production, but also an issue of environment and sustainability because more efficient use of aquatic resources will have a very positive impact on natural fisheries. Great progress will be made only when genomics research is well coupled with breeding programmes.

11.2 Bioinformatics challenges

Genomics relies on the ability to analyse large data sets through bioinformatic data mining, data analysis, data sharing, meta-analysis and data re-analysis. Of all the expertise, bioinformatics is key to the success of genome research. The development and application of genome-based technologies depend on the ability to analyse the impact of such applications on economics. Students of biology usually have limited background for informatics, and informatics students have limited background in biology. Training scientists with combined expertise is truly a great challenge. This challenge is paramount for aquaculture because aquaculture workers tend to have even less understanding of informatics.

This raises a question as to what kind of scientists should be trained today as the next generation of scientists. Decades ago, graduate students spent most of their graduate school studies on hands-on experiments. Data analysis was a minor part of the graduate life. Over time, more and more time is demanded of data analysis. For genomics students, they may have to spend most of their time doing data analysis using supercomputers. Yet, these students must also have all the biological aspects of training to understand the biology of the organisms they are working on. To be able to keep a balance between these things is truly a challenge.

11.3 Computational limitations

We are now living in the Big Data era, like it or not. The largest bottleneck will be computational challenges. With next-generation sequencers, terabytes of data can be readily generated in any of the simple experimentations. The key is the ability to analyse such large data sets. In most cases, this must be handled through the use of supercomputers, or high-performance computer clusters (HPC). This raises two lines of challenges: the first is that HPC computers mostly use Unix or Linux platforms that are less user-friendly and demand the users to have basic knowledge of command lines; the second is that the purchasing of HPC computers is very costly, most often costing over US$1 million. Even more difficult is the maintenance and update of the HPCs as they become sort of obsolete after a few years of service. This challenge is difficult to manage even in the United States of America, and it is apparent that such a challenge would be almost impossible for some poor developing countries. Though the opportunities to use cloud computing are available, it requires a certain level of infrastructure and information technology resources to be able to undertake bioinformatic analyses.

11.4 Funding challenges for aquaculture workers

Although funding limitation is a universal problem, it is a greater problem for aquaculture because aquaculture species represent small commodities, as compared with major plant or animal species. To study the genome or develop genome-based technologies, the cost would be the same when dealing with different species. However, funding levels are most often dictated by the commodity’s importance. Thus, aquaculture workers must find creative ways of getting the same work done with fewer resources.
11.5 Unbalanced research advances in the world

Genomic sciences are regarded as “big science” that requires big investment. As a result, there is a huge imbalance in the geographical distribution of genomic studies in the world, especially with aquaculture species. The major research projects of aquaculture genomics are in a dozen countries, such as Brazil, Canada, Chile, China, Greece, Japan, India, Norway, Singapore, Spain and the United States of America. Genome projects, however, are rare in Africa. In a way, the geographical distribution of fish genome projects is correlated with geographical distribution of the application of genome-based technologies in aquaculture species.

11.6 Challenges of working with small-scale farmers

It is relatively easy to deal with species for which the industries are vertically integrated. For instance, the poultry industry in the United States of America is vertically integrated well. As such, decisions can be made by the top management dealing with various problems and issues in a systems approach, including genetics, nutrition, culture technologies, harvest, processing and marketing. With aquaculture species, situations vary greatly from species to species. The best scenario is perhaps that of Atlantic salmon. In northern Europe such as Norway, breeding companies hold broodstocks that are continuously improved genetically with traditional selective breeding or with genomic selection. With many other species, however, breeding programmes may exist, but systematic selection using genome-based technologies have not been practised. This issue is more serious with small family-based farmers. In some cases, aquaculture is being practised in a primitive way, where fries are captured from the wild, and aquaculture is only to raise the captured fish to food size with artificial feed. No selection using the principles of genetics is involved. In this latter scenario, application of genome-based technologies is remote, if at all possible ever.

11.7 The push for a fast return on investment

Research, especially basic research, takes time. There is always a gap between research and application. This is particularly true with genome sciences, where it can take decades to generate sufficient enough information for possible applications. Therefore, there is a significant gap between research and application. This is well understood by the scientific community, but much less by the public and governments. Governments are often anxious to produce quick results. While that is understandable with public pressure for return of investment, it often backfires if application is sought prematurely. For instance, the application of marker-assisted selection before full understanding of major QTLs can cause harm rather than good for the genetic characteristics of the population under selection.

