Development of the Catfish 250K SNP Array for Genome-Wide Association Studies

by

Shikai Liu

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
August 3, 2014

Keywords: catfish, genome, SNP array, next-generation sequencing, genome-wide association studies, quantitative trait loci

Approved by

Zhanjiang Liu, Chair, Professor of Fisheries and Allied Aquacultures
David B. Rouse, Professor of Fisheries and Allied Aquacultures
Nannan Liu, Professor of Entomology and Plant Pathology
Joanna Wysocka-Diller, Associate Professor of Biological Sciences
Qi Li, Professor of Fisheries College, Ocean University of China

Abstract

Catfish is the primary aquaculture species in the United States. In recent years, the catfish industry has encountered unprecedented challenges including increased feeding costs, devastating diseases and fierce international competition. Traditional selective breeding have been conducted for catfish breeds with improved traits of fast-growth, high feed efficiency and high level of disease resistance. Genomic selection that utilizes whole genome-based markers to assist selective breeding holds premise with increased selection intensity and accuracy. However, genome-scale genetic markers are required for its implementation, which has been a major limitation to most farming animals including aquaculture species. In recent years, next-generation sequencing technologies have enabled efficient and cost-effective identification of genome-scale genetic markers, such as single nucleotide polymorphisms (SNPs). With the availability of a large number of SNPs, the challenge then is how to genotype these SNPs efficiently and economically. One of the most efficient approaches is to design a high-density array that includes hundreds of thousands of SNPs covering the entire genome. Toward genomic selection in catfish, my research, as presented here, encompasses these two major progresses with the generation of genome-scale SNPs and the development of a high-density SNP array in catfish.

Using the RNA-Seq approach, over two million gene-associated SNPs were identified from channel catfish and blue catfish, two of the most important catfish species, providing a large pool of SNP resources for designing SNP array. Criteria-based filtering resulted in hundreds of thousands of quality SNPs that are intra-specific in channel catfish, intra-specific in blue catfish, and inter-specific between the two species. This is the first application of next-generation sequencing technology in catfish for genome-wide SNP identification. With the large number of SNPs, it’s important to select SNPs to represent each gene because SNPs within same genes are always linked. In addition, pseudo-SNPs can be detected due to misalignment of paralogous sequences from duplicated genes. To generate the reference gene transcript sequences for SNP selection and to detect potential pseudo-SNPs derived from duplicated genes, the catfish transcriptome assembly and annotation was conducted by RNA-Seq of a doubled haploid channel catfish, which harbors two identical sets of chromosomes and therefore there should be no variations. A comprehensive set of catfish transcript sequences was obtained including over 14,000 full-length transcripts. A set of genes putatively duplicated in catfish genome were identified, which aided the detection of pseudo-SNPs. With these resources, the catfish 250K SNP array was developed with the state-of-the-art Affymetrix Axiom technology with inclusion of over 250,000 SNPs. This is the first high density SNP array developed for catfish, which should be valuable for both the catfish industry and research such as in genomic selection, genome-wide association studies, fine linkage mapping and haplotype analysis.

Acknowledgments

I would like to express my sincere appreciation to my major professor, Dr. Zhanjiang Liu, for his guidance in all stages of this work. I am grateful for the time and expertise offered by my committee members, Dr. David B. Rouse, Dr. Nannan Liu, Dr. Joanna Wysocka-Diller, Dr. Qi Li, and my outside reader, Dr. Bernhard Kaltenboeck. Furthermore, I would like to extend my appreciation to all the colleagues in the Fish Molecular Genetics and Biotechnology Laboratory, especially Dr. Huseyin Kucuktus and Ludmilla Kaltenboeck for their assistance and support. Above all, I want to thank my beloved wife, Qing, and my family for their encouragement and support.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Overview

Catfish is the major aquaculture species in the United States. In recent years, the catfish industry has encountered unprecedented challenges due to devastating diseases, high feed and energy costs and fierce international competition, of which diseases cause the largest loss to the catfish industry. Enteric septicemia of catfish (ESC) disease, caused by the bacterium, *Edwardsiella ictaluri*, is one of the most serious infectious diseases in catfish. Vaccines and antibiotics have been developed, but limited for their future use due to the drawbacks including side-effects and lacking long-term efficacy. Decades of traditional breeding have been made to select catfish breeds resistant to diseases. In addition, whole genome-based marker-assisted selection (genomic selection) is a promising alternative for traditional marker-assisted selection genetic enhancement. Genomic selection allows increased genetic gain, increased accuracy of selection, and increased selection response for difficult-to-measure traits such as disease resistance (EGGEN 2012; GODDARD and HAYES 2009; MEUWISSEN *et al.* 2001). To implement genomic selection, genome-wide genetic markers are required, which, however, has not been available in most non-model species until recently.

Single nucleotide polymorphisms (SNPs) are one of the most popular genetic markers. They are among the most abundant type of genetic variations and widely distributed in any genome. Because of their potential for high genotyping efficiency, automation, data quality, genome-wide coverage and analytical simplicity (MORIN *et al.* 2004), SNPs have rapidly become the marker of choice for genetics and genomics studies. In particular, SNPs are most suitable for genome-wide association studies (GWAS) when working with performance and production traits because linkage disequilibrium (LD) can be detected with high density SNP coverage of the genome. To identify large numbers of SNPs, high throughput sequence reads are needed to increase sequencing coverage and depth and ensure sequence accuracy. Such work has been costly and time-consuming when employing traditional Sanger sequencing technology. In recent years, genome-wide identification of genetic markers has been enabled by next-generation sequencing technologies. Next generation sequencing technologies such as Roche/454, Illumina/Solexa, and ABI/SOLiD sequencing are particularly adapted to producing high coverage of sequences within contigs (MARDIS 2008a; MARDIS 2008b).

With the availability of a large number of SNPs, the challenge then is how to genotype these SNPs efficiently and economically. One of the most efficient approaches to genotype large numbers of SNPs is to design a high-density array that includes hundreds of thousands of SNPs distributed throughout the genome.

My research, as presented here, was conducted with the objectives to generate large-scale SNP resources and to develop the first high-density SNP array in catfish. Accomplishing this work should enable genome-wide association studies in catfish, and set the foundation towards genomic selection for genetic enhancement. Literature reviews of genomic selection, genome-scale SNP identification using next-generation sequencing, SNP array development and genome-wide association studies are given below. Following this are three chapters of publication-based presentation of my findings, and a chapter of overall conclusion serving to summarize the work and provide directions for future research in this field.

Whole genome-based selection

Domestic channel and blue catfish exhibit significant phenotypic and genetic variation for performance traits such as growth rate, feed conversion efficiency, processing yield, and disease resistance (ARGUE *et al.* 2003; DUNHAM 2007; DUNHAM and ARGUE 1998; DUNHAM and ARGUE 2000; DUNHAM *et al.* 1999; DUNHAM *et al.* 1990; DUNHAM *et al.* 1985; DUNHAM and SMITHERMAN 1983a; DUNHAM and SMITHERMAN 1983b; DUNHAM *et al.* 1987; DUNHAM *et al.* 1986; DUNHAM *et al.* 1983; HALLERMAN *et al.* 1986; REZK *et al.* 2003; WOLTERS and JOHNSON 1995). The catfish breeding programs are currently focused on mass selection, family selection and intraspecific crossbreeding in combination with inter-specific introgression by channel catfish x blue catfish and production of higher generations of backcrosses. The introgression and backcrossing strategy based on the inter-specific hybrids has been regarded as one of the best strategies for genetic enhancement of disease resistance and processing yield. The F_1 hybrid exhibits great heterosis, but mass production of fingerlings has not been possible.

Traditional breeding approaches have been effective in selecting for animals with easy-to-measure traits, but more difficult-to-measure (and usually more important) traits cannot effectively be selected. In the genomics era that routine discovery of millions of SNPs has been enabled by next generation sequencing technologies and ever-decreasing cost of genotyping genetic variation in a massively parallel fashion on high-density SNP array can be achieved, genomic selection holds great promise for more widespread adoption than traditional marker-assisted selection (EGGEN 2012; GODDARD and HAYES 2009). Genomic selection estimate breeding values based on information from a large number of markers covering entire genome without have a precise knowledge of where specific genes are located. Therefore, it does not require prior knowledge of alleles or marker positions of loci and does not need to be implemented within families. The principal advantages of genomic selection include that it can be implemented in very early in life, not sex limited, reducing the generation interval, and can be extended to any traits that are recorded in a reference population. It especially enhances the improvement of difficult-to-measure trait by providing better selection accuracy and intensity, increasing genetic gain and explaining a much greater portion of the genetic variance than traditional marker-assisted selection (Eggen 2012; Goddard and Hayes 2009; Meuwissen *et al.* 2001).

Genomic selection has been early implemented in cattle breeding. It has been shown that the genetic gain per year could be doubled in dairy cattle using genomic selection, with a potential to reduce costs for providing bulls by more than 90% (Schaeffer 2006). In cattle, genomic selections are being implemented in over 15 countries on the national level (http://www.interbull.org/). The evident benefits achieved in cattle can be transplanted to other species, and it's anticipated that the broader use of genomic selection can be one of the major advances in breeding programs for all animal species (Eggen 2012), including aquaculture species. As genomic resources continue to be cost-effectively developed for livestock, crop and aquaculture species, the selection decisions can be made sooner, traits that are difficult to improve with traditional breeding method can then be effectively enhanced. Strategies for genomic selection in aquaculture breeding programs have been tested blindly (Sonesson and Meuwissen 2009), given that large numbers of genome-scale SNP markers and high-density SNP arrays are not yet available in vast majority, if not all, aquaculture species. However, as the cost of sequencing and SNP genotyping continues to decrease, higher density SNP arrays with several hundred thousand SNPs are already being developed, promising the implementation of genomic selection in aquaculture species in the very near future.

Genome-wide SNP identification

SNPs have rapidly become the marker of choice for many applications in genetics and genomics due to their potential for high genotyping efficiency, automation, data quality, genome-wide coverage and analytical simplicity (DAVEY *et al.* 2011; FRAZER *et al.* 2009; MORIN *et al.* 2004). In species where the whole genome has been sequenced, SNPs have been identified from genome sequencing efforts. In most cases, SNPs were identified by sequence variations between the two alleles of a single diploid individual whose genome was sequenced (ALTSHULER *et al.* 2000; ECK *et al.* 2009; NIELSEN *et al.* 2011). Next generation sequencing (NGS) has proven to be effective for genome-wide SNP discovery (DAVEY *et al.* 2011; KERSTENS *et al.* 2009; RAMOS *et al.* 2009). When a high quality reference genome sequence is available, genomic sequences of individuals can be aligned more easily to this reference genome to detect nucleotide variations (DAVEY *et al.* 2011; LI *et al.* 2009; NIELSEN *et al.* 2011). Numerous studies have applied NGS platforms to achieve sufficient coverage of the genome for high quality SNP discovery in the complex genomes of plants and animals (AHMAD *et al.* 2011; KERSTENS *et al.* 2009; LE and DURBIN 2011; RAMOS *et al.* 2009; STOTHARD *et al.* 2011; YOU *et al.* 2011). NGS platforms such as Roche/454, Illumina/Solexa, and ABI/SOLiD are particularly adapted to producing high coverage of sequences within contigs (DAVEY *et al.* 2011; FRAZER *et al.* 2009; MARDIS 2008a; MARDIS 2008b; NIELSEN *et al.* 2011).   By comparison, the Illumina sequencing platform is the most competitive because of its higher throughput. For instance, a single lane of HiSeq 2000 run can generate up to 75 Gb data which can provide sufficient coverage and depth for SNP detection in the majority of organisms, at the cost of less than $4000 (personal communication). The utility of massively parallel Illumina sequencing has been demonstrated effective for whole genome resequencing for accurate SNP discovery on a genome-wide scale

(Eck *et al.* 2009; Hillier *et al.* 2008; Li *et al.* 2009; Ramos *et al.* 2009). In Hillier et al.'s work, about nine fold-coverage Illumina sequence reads from the *Caernohabditis elegans* strain CB4858 were generated and aligned to the reference, and screened for SNPs and small indels. Orthologous validation yielded a high validation rate of 96.3% (Hillier *et al.* 2008). Another example in a recent investigation on turkey genome variation, alignment of NGS data of 32 individuals from different populations discovered 5.49 million SNPs (Aslam *et al.* 2012).

Among the SNPs throughout whole genome, gene-associated SNPs are especially important. Gene-associated SNPs are relatively evenly distributed across genome and can themselves be causative SNPs for traits. The identification of gene-associated SNPs in non-model species has been fuelled by mining large numbers of expressed sequence tags (ESTs) available in many species (Guryev *et al.* 2004; Muchero *et al.* 2009; Schmid *et al.* 2003; Snelling *et al.* 2005). Likewise, gene-associated SNPs derived from ESTs have been identified in several fish species, including Atlantic salmon (Moen *et al.* 2008), Atlantic cod (Hubert *et al.* 2010) and catfish (Wang *et al.* 2008; Wang *et al.* 2010). In spite of being relatively effective, SNP identification from ESTs is limited by sequence coverage and depth. For instance, of the 303,000 putative SNPs identified from catfish ESTs, only 48,594 were identified from contigs containing at least four ESTs and at least two sequences bearing the minor allele. The majority of the catfish EST contigs (56% of 45,306) contain only two or three sequences (Wang *et al.* 2010). Putative SNPs identified from such contigs would have the minor alleles represented by only one sequence. Such SNPs could represent sequencing errors and therefore, are not reliable (Wang *et al.* 2008).

To identify large numbers of gene-associated SNPs, high throughput expressed sequence reads are needed to increase coverage and depth and ensure sequence accuracy. Transcriptome sequencing using NGS (RNA-Seq) with multiple individuals has been demonstrated to be very effective for identification of gene-associated SNPs (NOVAES *et al.* 2008). In this regard, 454 sequencing has been widely applied for the identification of gene-derived SNPs in a number of species including both plant and animal (HALE *et al.* 2009; HELYAR *et al.* 2012; HOU *et al.* 2011; NOVAES *et al.* 2008a; PARCHMAN *et al.* 2010; VERA *et al.* 2008). While the 454 sequencing technology has been widely used, Illumina sequencing technology is being gradually accepted for its dramatically improved sequencing throughput and quality (CIRULLI *et al.* 2010; SURGET-GROBA 2010 and MONTOYA-BURGOS 2010). Furthermore, paired-end sequencing technology along with the longer sequence reads make it possible to assemble contigs of transcripts from Illumina short reads. Illumina-based RNA-Seq is now being widely used for gene-associated SNP identification in numerous non-model species (AHMAD *et al.* 2011; CANOVAS *et al.* 2010; LIU *et al.* 2011a; TRICK *et al.* 2009). Transcriptome assemblies from RNA-Seq short reads can be aided by the presence of reference genome and/or reference transcriptome sequences (PARCHMAN *et al.* 2010; TRICK *et al.* 2009). In this context, a large number of ESTs of catfish are available (WANG *et al.* 2010).

Development of SNP genotyping array

High throughput NGS platforms enable the genome-wide identification of SNPs efficiently and economically, but the subsequent challenge is how to utilize these SNP resources for genetic analysis. One of the most efficient approaches to genotype large numbers of SNPs is to design a high-density array that includes hundreds of thousands of SNPs distributed throughout the genome. A number of platforms are available for genotyping SNPs, including MassArray (Sequenom), SNPstream (Beckman Coulter), SnaPshot Multiplex System (Applied Biosystems) and iSelect HD Custom Beadchip (Illumina) which are based on the single-base extension for SNP calling; GoldenGate Assays (Illumina) which use allele-specific primer extension for SNP calling; TaqMan OpenArray (Applid Biosystems) and Dynamic Array (Fluidigm) that are designed from the TaqMan technology; MyGeneChip Custom Array (Affymetrix) that uses the chemistry of differential hybridization as the bases for SNP detection. Among them, the Sequenom MassArray, Illumina GoldenGate Assay and iSelect HD Custom BeadChip, and Affymetrix MyGeneChip Custom Array are widely used. More recently, Affymetrix adopted the Axiom myDesign Array Genotyping Services, which enable to create fully- or semi-customized genotyping arrays containing 1,500 to 2.6 million markers and have them processed with high-quality and fast turnaround time. Each of these platforms has significantly different requirements for SNP marker number, sample size, cost and automation. The appropriate platform for SNP genotyping needs to be chosen depends on specific project goals and budget levels.

SNP studies in human have usually been divided between genome-wide association studies and fine mapping. GWAS usually utilize Affymetrix MyGeneChip Array or Illumina iSelect HD Beadchip platforms to genotype several thousand samples for over half a million SNPs. Fine mapping studies follow up GWAS results to confirm and/or refine findings by scanning a SNP subset with a larger sample size (FLINT and ESKIN 2012; KRUGLYAK 2008; MCCARTHY *et al.* 2008). The majority of SNP genotyping platforms were initially developed for GWAS and fine mapping in human, and subsequently adopted into other model species including several agriculturally important livestock species. The human Affymetrix arrays for massively parallel genotyping of SNPs started with the Mapping 10K array (KENNEDY *et al.* 2003). After that, the Mapping 100K array set was developed as the first SNP array suitable for association studies and produced the first GWAS finding (KLEIN *et al.* 2005; MATSUZAKI *et al.* 2004). The Mapping 500K array was the first array set with sufficient density to enable highly powered GWAS, then the Affymetrix SNP 5.0 array was developed with essentially the same SNP content as the Mapping 500K, but the design involves multiple replicates of the most informative probes rather than a single copy of many different probe sequences (MCCARROLL *et al.* 2008). The Affymetrix SNP 6.0 array was designed based on screening over two million SNPs and selection of 906K SNPs to optimize the whole genome coverage (www.affymetrix.com). Recently, an Affymetrix Axiom-based SNP array was developed, optimizing to genotype individuals with European ancestry. The array contains 674,517 SNPs, and provides excellent genome-wide and gene-based SNP coverage (HOFFMANN *et al.* 2011). In parallel, Illumina launched the first Infinium product, the Human-1 Genotyping BeadChip, includes over 100,000 markers on a single BeadChip in middle of 2005. The HumanHap300 Genotyping BeadChip, which leverages tag SNP content to deliver over 300,000 markers on a single microarray, was developed a year later. Illumina launched the next BeadChip in its Infinium whole-genome genotyping, the HumanHap550, with over 550,000 SNP markers. The third BeadChip in its Infinium whole-genome genotyping offering, Infinium HumanHap650Y BeadChip, was designed with over 650,000 SNPs. The Human1M and Human450S BeadChips were developed to combining an unprecedented level of content for both whole-genome and copy number variation analysis (http://www.illumina.com/).

Several high throughput SNP genotyping arrays have been developed in livestock species including cattle, pig, horse, sheep, dog and chicken. The BovineSNP50 Beadchip powered by Illumina Infinium HD assay was first developed for detecting variation in cattle breeds (MATUKUMALLI *et al.* 2009). The BovineSNP50 features 54K informative SNP probes that uniformly span the entire genome. The evaluation on a panel of 576 animals from 21 cattle breeds and six outgroup species revealed 74%-86% polymorphic SNPs within individual breeds (MATUKUMALLI *et al.* 2009). Two higher-density genotyping arrays have recently become available to the bovine genomics community (RINCON *et al.* 2011). The High-Density Bovine BeadChip (BovineHD) array (777K) using Illumina Infinium HD assay, and the Axiom Genome-Wide BOS 1 Array (BOS 1) (648K) using Affymetrix Axiom technology. The performance to these two bovine high-density genotyping platforms was tested in Holsteins and Jerseys. The results indicated that both Affymetrix BOS 1 and Illumina BovineHD genotyping platforms were well designed and provided high-quality genotypes and similar genome coverage (RINCON *et al.* 2011). A high-density SNP genotyping assay (PorcineSNP60) in the pig was designed using SNPs identified by next generation sequencing technology (RAMOS *et al.* 2009). The results of this study indicated the utility of NGS to identify large numbers of reliable SNPs for array design. An equine SNP genotyping array (EquineSNP50 BeadChip) was developed and evaluated on a panel of samples representing 14 domestic horse breeds and 18 evolutionarily related species (MCCUE *et al.* 2012). The utility of the SNP array in GWAS was confirmed by mapping the known recessive chestnut coat color locus (MC1R) and defining a conserved haplotype of ~750kb across all breeds. Additionally, the Illumina OvineSNP50 BeadChip that provided 42,469 SNP markers was developed by the International Sheep Genomics Consortium (ISGC). In dog, Illumina developed the CanineSNP20 BeadChip and then a recent upgraded version of CanineHD BeadChip. The Illumina CanineHD BeadChip was developed to contain over 170,000 evenly spaced and validated SNPs enabling the interrogation of genetic variation in any domestic dog breed (MOGENSEN *et al.* 2011). A 60K chicken SNP array powered by Illumina iSelect BeadChip was designed to consist of 57,636 SNPs (GROENEN *et al.* 2011). The evaluation indicated that 94% of SNPs could be genotyped and shown to be segregating in chicken populations. More recently, a 600K chicken SNP array with Affymetrix Axiom technology was developed (KRANIS *et al.* 2013).

Although medium- to high-throughput SNP genotyping is still in its early stage, several projects have been conducted in aquaculture species. The Illumina GoldenGate Assay was used to evaluate 384 rainbow trout SNPs of a large set of SNPs discovered by reduced representation shotgun sequencing. A total of 192 samples were genotyped, resulting in a validation rate of 48% for the tested SNPs (SANCHEZ *et al.* 2009). The GoldenGate Assay was also used to genotype 192 catfish for 384 EST-derived SNPs in catfish (WANG *et al.* 2008). The research provided assessments of factors affecting SNP validation rates including contig size, minor allele frequency and flanking sequence quality. The Sequenom MassArray platform was used in several studies in the context of QTL analysis, linkage mapping and map integration. Boulding et al. (2008) examined associations between SNPs and QTLs for adaptive traits in juvenile Atlantic salmon. Genotyping of 980 fish for 129-320 SNPs using the MassArray platform enabled the detection of 79 significant associations between SNPs and QTLs for the adaptive traits (BOULDING *et al.* 2008). An Atlantic cod genetic linkage map was constructed by genotyping 1,146 offspring from 12 full-sib families for 257 SNPs on the MassArray system (MOEN *et al.* 2009). A total of 174 SNPs were successfully placed on the linage map. Using the MassArray system, a gene-associated SNP linkage map was also created for Pacific white shrimp by genotyping 144 individuals for 825 SNPs (DU *et al.* 2010). In this case, a total of 418 SNPs were incorporated into the linkage map. The MassArray system was also used for the Atlantic salmon physical map-linkage map integration (LORENZ *et al.* 2010). They resequenced bacterial artificial chromosome (BAC) end sequences from 14 individuals to discover a total of 180 SNPs. They then genotyped 376 fish for these SNPs and positioned 110 SNPs on the existing linkage map, then successfully anchored 73 BAC contigs to the Atlantic salmon linkage map.

Recently, a custom Illumina iSelect SNP-array containing approximately 6K SNP markers from Atlantic salmon has been developed and used to analyze genetic differences between farmed and wild Atlantic salmon (KARLSSON *et al.* 2011). This SNP array was also used to construct a relatively dense SNP linkage map (LIEN *et al.* 2011) and identify QTL related to body weight of this species (GUTIERREZ *et al.* 2012). Linkage mapping based on genotyping of 3,297 Atlantic salmon from 143 families resulted in a linkage map with a total of 5,650 SNPs (LIEN *et al.* 2011). Based on this linkage map, QTL analysis by genotyping 5 Atlantic salmon families identified 5 significant QTL associated with body-weight traits (GUTIERREZ *et al.* 2012).