11.8 Ethical, legislative and regulatory issues

The new line of genomic technologies, especially genome editing technologies, brings risks and ethical challenges, as outlined in FAO’s Statement on Biotechnology (2000). While there is little controversy about many aspects of biotechnology and its application, genetically modified organisms (GMOs) have become the target of a very intensive and, at times, emotionally charged debate. It is generally well recognized that genetic engineering has the potential to help increase production and productivity. However, there is concern about the potential risks posed by certain aspects of biotechnology. These risks fall into two basic categories: the effects
on human and animal health and the environmental consequences. Caution must be exercised in order to reduce the risks of transferring toxins from one life form to another, of creating new toxins or of transferring allergenic compounds from one species to another, which could result in unexpected allergic reactions. Risks to the environment include the possibility of outcrossing, which could lead, for example, to the development of more aggressive weeds or wild relatives with increased resistance to diseases or environmental stresses, upsetting the ecosystem balance. Biodiversity may also be lost, as a result of the displacement of traditional cultivars by a small number of genetically modified cultivars, for example.

There is a strong case to support a science-based evaluation system that would objectively determine the benefits and risks of each individual GMO. This calls for a cautious case-by-case approach to address legitimate concerns for the biosafety of each product or process prior to its release. The possible effects on biodiversity, the environment and food safety need to be evaluated, and the extent to which the benefits of the product or process outweigh its risks assessed. The evaluation process should also take into consideration experience gained by national regulatory authorities in clearing such products. Careful monitoring of the post-release effects of these products and processes is also essential to ensure their continued safety to human beings, animals and the environment.

Although the nature of the issues is perhaps less serious with aquaculture species, ethical and legislative issues are equally serious. For instance, the TALEN or CRISPR technologies allow introduction or knockout of any gene without much difficulty with any fish or shellfish species. The altered genome is able to pass on the genetic material to future generations. While it is clear that the genome editing technologies are different from gene transfer technologies, it is widely believed that government agencies should be regulating any commercial products generated using gene-editing technologies. At this time, it is not clear how stringent such regulation should be because the scientific community is still at the early stages of discussion.

An important aspect concerning the regulation of biotechnology is the issue of Intellectual Property Rights (IPR) are defined as the right to control the commercial exploitation of the projected subject matter for a specific period (FAO 2016). Different forms of IPR exist, such as copyrights or patents, each having different requirements. It was noted that minimum standards for protecting IPR are set by the World Trade Organization Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS) for signatory countries. The use of IPR in agricultural biotechnology has been controversial, especially in developing countries, and have been criticized for a variety of reasons including conflicts with farmers’ traditional practices to reuse seed; excessively broad patent claims; patentability of genetic material and plant varieties; uncertainty regarding the scope of research use; high transaction costs; pricing of improved varieties (high because it has to cover costs of licensing of IPR); and appropriation of traditional knowledge and sovereign genetic resources. In many instances, it is not IPR per se that impede diffusion of agricultural biotechnologies in developing countries but other confounding issues are involved. IPR are legal tools to arrange a licensing agreement; there are other legal tools available to handle conflicts raised within IPR (e.g. consumer protection legislation), which should be activated whenever needed. National biotechnology policies and legislation should encompass aspects related not only to the safety of biotechnology products but also to their ownership.

Levels of regulation on agricultural biotechnology, specifically genetically engineered organisms, are diverse between countries. In most developed countries genetically engineered organ-
isms are well regulated by various government agencies. As one example, in the United States of America, both the Food and Drug Administration (FDA) and the United States Department of Agriculture (USDA) regulate GMOs. Within the USDA’s Animal and Plant Health Inspection Service, there is the Biotechnology Regulatory Services (https://www.aphis.usda.gov/aphis/our-focus/biotechnology). The FDA is responsible for ensuring the safety and proper labelling of all plant- and animal-derived food and feed, including those developed through genetic engineering. The typical struggle GMOs have to go through is well illustrated by the decade-long fight over the marketing of transgenic salmon by AquaBounty Technologies.

While similar regulatory agencies exist in many countries, regulation may vary across the world. International regulation of biotechnology is well described by Matthias Herdegen (http://legal.un.org/avl/pdf/is/Herdegen_slideshow.pdf). Concerns for international regulation are focused on risk, free trade and development perspectives in the field of human rights, environment protection and international trade law. For environment protection, the United Nations Convention on Biological Diversity and the Cartagena Protocol on Biosafety are guiding documents for actions. However, with the advanced genome editing technologies, it is probably necessary to have stronger international regulations, as well as those regulations placed by governments in various countries.

The increased convergence between biological and other sciences, higher investment requirements, the higher profile of intellectual property and biosafety issues, the changed role of the private sector both in the development of the technologies and the technology delivery systems, are all aspects that should be clearly present in an effective policy development process. It is important to consider also that the development of new biotechnologies is progressing extremely quickly, and the state-of-the-art is changing faster and faster. As such, flexible and forward-looking regulatory and legislative frameworks will increasingly be needed.