Genome-wide association studies

Along with the development of high throughput SNP genotyping technologies, GWAS has been extensively conducted in human disease research. For instance, in 2005, an association was found between age-related macular degeneration (ARMD) and a variation in the gene for complement factor H (CFH). Use of variation in the CFH gene, along with four other variants, can predict half the risk of ARMD between siblings, and this work was regarded as among the most successful examples of GWAS (KLEIN *et al.* 2005). Similarly, a GWAS involving genotyping of around 400K SNPs in a French case–control cohort allowed detection of an association between type 2 diabetes and a variation in several SNPs in the genes TCF7L2, SLC30A8 and others (SLADEK *et al.* 2007), which can explain a substantial portion of disease risk. In 2007, the Wellcome Trust Case Control Consortium carried out GWA studies for 14,000 cases of seven common diseases including coronary heart disease, type 1 diabetes, type 2 diabetes, rheumatoid arthritis, Crohn's disease, bipolar disorder, and hypertension. This study was successful in uncovering many new genes underlying these diseases (BURTON *et al.* 2007). In recent years, numerous GWA studies utilizing high-density SNP arrays have been conducted for human diseases (ANTTILA *et al.* 2010; BARRETT *et al.* 2009; BURTON *et al.* 2007; CAI *et al.* 2011; CHAMBERS *et al.* 2009; CHIO *et al.* 2009; EELES *et al.* 2009; FLETCHER *et al.* 2011; GUDMUNDSSON *et al.* 2009; HAN *et al.* 2009; HAROLD *et al.* 2009; HU *et al.* 2011; HUANG *et al.* 2010; HUANG *et al.* 2012; KHOR *et al.* 2011; KIM *et al.* 2011; LIN *et al.* 2012a; LIN *et al.* 2012b; LU *et al.* 2012; MANOLIO 2010; MELUM *et al.* 2011; NEWTON-CHEH *et al.* 2009; PAINTER *et al.* 2011; QI *et al.* 2010; RAPLEY *et al.* 2009; SHI *et al.* 2011; SIMON-SANCHEZ *et al.* 2009; SLADEK *et al.* 2007; SUN *et al.* 2011; VITHANA *et al.* 2012; WANG *et al.* 2011; WU *et al.* 2011; WU *et al.* 2012a; WU *et al.* 2012b; XU *et al.* 2012; ZHANG *et al.* 2009) see reviews (HARDY and SINGLETON 2009; HIRSCHHORN and DALY 2005; KRUGLYAK 2008; MANOLIO 2010; MCCARTHY *et al.* 2008) and for several agriculturally important traits in crops (FAMOSO *et al.* 2011; HUANG *et al.* 2010; HUANG *et al.* 2012).

In recent years, along with the development of high-density SNP arrays, GWAS were extended to the field of domestic animals including cattle, pig, horse, sheep, dog, and chicken. A recent review paper provides a comprehensive summery of GWA studies conducted in domestic animals (ZHANG *et al.* 2012). In cattle, tens of GWA studies have been conducted for several economically important traits, including mile yield and quality, fertility, growth, meat quality, carcass yield and bovine diseases (BOLORMAA *et al.* 2011; FINLAY *et al.* 2012; JIANG *et al.* 2010; MAI *et al.* 2010; MEREDITH *et al.* 2012; PANT *et al.* 2010; SAHANA *et al.* 2010; SAHANA *et al.* 2011; SCHOPEN *et al.* 2011; SETTLES *et al.* 2009; UEMOTO *et al.* 2010; VAN HULZEN *et al.* 2012). The majority of these studies employed the bovineSNP50 for SNP genotyping, while one recent study using llumina BovineHD (777K) for Dominant White Phenotype and Bilateral Deafness has been conducted (PHILIPP *et al.* 2011). The results of this study revealed a most significant associated-region on bovine chromosome 22 and a missense mutation in exon 7 of MITF gene was responsible for the trait. The first example of GWAS in pig was reported by DUIJVESTEIJN *et al.* (2010). They used the PorcineSNP60 BeadChip to genotype 987 pigs divergent for androstenone concentration from a commercial Duroc-based sire line. The results of this study revealed that 37 SNPs on chromosomes 1 and 6 significantly affect the androstenone levels in fat tissue (DUIJVESTEIJN *et al.* 2010). Soon after, GWAS were conducted for several other economically important traits including skatole levels, boar taint, fertility, backfat, body confirmation, fatness and coat color (GRINDFLEK *et al.* 2011; ONTERU *et al.* 2012; ONTERU *et al.* 2011; PONSUKSILI *et al.* 2011; RAMOS *et al.* 2011). The Illumina EquineSNP50 BeadChip genotyping array enabled several GWAS on traits of racing distance and some disorder and diseases in horse (BROOKS *et al.* 2010; DUPUIS *et al.* 2011; HILL *et al.* 2010; ORR *et al.* 2010). The first report of GWAS in sheep was conducted on horn types using Illumina OvineSNP50 BeadChip (JOHNSTON *et al.* 2011). This study determined an autosomal gene, RXFP2, as the main genetic candidate for horn-type, accounting for up to 76% of the additive genetic variation in this trait (JOHNSTON *et al.* 2011). Another study of GWAS in sheep using the same SNP array revealed that a nonsense mutation on exon 6 of dentin matrix protein 1 (DMP1) was responsible for inherited rickets in Corriedale sheep (ZHAO *et al.* 2011). The special roles of dog related to human made it popular for genomics and genetics studies. A number of GWA studies on dog diseases have been conducted including Degenerative myelopathy (DM), Canine atopic dermatitis (cAD), and Arrhythmogenic right ventricular cardiomyopathy (ARVC) (AWANO *et al.* 2009; HALE *et al.* 2009; MEURS *et al.* 2010). A recent GWAS analysis using Illumina CanineHD BeadChip was carried out to identify genetic variants associated with intervertebral disc calcification in Dachshunds (MOGENSEN *et al.* 2011). The first GWAS in chicken was conducted on fatness using 3K SNPs in two F2 populations (ABASHT and LAMONT 2007). After the development of chicken 60K Illumina iSelect SNP array, several recent GWA studies were published for various traits including body weight, growth, egg production and quality, and diseases in chicken (GU *et al.* 2011; LI *et al.* 2012; WELLS *et al.* 2012; XIE *et al.* 2012).

In spite of being very powerful, GWAS has not been applied to aquaculture species. The primary reason was the lack of genome-wide polymorphic markers until recently. Now that a large number of SNPs are available for a number of aquaculture species, future applications in GWAS for aquaculture species is clearly technically feasible. It's anticipated that GWA studies for production traits in aquaculture species will soon be available given the ever-decreasing cost for sequencing and high-throughput SNP genotyping.

Chapter 2

Generation of genome-scale gene-associated SNPs in catfish for the construction of a high-density SNP array

Abstract

Single nucleotide polymorphisms (SNPs) have become the marker of choice for genome-wide association studies. In order to provide the best genome coverage for the analysis of performance and production traits, a large number of relatively evenly distributed SNPs are needed. Gene-associated SNPs may fulfill these requirements of large numbers and genome wide distribution. In addition, gene-associated SNPs could themselves be causative SNPs for traits. The objective of this project was to identify large numbers of gene-associated SNPs using high-throughput next generation sequencing. Transcriptome sequencing was conducted for channel catfish and blue catfish using Illumina next generation sequencing technology. Approximately 220 million reads (15.6 Gb) for channel catfish and 280 million reads (19.6 Gb) for blue catfish were obtained by sequencing gene transcripts derived from various tissues of multiple individuals from a diverse genetic background. A total of over 35 billion base pairs of expressed short read sequences were generated. Over two million putative SNPs were identified from channel catfish and almost 2.5 million putative SNPs were identified from blue catfish. Of these putative SNPs, a set of filtered SNPs were identified including 342,104 intra-specific SNPs for channel catfish, 366,269 intra-specific SNPs for blue catfish, and 420,727 inter-specific SNPs between channel catfish and blue catfish. These filtered SNPs are

distributed within 16,562 unique genes in channel catfish and 17,423 unique genes in blue catfish. For aquaculture species, transcriptome analysis of pooled RNA samples from multiple individuals using Illumina sequencing technology is both technically efficient and cost-effective for generating expressed sequences. Such an approach is most effective when coupled to existing EST resources generated using traditional sequencing approaches because the reference ESTs facilitate effective assembly of the expressed short reads. When multiple individuals with different genetic backgrounds are used, RNA-Seq is very effective for the identification of SNPs. The SNPs identified in this report will provide a much needed resource for genetic studies in catfish and will contribute to the development of a high-density SNP array. Validation and testing of these SNPs using SNP arrays will form the material basis for genome association studies and whole genome-based selection in catfish.

Introduction

Single nucleotide polymorphisms (SNPs) are alternative bases at any given position of DNA. They are among the most abundant type of genetic variations and widely distributed within genomes. Theoretically, SNPs can have four alleles in the population, but they most often exist as bi-allelic markers. Because of their potential for high genotyping efficiency, automation, data quality, genome-wide coverage and analytical simplicity (MORIN *et al.* 2004), SNPs have rapidly become the marker of choice for many applications in genetics and genomics. In particular, SNPs are most suitable for whole genome association studies because linkage disequilibrium can be detected with high density SNP coverage of the genome when working with performance and production traits. For instance, simultaneous analysis of thousands of SNPs have enabled genome-wide association studies for complex traits in chicken (ABASHT and LAMONT 2007), pig (DU *et al.* 2009; DUIJVESTEIJN *et al.* 2010) cattle (KHATKAR *et al.* 2008; KIM *et al.* 2009; MEUWISSEN *et al.* 2001) horse (BROOKS *et al.* 2010) and sheep (BECKER *et al.* 2010; KIJAS *et al.* 2009). However, such studies have not been possible with most aquaculture species including catfish because large numbers of SNPs have not been available.

In species where the whole genome has been sequenced, SNPs have been identified from genome sequencing efforts. In most cases, SNPs were identified by sequence variations between the two alleles of a single diploid individual whose genome was sequenced (ECK *et al.* 2009). More recently, the identification of SNPs in non-model species has been fuelled by mining large numbers of expressed sequence tags (ESTs) available in many species. Likewise, gene-associated SNPs derived from ESTs have been identified in several fish species, including Atlantic salmon (MOEN *et al.* 2008), Atlantic cod (HUBERT *et al.*) and catfish (HE *et al.* 2003; WANG *et al.* 2010; WANG *et al.* 2008). In spite of being relatively effective, SNP identification from ESTs is limited by sequence coverage and depth. For instance, of the 303,000 putative SNPs identified from catfish ESTs, only 48,594 were identified from contigs containing at least four ESTs and at least two sequences bearing the minor allele. The majority of the catfish EST contigs (56% of 45,306) contain only two or three sequences (WANG *et al.* 2010). Putative SNPs identified from such contigs would have the minor alleles represented by only one sequence. Such SNPs could represent sequence errors and therefore, are not reliable (WANG *et al.* 2008).

To identify larger numbers of gene-associated SNPs, higher throughput expressed sequence reads are needed to increase coverage and depth and ensure sequence accuracy. Next generation sequencing technologies such as Roche/454, Illumina/Solexa, and ABI/SOLiD sequencing platforms are particularly adapted to producing high coverage of expressed sequences within contigs (MARDIS 2008a; MARDIS 2008b). Transcriptome analysis using next generation sequencing with multiple individuals has been demonstrated to be very effective for SNP identification (NOVAES *et al.* 2008). Recently, 454 sequencing was applied for the identification of gene-derived SNPs in a number of species such as eucalyptus grandis (NOVAES *et al.* 2008), pine tree (PARCHMAN *et al.*2010), butterfly (VERA *et al.* 2008), lake sturgeon (HALE *et al.* 2009) and coral (MEYER *et al.* 2009).

While the 454 sequencing technology has been widely used for transcriptome analysis, Illumina sequencing technology is being gradually accepted for its dramatically improved sequencing throughput and quality (CIRULLI *et al.* 2010; SURGET-GROBA 2010 and MONTOYA-BURGOS 2010). Paired-end sequencing technology along with the longer sequence reads make it possible to assemble contigs of transcripts from Illumina short reads. Such assemblies are aided by the presence of reference genome and/or reference transcriptome sequences (PARCHMAN *et al.* 2010 ; TRICK *et al.* 2009). In this context, a large number of ESTs of catfish are available. The objective of this study is to conduct transcriptome sequencing from multiple individuals of both channel catfish (*Ictalurus punctatus*) and blue catfish (*I. furcatus*) in order to identify gene-associated SNPs for the development of SNP arrays in catfish.

Materials and Methods

Sample and RNA isolation

Channel catfish of 47 individuals from five different aquaculture populations/fingerling sources (8 Marion Select, 10 Pearson, 11 Moyer, 10 Holland, 8 Noble) and blue catfish of 19 individuals from two different strains (7 Rio Grande and 12 D&B) were used for this study. Samples of 11 tissues including brain, gill, head kidney, intestine, liver, muscle, skin, spleen, stomach, heart, and trunk kidney were collected. The fish were euthanized with tricaine methanesulfonate (MS 222) at 300 mg/l before tissue collection. Tissue samples from each species were collected, pooled, immediately placed in 5 ml RNA later™ (Ambion, Austin, TX, USA) and kept at 4 ℃ for 2-4 days until RNA extraction. Equal weight of each tissue from individuals of each species were combined, ground to a fine powder with mortar and pestle in the presence of liquid nitrogen and thoroughly mixed. A fraction of the tissue samples was used for RNA isolation. Total RNA was isolated using the RNeasy plus Mini Kit (Qiagen, Valencia, CA, USA) with DNase I (Invitrogen, USA) treatment following the manufacturer's protocol.

Illumina sequencing

Sequencing was conducted commercially in HudsonAlpha Genomic Services Lab (Huntsville, AL, USA). Briefly, 100 ng of total RNA was used to prepare amplified cDNA using Ovation RNA-seq, a commercially available kit optimized for RNA sequencing (NuGEN Technologies, San Carlos, CA). The produced double-stranded cDNA was subsequently used as the input to the Illumina library preparation protocol starting with the standard end-repair step. The end-repaired DNA with a single 'A'-base overhang is ligated to the adaptors in a standard ligation reaction using T4 DNA ligase and 2 μM-4 μM final adaptor concentration, depending on the DNA yield following purification after the addition of the 'A'-base. Following ligation, the samples were purified and subjected to size selection via gel electrophoresis to isolate 350 bp fragments for ligation-mediated PCR (LM-PCR). Twelve cycles of LM-PCR were used to amplify the ligated material in preparation for cluster generation. For each species of channel catfish and blue catfish, the prepared cDNA library was sequenced with 36-bp paired-end reads on one flow cell lane of the Illumina Genome Analyzer II platform and 100-bp paired-end reads on one flow cell lane of the Hiseq 2000 platform, respectively. The image analysis, base calling and quality score calibration were processed using the Illumina Pipeline Software v1.4.1

according to the manufacturer's instructions. Reads were exported in the FASTQ format and has been deposited at the NCBI Sequence Read Archive (SRA) under accession number SRA025099.

Assembly of expressed short reads

Sequence analysis was performed using the high-throughput sequencing module of CLC Genomics Workbench (version 4.0.2; CLC bio, Aarhus, Denmark). The raw reads were cleaned by trimming of adaptor sequences, ambiguous nucleotides ('N' in the end of reads) and low quality sequences with average quality scores less than 20. Trimmed reads less than 15 bp were also discarded from further analysis, the remaining reads were used in subsequent assembly. The approach of assembly in this study was based on a combination of reference assembly and *de novo* assembly. A reference-based assembly was firstly executed using a set of catfish unique sequences generated from ~500,000 Sanger-ESTs of both channel and blue catfish as a reference. For the reference assembly, the default local alignment settings were used to rank all potential matches, with mismatch cost of 2, deletion cost of 3 and insertion cost of 3. The highest scoring matches that shared ≥ 80% similarity with the reference sequence across ≥ 50% of their length were included in the alignment. This permissive alignment ensured that even reads derived from highly mutated orthologs between channel catfish and blue catfish would not be discarded. Reads that were not assembled into contigs in the reference assembly were entered into a subsequent *de novo* assembly with a higher stringency minimum match similarity (90%). Three separate assemblies were generated: channel catfish assembly, blue catfish assembly, and all catfish assembly (Figure 1).

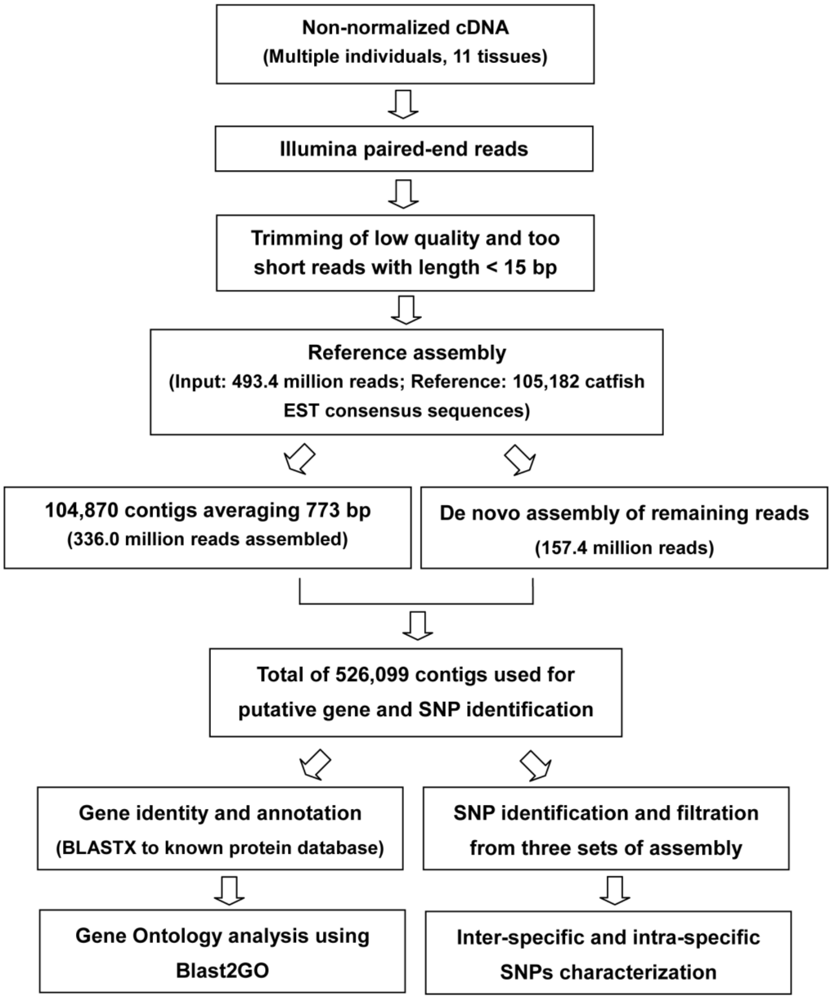
Non-normalized cDNA
(Multiple individuals, 11 tissues)
Illumina paired-end reads
Trimming of low quality and too short reads with length < 15 bp
Reference assembly
(Input: 493.4 million reads; Reference: 105,182 catfish EST consensus sequences)
104,870 contigs averaging 773 bp
(336.0 million reads assembled)
De novo assembly of remaining reads
(157.4 million reads)
Total of 526,099 contigs used for putative gene and SNP identification
Gene identity and annotation
(BLASTX to known protein database)
SNP identification and filtration from three sets of assembly
Gene Ontology analysis using Blast2GO
Inter-specific and intra-specific SNPs characterization

Figure 1. Schematic presentation of the catfish transcriptome analysis.

Gene identification and annotation

Unique consensus sequences from the all catfish assembly were compared against the Uniprot database and the zebrafish Refseq protein database (NCBI) using BLASTX (cutoff E-value of 1E-10) to obtain the putative gene identity. To estimate the proportion of annotated contigs that matched to unique genes in the known protein database, all BLASTX hits were filtered for redundancy in protein accessions. Assignment of Gene Ontology terms to annotated unique sequences was conducted using the program Blast2GO (CONESA *et al.* 2005). Ontology was categorized with respect to Biological Process, Molecular Function, and Cellular Component.

SNP and microsatellite markers identification

Assembled contigs were scanned for SNPs utilizing SNP detection software included in CLC Genomics Workbench (CLC bio, Aarhus, Denmark). The central base quality score of ≥ 25 and average surrounding base quality score of ≥ 20 were set to assess the quality of reads at positions for SNP detection. Under the criteria of minimum coverage (read depth) of four and the minimum variant frequency of two, the variations compared to the reference sequence were counted as SNPs. Three lists of SNPs were generated from channel catfish, blue catfish and all catfish assembly, respectively. The identification of intra-specific SNPs for both channel and blue catfish, and inter-specific SNP between channel and blue catfish was achieved by comparing these three lists of SNPs. Inter-specific SNPs were defined as those that have sequence variations between channel catfish and blue catfish, but no sequence variations within channel catfish or within blue catfish; similarly, intra-specific SNPs were identified within channel catfish or within blue catfish; and intra-specific SNPs for both channel catfish and blue catfish were identified within both channel catfish and blue catfish at the same SNP position.

All the unique sequences were used to search for microsatellite makers using Msatfinder (THURSTON MI) with a repeat threshold of eight di-nucleotide repeats or five tri-, tetra-, penta-, or hexa- nucleotide repeats. The presence of at least 50-bp sequence on both sides of the microsatellite repeats were considered sufficient for primer design (ROZEN and SKALETSKY 2000; SOMRIDHIVEJ *et al.* 2008).

Quality SNP screening

In order to identify quality SNPs, putative SNPs identified as mentioned above were further screened following specific criteria based on the read depth, minor allele frequency, the quality of flanking regions and absence of other SNPs within 15-bp flanking regions: only those SNPs with minor allele sequences representing no less than 10% of the reads aligned at the polymorphic loci were declared as quality SNPs; no extra SNPs or indels within 15-bp flanking regions were allowed; SNPs located in repetitive regions were also not considered. Potential repetitive elements were detected by RepeatMasker (http://www.repeatmasker.org/). SNPs located in repetitive regions were checked and ruled out using custom scripts. For practical application in SNP genotyping assays, only bi-allelic SNPs were considered in this study. To get a snapshot of the SNP distribution across the catfish genome, SNP- containing contigs with BLAST hits to the Ensembl zebrafish transcripts database were plotted along the zebrafish chromosomes.

Results

Generation of expressed short reads

Illumina sequencing was conducted to generate short sequence reads of expressed sequences. Two cDNA libraries were made from pooled RNA samples prepared from a total of 11 tissues of 47 channel catfish and 19 blue catfish, respectively, representing major strains used in commercial production. The cDNAs were sequenced with one lane each using Illumina GA-II and Illumina HiSeq 2000 that generated 48.6 million 36-bp paired-end reads and 173.9 million 100-bp paired-end reads for channel catfish, and 66.9 million 36-bp paired-end reads and 216.6 million 100-bp paired-end reads for blue catfish (Table 1). After removal of ambiguous nucleotides, low-quality sequences (quality scores < 20) and sequences less than 15 bp, sequences totaling 15.6 billion base pairs for channel catfish and 19.6 billion base pairs for blue catfish were generated (Table 1).