11.9 Public perception
There is a clear need for a common understanding on vocabulary and definitions of the terminology used for agriculture biotechnologies (e.g. what is meant by genetic modification, genetic engineering or a GMO?) for an informed discussion, and this common understanding is crucial for public perception (FAO 2016).

Public perception and acceptance has always been an issue to date with genetically modified organisms (GMOs). However, in most cases, GMOs have been created by genetic engineering. The term “agricultural biotechnologies” is clearly broader than GMOs. Any aquaculture products produced using genome editing technologies would encounter such an issue of public perception. Although the exact intensity and the nature of the concerns are unknown at present, it is expected that some levels of negative public perception may exist for aquaculture products created using genome editing technologies. As such, the level of public acceptance for products created by genome editing technologies may be lower than those produced using traditional breeding techniques.

11.10 Technology transfer
Current investment in biotechnological research tends to be concentrated in the private sector and oriented towards agriculture in higher-income countries where there is purchasing power for its products. In view of the potential contribution of biotechnologies for increasing food
supply and overcoming food insecurity and vulnerability, efforts should be made to ensure that developing countries, in general, and resource-poor farmers, in particular, benefit more from biotechnological research, while continuing to have access to a diversity of sources of genetic material.

Although technologies are different, it appears that the technology transfer part of the equation has not been changed by the development of genome-based technologies. It will all depend on the patents and other protective measures placed on a particular product or technology.

12. CONCLUDING REMARKS

The historical Human Genome Project was officially launched in 1990, and the project lasted more than a dozen years when the human genome was sequenced and assembled in 2004. This multi-billion dollar project, being started with a simple objective of just the sequence of the entire DNA composition or the genome of a single person, ended with massive achievements in biology, ranging from technology development, to methodology breakthrough, to the emergence of the systems biology that takes an entirely different approach changing the traditional way of doing science by proposing a working hypothesis, to experimentally demonstrate the processes, to drawing conclusions proving or denying the original hypothesis. Genomic sciences have made drastic advances in the past ten years, largely because of the application of the next-generation sequencing technologies. It is not just the high throughput that has revolutionized the way science is conducted, but also the rapidly reducing cost for sequencing has made the technologies applicable to all aspects of molecular biological research as well as to all organisms, including aquaculture and fisheries species. From the billion dollar project of the human genome project, to the genome projects of agriculture animals with a budget of ten million dollars or so, down to a million dollars just a few years ago, to the current cost level of just tens of thousands of dollars for a de novo sequencing project, the potential for research using genomic approaches has become unlimited. Today, commercial services are available worldwide for projects, whether they are new sequencing projects for a species, or re-sequencing projects for many individuals. The key issue is to achieve a balanced output of quality and quantity with minimal costs.

Rapid technological advances provide huge opportunities to apply modern genomics to enhance aquaculture production and performance traits. It is now not too difficult to map and sequence a genome of an aquaculture species. However, understanding the genes controlling economically important traits takes tremendous levels of additional research. Once the important genes for performance and production traits are identified, they can be selected through marker-assisted selection or genomic selection. These genes can be modified to destroy a “bad” gene, edited to have the beneficial allele, or inserted in the genome at exact locations as specified by the scientists. Such powerful genome editing technologies opens unlimited potential for genetic enhancement of aquaculture species.

However, the world is facing a number of new challenges, especially in the area of bioinformatics. This challenge may be paramount for aquaculture researchers and educators. Aquaculture students may be well educated with aquaculture, but may have limited background in computer science, or not sophisticated enough for the bioinformatics analysis of large data sets. The large data sets in tera-scales themselves pose great computational challenges. Therefore, it requires new ways of thinking in terms of education and training of the next generation of scientists.
For instance, a few laboratories in the world may be sufficient for the production of data, but an infinite number of laboratories may be required for data analysis or bioinformatics data mining to link data with biology. In addition, aquaculture industries are less well organized, and working with small family-based farmers will pose tremendous levels of challenges when applying expensive and novel technologies. Consumer acceptance and public perception may become an issue of focus, and government and international regulations may become necessary for the broad application of genome-based technologies beyond research.
References


Baranski, M. et al. 2014. The development of a high density linkage map for black tiger shrimp (Penaeus monodon) based on cSNPs. PLOS ONE, 9: e85413.


Gonzalez-Pena, D. 2016. Genome-wide association study for identifying genome loci that affect fillet yield in rainbow trout (Oncorhynchus mykiss). Plant and Animal Genome XXIV Conference. Plant and Animal Genome. 8–13 January 2016.


Yue, G.H. 2014. Recent advances of genome mapping and markerFaisshsisted selection in *and Fisheries*, 15.3: 376–396.