Table 1. Generation of Illumina expressed short reads. Eleven tissues were used for RNA preparation including brain, gill, head kidney, intestine, liver, muscle, skin, spleen, stomach, heart, and trunk kidney. *Paired-end reads were generated in different lengths of either 36 bp or 100 bp as a result of different sequencers, Illumina GA-II or HiSeq 2000.

Species	No. of fish	Sequencer	Sequence length*	Reads ($X10^6$)	Bases sequenced ($X10^9$)	Reads after trimming ($X10^6$)	Bases after trimming ($X10^9$)
Channel catfish	47	Illumina GA-II	36 bp	48.6	1.8	47.2	1.7
		HiSeq 2000	100 bp	173.9	17.4	171.6	13.9
Blue catfish	19	Illumina GA-II	36 bp	66.9	2.3	62.1	2.2
		HiSeq 2000	100 bp	216.6	21.7	212.5	17.4
Total	-	-	-	506.0	43.2	493.4	35.2

Assembly of the expressed short reads

Assembly of the expressed short reads was conducted in several ways. First, reference assemblies of channel catfish expressed short reads and blue catfish expressed short reads were conducted separately using all existing catfish ESTs as a reference. Such assemblies would allow establishment of contigs for channel catfish expressed short reads and blue catfish expressed short reads separately to allow identification of intra-specific SNPs that are anchored (scaffold) by longer EST reference sequences. Such an assembly is superior to the total *de novo* assembly of the expressed short reads which generates very large numbers of contigs, over 800,000 (data not shown). As shown in Table 2, over two thirds of the expressed short reads were assembled with the reference assemblies. Over 152 million reads of channel catfish (69.8%) and 183 million reads of blue catfish (66.7%) were assembled into 103,650 and 104,475 contigs, respectively. The contigs were reasonably long with an average contig length of 670 bp and 775 bp, respectively, for channel catfish and blue catfish (Table 2).

Table 2. Reference assembly of expressed short reads. *Number of reads per contig. #Total number of assembled read bases/Total number of bases in consensus sequence

Species	No. of reads used for assembly	No. of reads assembled	% sequences assembled	No. of contigs	Average contig length (bp)	Average contig size*	Average coverage #
Channel catfish	218.8 x 10^6	152.6 x 10^6	69.8%	103,650	670	1,473	137.4
Blue catfish	274.6 x 10^6	183.8 x 10^6	66.7%	104,475	775	1,760	164.2

Despite generating an efficient reference assembly, over 66 million channel catfish reads and 90 million blue catfish reads were not assembled with the reference assembly. These reads could represent additional genes that were not represented by the EST reference sequences, or they could come from gene regions that were not represented by the EST references. In order to make them useful resources for SNP identification, *de novo* assembly of these remaining reads was conducted. As shown in Table 3, over 70% of these unassembled expressed short reads could be assembled *de novo*, generating 420,165 contigs and 420,953 contigs for channel catfish and blue catfish, respectively. However, the average contig length was much shorter than those in the reference assembly, with 298 bp and 315 bp for channel catfish and blue catfish, respectively. These contigs are also useful resources for the identification of intra-specific SNPs.

Table 3. *De novo* assembly of the unassembled expressed short reads. All the newly generated expressed short reads were first assembled using reference assembly (Table 2), and those that were not assembled, i.e., they did not align *in silico* to the existing catfish ESTs, were used for the *de novo* assembly. *Number of reads per contig. #Total number of assembled read bases/Total number of bases in consensus sequence.

Species	No. of reads used for assembly	No. of reads assembled	% sequences assembled	No. of contigs	Average contig length (bp)	Average contig size*	Average coverage #
Channel catfish	66.2 x 10^6	46.8 x 10^6	70.7%	420,165	298	111	19.7
Blue catfish	90.8 x 10^6	64.3 x 10^6	70.8%	420,953	315	153	26.4

After separate analyses of channel catfish and blue catfish expressed short reads, reference and *de novo* assemblies were conducted using combined channel catfish and blue catfish expressed short reads in order to identify inter-specific SNPs. A total of 493.4 million expressed short reads from both channel catfish and blue catfish (all catfish) were used. The reference assembly of all catfish expressed short reads placed 336.0 million reads (68%) into 104,870 contigs, with an average contig length of 686 bp. Similarly, the *de novo* assembly of all catfish expressed short reads generated 421,229 contigs, with an average contig length of 340 bp (Table 4). These contigs should be useful for the identification of inter-specific SNPs.

Table 4. Summary of assembly using all catfish expressed short reads. [1]All expressed short reads from both channel catfish and blue catfish were first assembled using existing ESTs as references. [2]Those that were not assembled into contigs with the reference ESTs were then assembled *de novo*. [#]Total number of assembled read bases/Total number of bases in consensus sequence.

Assembly	No. of reads used	No. of reads assembled	% sequences assembled	No. of contigs	Avg. contig length	Max length	No. of contigs (>1kb)	Avg. coverage[#]
Reference[1]	493.4 x 10^6	336.0 x 10^6	68.1%	104,870	686	6,849	17,756	330.8
De novo[2]	157.4 x 10^6	107.2 x 10^6	68.2%	421,229	340	4,615	4,133	44.1

Putative gene identity and annotation

Before SNP identification, we conducted analysis of putative gene identities to help assess how many genes may be included in the assemblies. To determine the putative gene identities, unique consensus sequences from the all catfish reference assembly and *de novo* assembly were searched against the Uniprot database and NCBI zebrafish Refseq protein database using BLASTX with a cutoff E-value of 1E-10. Of 104,870 all catfish contigs from the reference assembly, 32,350 (30.9%) had BLAST hits to the Uniprot database, and matched 17,766 unique protein accessions. As expected, a lower percentage of the contigs from *de novo* assembly had

BLAST hits to Uniprot proteins. Of the 421,229 contigs, 24,168 (5.7%) had BLAST hits to the Uniprot database, with matches to 12,331 unique proteins. Altogether, of the 526,099 contigs, 56,518 (10.7%) had significant BLAST hits to the Unitprot database, and matched 24,440 unique protein accessions. Larger numbers of contig hits but fewer matches to unique proteins were observed when compared to the zebrafish Refseq protein database (Table 5). Altogether, 66,285 (12.6%) had BLAST hits to known proteins in zebrafish Refseq protein database that matched 19,899 unique protein accessions. This seemingly low percentage of contigs with BLAST hits is partially due to a high proportion of short contigs in the assembly of expressed short reads, although the percentage of the unique proteins of zebrafish hit by the unique catfish sequences in this study is comparable to levels reported in our previous catfish EST project (WANG *et al.* 2010). Longer contigs were more likely to have BLAST hits to the annotated protein databases, 80% of our contigs with BLAST hits were over 350 bp in length, similar to observations in previous studies (WANG *et al.* 2010; NOVAES *et al.* 2008; PARCHMAN *et al.*2010). Nonetheless, BLAST searches identified a total of 24,440 unique protein accessions including 6,674 genes that were identified for the first time here from the catfish transcriptome.

Table 5. Summary of BLASTX searches to annotated protein databases. Contigs of two assemblies, the reference assembly with 104,870 contigs and the *de novo* assembly with 421,229 contigs, were used to search the Uniprot database and the zebrafish Refseq protein database to assess the number of related genes represented by catfish expressed sequences.

	Contigs hit Uniprot	% contigs with hits	Unique protein hits	Contigs hit zebrafish Refseq	% contigs with hits	Unique zebrafish Refseq hits
Reference assembly	32,350	30.9%	17,766	36,597	34.9%	14,874
De novo assembly	24,168	5.7%	12,331	29,688	7.1%	10,781
Total	56,518	10.7%	24,440	66,285	12.6%	19,899

To assess the coverage of the catfish transcriptome achieved by our sequencing effort, the distribution of gene ontology (GO) annotations in catfish was compared with that of zebrafish. The unique genes from the catfish and the zebrafish annotated database were analyzed using generic GO-slim terms with Blast2GO (LOMAX 2005). The percentages of annotated catfish sequences assigned to GO-slim terms are very similar to those of zebrafish genes (Figure 2), suggesting a generally similar distribution of genes in different functional categories, and the depth of the coverage of the transcriptome.

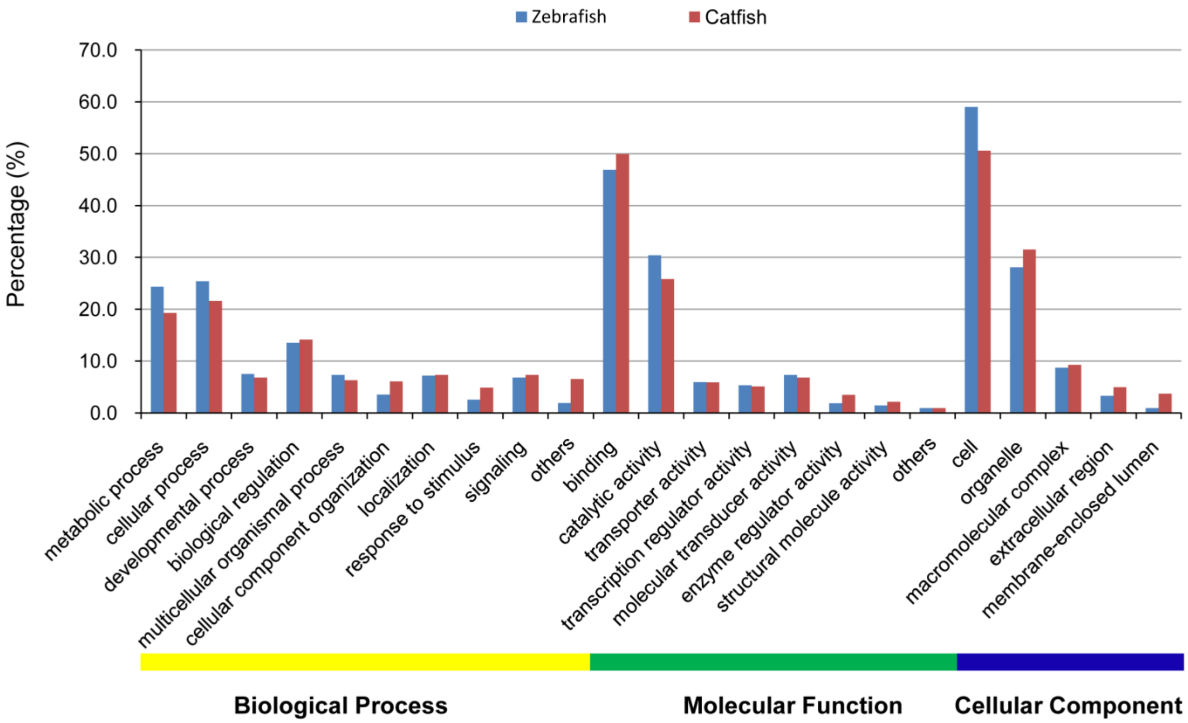

Figure 2. Similarity of GO-term assignments for catfish and zebrafish genes. Proportions of GO-terms assigned to annotated contigs from catfish assembly compared with the proportions found in the zebrafish genome annotation which serves as an indicator of the extent to which the catfish transcriptome has been characterized.

SNP identification

As summarized in Table 6, a total of 2,030,410 intra-specific SNPs were identified from the channel catfish sequence assembly; 2,497,806 intra-specific SNPs were identified from the blue catfish sequence assembly; and 4,236,135 SNPs were identified from the all catfish sequence assembly (intra-specific blue + intra-specific channel – intra-specific both + inter-specific). Almost two thirds of the putative SNPs were transitions.

Table 6. Summary of putative SNP identification. Putative SNPs include all base variations involved in the sequence assemblies with at least four sequences present at the SNP position with minor allele sequences represented at least twice. *All catfish* represents both intra-specific and inter-specific SNPs. Note that the total SNPs from all catfish assembly is fewer than the sum of total SNPs from channel catfish and blue catfish due to shared SNP positions in the two catfish species.

	Channel catfish	Blue catfish	All catfish
Contigs under analysis	523,815	525,428	526,099
Total SNPs	2,030,410	2,497,806	4,236,135
Transitions	1,311,220	1,616,477	2,751,244
Transversions	719,190	881,329	1,484,891
SNP/100 bp	1.6	1.8	3.0

Our previous research suggested that SNPs identified from contigs with at least four sequences at the SNP sites with the minor allele being represented at least twice are more reliable (WANG *et al.* 2008). In this study, putative SNPs were further screened following specific criteria based on the read depth, minor allele frequency, the quality of flanking regions and absence of additional SNPs in the 15-bp flanking regions (see Methods). With these criteria, a total of 342,104 putative filtered SNPs were identified from channel catfish; 366,269 putative filtered SNPs were identified for blue catfish (Table 7); of these 25,143 putative filtered SNPs were identified from same positions in both channel catfish and blue catfish, while 420,727 putative filtered inter-specific SNPs were identified. The number of intra-specific SNPs identified from same positions in both channel catfish and blue catfish may be underestimated, due to failure to capture sequences from one or both species in the current sequence data. A total of 146,573 filtered intra-specific SNPs in channel catfish were identified from positions where there were fewer than four blue catfish sequences, and similarly, 174,034 filtered intra-specific

SNPs in blue catfish were identified from positions where there were fewer than four channel catfish sequences. Obviously, the failure to obtain sequences from one or both species at same positions would also cause the underrepresentation of inter-specific SNPs.

Table 7. Quality SNPs selected from the putative SNPs with a set of criteria. [1]SNPs identified at positions where there were SNPs within channel catfish; [2]SNPs identified at positions where there were SNPs within blue catfish; [3]SNPs identified at positions where there were no intra-specific channel catfish SNPs or intra-specific blue catfish SNPs, but the bases differed between the two species.

	Intra-specific SNPs		Inter-specific SNPs[3]
	Channel catfish[1]	Blue catfish[2]	
Total SNPs	342,104	366,269	420,727
Transitions	208,517	230,031	262,048
Transversions	133,587	136,238	158,679
No. of contigs with SNPs	168,458	190,197	232,972
No. of contigs with Uniprot hits & SNPs	28,067	30,376	32,515
No. of unique known genes containing SNPs	16,562	17,423	18,085

Since the information on minor allele frequency (MAF) is an important consideration in choosing which SNPs to be included in SNP arrays, the minor allele frequencies of SNPs in the discovery populations were estimated from the sequence data. As shown in Figure 2, the majority of SNPs have sequence derived minor allele frequencies more than 15%, and the average MAFs were 0.28, 0.26 and 0.31 in putative filtered SNPs identified for channel catfish, blue catfish and inter-species, respectively (Figure 3).

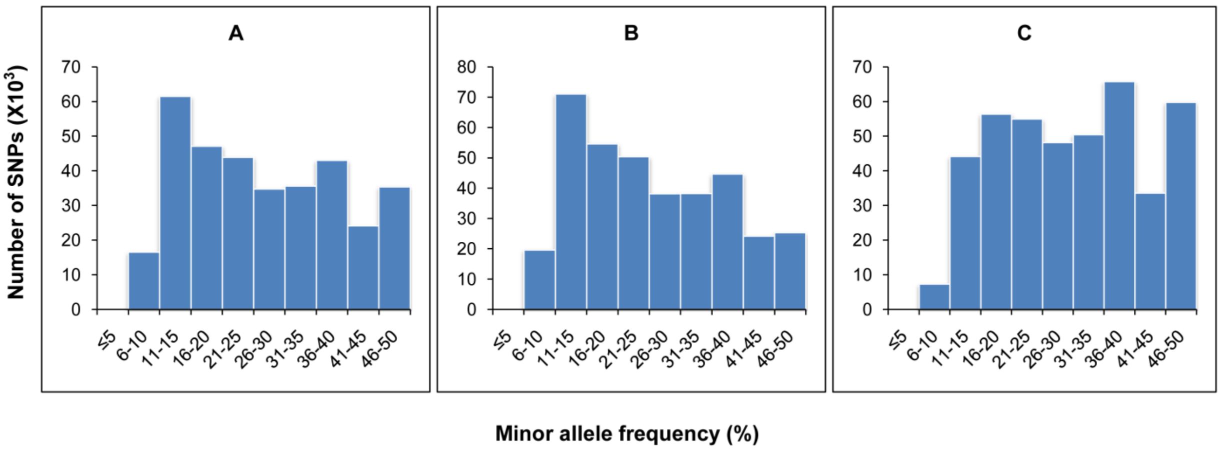

Figure 3. Distribution of minor allele frequencies of SNPs. A: Intra-specific SNPs in channel catfish; B: Intra-specific SNPs in blue catfish and C: Inter-specific SNPs between the two species. The X-axis represents the SNP sequence derived minor allele frequency in percentage, while the Y-axis represents the number of SNPs with given minor allele frequency. Note that the majority of SNPs have minor allele frequencies more than 10%.

While the number of SNPs is important, their distribution in contigs and genes within the genome is also important when used for genetic analysis. A total of 168,458 channel catfish contigs and 190,197 blue catfish contigs were found to contain putative filtered SNPs, of which 13,414 contigs contain SNPs at same positions in both channel catfish and blue catfish. The number of unique Uniprot accessions hit by contigs containing SNPs was 16,562 for channel catfish, and 17,423 for blue catfish, suggesting that putative filtered SNPs were identified from the vast majority of catfish genes.

One important aspect of using the inter-specific hybrid system is to identify inter-specific SNPs. From this work, a total of 232,972 contigs were identified to contain 420,727 inter-specific SNPs, i.e., sequence variations between the two species, channel catfish and blue catfish. These SNPs were from at least 18,085 distinct genes as determined by unique hits to the Uniprot protein database (Table 7).

Microsatellite markers identification

The 526,099 catfish contigs were surveyed to identify microsatellite markers. A total of 57,379 microsatellites were initially identified from 49,883 contigs. The majority of the microsatellites are dinucleotide repeats (Table 8). Of these microsatellites, 39,516 distributed within 34,539 contigs had sufficient flanking sequences on both sides for primer design. These microsatellites should be useful for genetic linkage mapping and other genetic studies.

Table 8. Summary of microsatellite markers identification.

Number of contigs of sequences surveyed	526,099
Number of contigs containing microsatellites	49,883
Total number of microsatellites identified	57,379
Di-nucleotide repeats	31,657
Tri-nucleotide repeats	16,925
Tetra-nucleotide repeats	8,235
Penta-nucleotide repeats	506
Hexa-nucleotide repeats	56
Number of microsatellites with sufficient flanking sequences	39,516
Number of contigs containing microsatellites with sufficient flanking sequences	34,539

Assessment of SNP distribution

SNPs distribution along the chromosomes of a genome is important for consideration of genome coverage using SNP markers. In the absence of a whole genome sequence assembly in catfish, we have taken a comparative genomic approach to plot the SNPs from expressed sequences onto the zebrafish genome sequence assembly. Contigs containing SNPs were used as queries against zebrafish transcripts to plot their putative genomic locations based on homology. As shown in Figure 3, the catfish expressed SNPs represent genes that are widely distributed along the chromosomes of all 25 zebrafish chromosomes. There are few gaps over one million base pairs in this comparative alignment (Figure 4).

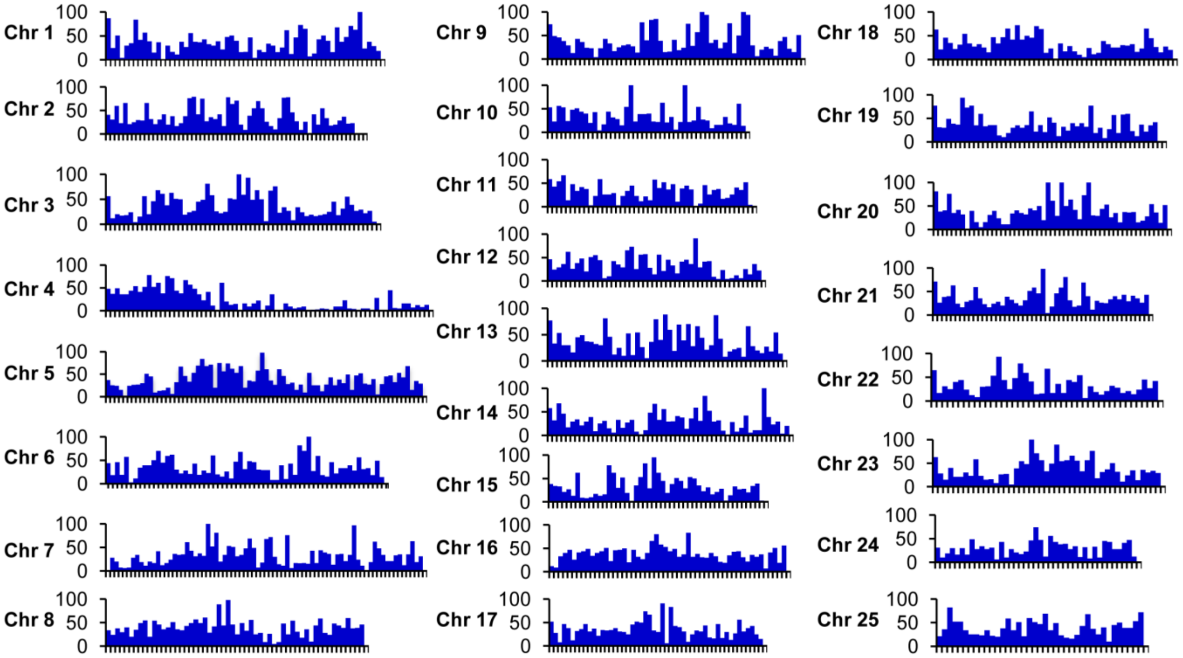

Figure 4. Comparative analysis of the genes containing SNPs on zebrafish genome. Each of the 25 zebrafish chromosomes was laid out in the X-axis with one million base pairs intervals, and the number of genes contained with filtered SNPs residing in the interval was plotted on the Y-axis.

Discussion

In this work, we have conducted RNA-Seq analysis with pooled RNA samples from multiple individuals of both channel catfish and blue catfish to develop large numbers of high-quality SNPs. A total of 493.4 million reads allowed generation of a total of over 35 billion base pairs of expressed sequences. Previous to this report, a total of approximately 290 million base pairs of expressed sequences of catfish had been generated using traditional Sanger sequencing. This work represents more than 100 times more transcript sequences than the total previously submitted to GenBank. Our results demonstrate the efficiency and cost-effectiveness of next generation sequencing technologies in generating expressed sequences.

One great challenge of transcriptome analysis using Illumina sequencing is the short read length. In this study, we have used both the Illumina GA-II and HiSeq 2000 sequencing platforms that generated read lengths of 36 bp or 100 bp. *De novo* assembly of the expressed short reads proved to be problematic even with gene-associated sequences. For instance, a total *de novo* assembly of the 218.8 million short reads from channel catfish would lead to over 800,000 contigs. Similarly, *de novo* assembly of 274.6 million short reads from blue catfish would lead to over 1,000,000 contigs. Such large numbers of short contigs may make subsequent applications of the EST or SNP resources less effective. However, such challenges are significantly alleviated when a large EST resource is available, as demonstrated by drastic reduction of contig numbers with the reference assembly in this study.

A second challenge is the over representation of highly expressed gene tags in transcriptome analysis. As shown in Figure 5, a small number of contigs (254) accounted for 32.6% of total reads. Clearly, there is a huge proportion of repeated sequencing and over representation of abundantly expressed genes. Obviously, this problem can be reduced by normalization of the cDNA. However, when a good EST reference is available, such a seemingly large problem is not as serious as the numbers indicate. As shown in Table 5, BLAST analysis of the reference assembly and *de novo* assembly contigs revealed that 24,440 unique Uniprot accessions were represented, suggesting that the expressed short reads provided good coverage of the catfish transcriptome. Additionally, previous, extensive EST sequencing of normalized and subtracted cDNA libraries resulted in 105,182 unique consensus sequences from channel catfish and blue catfish. Our sequencing here covered 104,870 of those contigs (99.7%), produced significant hits to 6,674 previously uncaptured genes, and covered thousands of additional transcript regions currently without annotation.

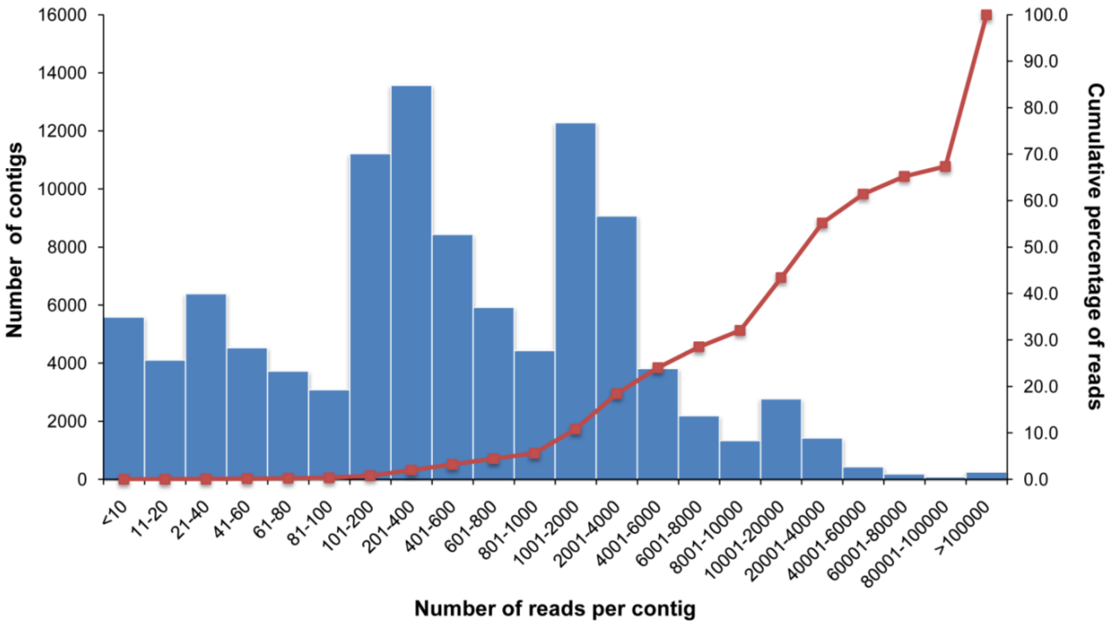

Figure 5. Frequency of contigs of various sizes from the all catfish reference assembly. The X-axis represents contig size (number of reads per contig). The curved line denotes the cumulative percentage of reads assembled. Note that a small number of very large contigs account for the majority of total reads. For instance, less than 0.3% of the contigs with over 100,000 reads per contig represent over 32% of all sequence reads assembled.

Pooling of RNA samples from multiple individuals followed by transcriptome analysis using next generation sequencing is among the most efficient methods for SNP identification. Through many years of efforts, a total of approximately 303,000 putative catfish SNPs were previously identified (WANG *et al.* 2010). However, this study alone allowed identification of over 2 million SNPs from channel catfish and almost 2.5 million SNPs from blue catfish. This efficiency is even more obvious when considering filtered (high-quality) SNPs. While only 48,594 filtered SNPs were identified among all catfish ESTs (WANG *et al.* 2010), this work resulted in 342,104 filtered SNPs within channel catfish and 366,269 within blue catfish. In addition, more than 420,000 filtered SNPs were identified as inter-specific SNPs, and are valuable in genetics and breeding studies involving hybrid catfish.

One major challenge for SNPs is the problem caused by paralogous sequence variants (PSVs) and multisite sequence variants (MSVs) (GUT and LATHROP 2004). Putative SNPs detected may be false positives, potentially arising from sequencing errors or misassembly of PSVs or MSVs. Paralogs that share high levels of sequence similarity may have been assembled in the same contig due to the short read length of Illumina reads. A higher stringency of assembly may better discriminate between paralogs, but complete discrimination may prove to be difficult due to the lack of a reference genome sequence. On the other hand, a higher stringency of assembly would lead to the separate assembly of haplotypes from highly polymorphic genes (NOVAES *et al.* 2008). Therefore, in order to select SNPs with high confidence, putative SNPs were screened based on several factors including surrounding sequence quality, absence of additional SNPs in the flanking regions, sequence depth and minor allele frequency. SNPs detected within contigs or regions of high sequence depth are more likely to be false positives. Therefore, setting a minimum minor allele frequency (e.g. 10%) for larger contigs may help reduce false SNP calling based on sequence errors. Additionally, multiple SNPs located close to one another (< 15 bp) often represent sequence errors and prevent the design of primers and probes for SNP genotyping. A requirement of no additional SNPs in the 15-bp flanking region around a putative SNP was therefore applied.

Given the large numbers of SNPs generated that meet these minimal requirements, more stringent parameters can be applied in picking SNP sets for different applications. Average depth at putative SNP positions is greater than 100 sequences, providing high confidence in accuracy of identified SNPs within the pooled samples. Re-sequencing or limited validation of these samples by low-throughput SNP genotyping is costly and is unlikely to generate additional information. Ultimately, SNPs need to be validated by genotyping in a variety of reference mapping families and trait-selected populations using a high-density screening array. In catfish, the use of homozygous gynogenetic catfish (WALDBIESER *et al.* 2010) as controls will allow detection of false positives caused by PSVs or MSVs.

Genome-wide association studies of complex traits require a large number of SNPs. However, for research communities focused on non-model organisms, it is cost-prohibitive to genotype all SNPs in an association study with the throughput of current technologies. Selection of uniformly distributed SNPs across the genome for association studies is therefore very important (ZHANG and SUN 2005). Gene-associated SNPs identified in this study, as anticipated, appear to be widely distributed across the catfish genome based on comparative analysis with zebrafish. About 30% of all contigs with identified SNPs had one SNP and 66% had three or fewer SNPs per contig (Figure 6). In absence of a whole genome assembly, the assessment of the exact pattern of the SNP distribution in the catfish genome is not possible. However, when the contigs containing filtered SNPs were plotted to the zebrafish genome by BLAST analysis, they had a good coverage of all regions of all 25 zebrafish chromosomes (Figure 4). While chromosome breakage, fusions, and rearrangements between catfish and zebrafish have occurred during genome evolution, at the genomic scale it is reasonable to assume that these widely distributed genes in the zebrafish genome will have a similar genomic distribution in catfish.

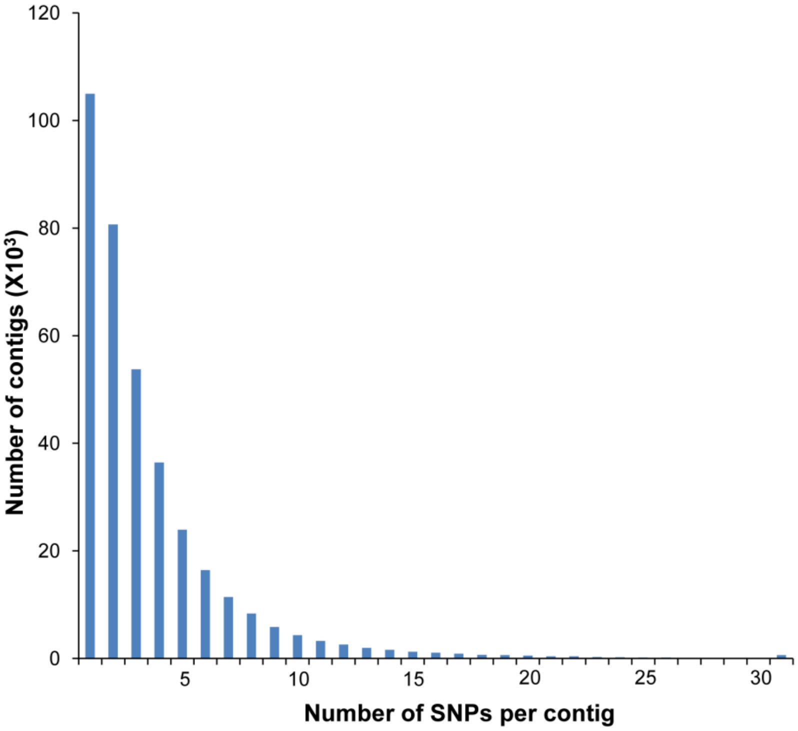

Figure 6. Distribution of filtered SNPs per contig. Histograms depict frequency of contigs with a given number of SNPs identified. Note that the majority of contigs have 5 or fewer SNPs per contig.

Conclusions

The approach to sample animals of diverse genetic backgrounds and sequence to sufficient depth for reliable SNP identification allowed the ability to detect many common SNPs across the entire genome. We have demonstrated that transcriptome analysis of pooled RNA samples from multiple individuals using Illumina sequencing technology is both technically efficient and cost-effective for generating expressed sequences. Such an approach is most effective when coupled to existing EST resources generated using traditional sequencing approaches because the reference ESTs facilitate effective assembly of the expressed short reads. The SNPs identified in this report will provide a much needed resource for genetic studies in the catfish scientific community and will contribute to the development of high density, cost-effective genotyping platforms. Validation and testing of SNPs using high-density arrays will subsequently lead to the production of a SNP array with well-spaced SNPs providing a powerful genotyping tool for the study of performance and production traits in catfish.

Acknowledgements

This project was supported by a grant from USDA AFRI Animal Genome Basic Genome Reagents and Tools Program (USDA/NRICGP award# 2009-35205-05101). Zunchun Zhou, Jianguo Lu, Fanyue Sun, Shaolin Wang, Hong Liu, Yanliang Jiang, Huseyin Kucuktas, Ludmilla Kaltenboeck, Eric Peatman and Zhanjiang Liu are coauthors on the manuscript. The authors want to thank Greg Whitis, Alabama Fish Farming Center (Greensboro, AL) for help in collecting representative fish from commercial populations. We thank all the lab colleagues who provided assistance with sample collection. Thanks go to Dr. Shawn Levy and team at HudsonAlpha Institute for sequencing efforts. Many thanks are given to Dr. Zhiliang Hu and Dr. Thomas Parchman for the helpful discussion and assistance during the data analysis. We are grateful for the Alabama Supercomputer Center for providing the computer capacity to do the bioinformatics analysis.

Chapter 3

Efficient assembly and annotation of catfish transcriptome: generation of full-length transcripts and detection of pseudo-SNPs

Abstract

Upon the completion of whole genome sequencing, thorough genome annotation that associates genome sequences with biological meanings is essential. Genome annotation depends on the availability of transcript information as well as orthology information. In teleost fish, genome annotation is seriously hindered by genome duplication. Because of gene duplications, one cannot establish orthologies simply by homology comparisons. Rather intense phylogenetic analysis or structural analysis of orthologies is required for the identification of genes. To conduct phylogenetic analysis and orthology analysis, full-length transcripts are essential. Generation of large numbers of full-length transcripts using traditional transcript sequencing is very difficult and extremely costly. In this work, we took advantage of a doubled haploid catfish, which has two sets of identical chromosomes and in theory there should be no allelic variations. As such, transcript sequences generated from next-generation sequencing can be favorably assembled into full-length transcripts. Deep sequencing of the doubled haploid channel catfish transcriptome was performed using Illumina HiSeq 2000 platform, yielding over 300 million high-quality trimmed reads totaling 27 Gb. Assembly of these reads generated 370,798 non-redundant transcript-derived contigs. Functional annotation of the assembly allowed identification of 25,144 unique protein-encoding genes. A total of 2,659 unique genes

were identified as putative duplicated genes in the catfish genome because the assembly of the corresponding transcripts harbored PSVs or MSVs (in the form of pseudo-SNPs in the assembly). Of the 25,144 contigs with unique protein hits, around 20,000 contigs matched 50% length of reference proteins, and over 14,000 transcripts were identified as full-length with complete open reading frames. The characterization of consensus sequences surrounding start codon and the stop codon confirmed the correct assembly of the full-length transcripts. The large set of transcripts assembled in this study is the most comprehensive set of genome resources ever developed from catfish, which will provide the much needed resources for functional genome research in catfish, serving as a reference transcriptome for genome annotation, analysis of gene duplication, gene family structures, and digital gene expression analysis. The putative set of duplicated genes provide a starting point for genome scale analysis of gene duplication in the catfish genome, and should be a valuable resource for comparative genome analysis, genome evolution, and genome function studies.

Introduction

Recent advances in next-generation sequencing enabled an array of whole genome sequencing or re-sequencing projects in both model and non-model species. Such efforts have produced a wealth of genome resources. However, thorough genome analysis thereafter is essential to associate genome sequences with biological meanings (ADAMIDI *et al.* 2011; BRUNO *et al.* 2010). An important step in genome analysis is to decipher the complete protein coding sequence (CDS) region of each gene. In eukaryotes, prediction of CDS regions in genomic sequence is complicated by the intron interruptions and the low proportion of protein coding regions in the genome. It is still problematic at present to predict the correct distribution of CDS regions solely based on genomic sequences (FURUNO *et al.* 2003). To obtain information about the portion of a genome that is transcribed as RNAs and then translated into proteins, a comprehensive set of full-length transcripts is needed (DENOEUD *et al.* 2008).

In teleost fish, genome annotation is further seriously hindered by genome duplication. Because of gene duplications, it's unable to establish orthologies simply by homology comparisons. Rather intense phylogenetic analyses or syntenic analyses of orthologies are required for the identification of genes. To conduct phylogenetic analysis and orthology analysis, full-length coding regions of transcripts are essential. Furthermore, the nature of highly diversified and duplicated genome of fish species hindered sequence assembly and complicated the genome annotation as well as SNP identification. Previous efforts aiming at SNP discovery for catfish rendered the issue about discrimination of false positive SNPs derived from paralogous sequences and multisite sequences (PSVs/MSVs) (LIU *et al.* 2011).

Obtaining large numbers of full-length transcripts is not an easy task. In most cases, full-length transcripts were obtained by Sanger-based sequencing of full-length cDNA clones. This strategy works well for many highly abundantly expressed and short transcripts because in such cases: 1) clones containing full-length transcripts can be readily identified, and 2) complete sequencing can be achieved through sequencing from both ends of the clones. However, such a task for rarely expressed genes and large transcripts can be troublesome because the identification of full-length cDNA containing clones is itself a huge challenge, and even if the clones are identified, complete sequencing of long inserts using Sanger sequencing can still be time-consuming and expensive. Recently, next-generation sequencing has been recognized as one solution for transcriptome sequencing at a reasonable cost (SANDMANN *et al.* 2011; WANG *et al.* 2009). High-throughput sequencing of cDNA (RNA-Seq) does not rely on prior knowledge, enabling interrogation of all transcripts including potentially novel transcripts uncaptured in

Sanger-based sequencing (TORRES *et al.* 2008). RNA-Seq can provide sufficient sequencing coverage on whole transcriptome scale to ensure the precision of each single base and integrality of full-length transcripts (GRABHERR *et al.* 2011; SANDMANN *et al.* 2011). However, huge amount of short reads generated from RNA-Seq make the transcriptome assembly difficult, which is not only impeded by repeats but also by alternatively spliced transcripts. Moreover, while genomic sequencing coverage is generally uniform across the genome, transcriptome sequencing coverage is highly variable, depending on gene expression levels, excluding the use of coverage information to resolve repeated motifs (ZERBINO and BIRNEY 2008). Since a transcriptome assembly with good quality is essential for all the downstream analysis, extra efforts are required to improve the transcriptome assembly. An optimized assembly strategy can be obtained by combinatory use of different assembly softwares, especially the ones using multiple *k*-mers to resolve the problem of biased coverage of transcriptome sequencing. A higher *k*-mer length will theoretically results in a more contiguous assembly of highly expressed transcripts. On the other hand, poorly expressed transcripts will be better assembled if lower *k*-mer lengths are used (ZERBINO and BIRNEY 2008). Therefore, multiple *k*-mer approach can be used to increase the assembly sensitivity and contiguity. With the fast development of assemblers able to efficiently handle a greater number of sequence reads (SIMPSON *et al.* 2009; ZERBINO and BIRNEY 2008), short-reads can be of considerable utility for assembling transcriptomes of non-model organisms (GIBBONS *et al.* 2009; GRABHERR *et al.* 2011; MIZRACHI *et al.* 2010). The most significant issue with RNA-Seq for the assembly of transcriptome, however, is the allelic variation. Most vertebrate species are diploid organisms and therefore two sets of chromosomes are involved in the generation of transcripts. Even if only one individual is used for RNA-Seq, the assembly of related transcripts from the two sets of chromosomes creates “haplotypes” that do not exist, and this alone prohibits assembly of full-length transcripts from multiple sequences. To overcome this huge problem, the ideal approach is to create an individual with doubled haploid genome and therefore there are no allelic variations such that the short reads from RNA-Seq can be assembled, not only technically feasible by the software packages, but also biologically meaningful as they are transcribed from the same sequences.

In catfish, years of efforts have resulted in nearly 500,000 quality ESTs (WANG *et al.* 2010), but a limited number of full-length transcripts were obtained through the traditional clone-by-clone approach using Sanger sequencing (CHEN *et al.* 2010). Recent efforts using RNA-Seq allowed the identification of a large number of transcripts in catfish (LIU *et al.* 2011), but the assembly of full-length transcripts was hindered by the reasons discussed above. In order to circumvent this problem and generate a large set of full-length coding sequences to support the genome sequencing and annotation in catfish, doubled haploid catfish was produced that has been demonstrated to harbor two sets of homozygous chromosomes (WALDBIESER *et al.* 2010). Such homozygous catfish is ideal material for the generation and assembly of full-length transcripts using RNA-Seq. Here we report deep sequencing of catfish transcriptome by RNA-Seq using this gynogenetic homozygous catfish. The two main objectives of this study were to develop a comprehensive set of reference transcript sequences for genome-scale gene discovery and expression studies in catfish; and to obtain a large number of full-length transcripts for whole genome annotation, duplicate gene identification, and facilitating detection of false SNPs derived from PSVs/MSVs.

Materials and Methods

Sample and RNA isolation

To better characterize the catfish transcriptome and improve discovery of variations derived from duplicated genes, a haploid transcriptome is needed. Doubled haploid channel catfish, that have two identical copies of each chromosome, were established using gynogenesis (WALDBIESER *et al.* 2010). Tissues were collected from a single doubled haploid female channel catfish adult for this study. The fish were euthanized with tricaine methanesulfonate (MS 222) at 300 mg/l before tissue collection. Samples of 19 tissues including head kidney, fin, pancreas, spleen, gill, brain, trunk kidney, adipose, liver, stomach, gall bladder, ovary, intestine, thymus, skin, eye, swim bladder, muscle, and heart were collected. Tissues were flash-frozen in liquid nitrogen and shipped on dry ice then stored at -80 ℃ until RNA extraction. Tissue samples were ground to a fine powder with mortar and pestle in the presence of liquid nitrogen and thoroughly mixed. A fraction of the tissue samples was used for RNA isolation. Total RNA was isolated using the RNeasy plus Mini Kit (Qiagen, USA) followed by DNase I (Invitrogen, USA) treatment according to the manufacturer's protocol as in previous study (LIU *et al.* 2011). Equal amount of total RNA from each tissue was combined and sent out for commercial sequencing.

Illumina sequencing

Sequencing was conducted commercially in HudsonAlpha Genomic Services Lab (Huntsville, AL, USA) similarly as in previous study (LIU *et al.* 2011). Briefly, 100 ng of total RNA was used to prepare amplified cDNA using Ovation RNA-Seq, a commercially available kit optimized for RNA sequencing (NuGEN Technologies, San Carlos, CA). The produced double-stranded cDNA was subsequently used as the input to the Illumina library preparation protocol starting with the standard end-repair step. The end-repaired DNA with a single ‘A’-base overhang is ligated to the adaptors in a standard ligation reaction using T4 DNA ligase and 2 μM-4 μM final adaptor concentration, depending on the DNA yield following purification after the addition of the ‘A’-base. Following ligation, the samples were purified and subjected to size selection via gel electrophoresis to isolate 350 bp fragments for ligation-mediated PCR (LM-PCR). Twelve cycles of LM-PCR were used to amplify the ligated material in preparation for cluster generation. The prepared cDNA library was sequenced for 100-bp paired-end reads on three flow cell lanes using the Hiseq 2000 platform, but one of which was partitioned for including three other samples, generating less number of sequences as in other two lanes for catfish. The image analysis, base calling and quality score calibration were processed using the Illumina Pipeline Software v1.4.1 according to the manufacturer's instructions. Reads were exported in FASTQ format and has been deposited at the NCBI Sequence Read Archive (SRA) under accession number SRA047025.

Assembly of expressed short reads

The raw reads were cleaned by trimming of adaptor sequences, ambiguous nucleotides (‘N’ in the end of reads) and low quality sequences with average quality scores less than 20. Reads less than 15 bp after trim were also discarded, the remaining reads were used in subsequent assembly. In order to obtain a comprehensive and reliable assembly, three different assemblers including CLC Genomics Workbench (version 4.2; CLC bio, Aarhus, Denmark), ABySS (version 1.2.6) and Velvet (version 1.1.02) were used for *de novo* assembly. In brief, the CLC *de novo* assembly was performed with a choice of an optimized *k*-mer length based on the input data by default settings. In case of ABySS and Velvet assembly, multiple *k*-mer approach with every other *k*-mer values from 21 to 95 for ABySS and from 45 to 95 for Velvet were used so as to maximize assembly contiguity and sensitivity. Subsequently, the multiple *k*-mer assemblies from ABySS and Velvet were merged by running the first stage of the trans-ABySS analysis pipeline (version 1.2.0) (ROBERTSON *et al.* 2010), respectively. Afterwards, these three assemblies were combined to produce the final non-redundant assembly. As anticipated, some identical contigs were generated from more than one assemblies introducing duplicates. The CD-HIT-EST (LI and GODZIK 2006) was used to remove redundancy and retain the longest possible contigs. The short redundant contigs were removed, and the remaining contigs composed the final assembly of non-redundant transcripts.

Functional annotation and identification of putative full-length transcripts

All the non-redundant transcripts from final assembly were searched against NCBI zebrafish RefSeq protein database and Uniprot/Swiss-Prot database using BLASTX with E-value ≤ 1e-10. The ORFs were predicted with the software orfPredictor (MIN *et al.* 2005) by using BLASTX as a guide for the prediction. The BLASTX-aided method detects ORFs by finding the starting methionine and stop codon in catfish transcripts relative to the same features in the most closely related species identified by BLASTX. In the cases where the catfish transcripts did not show high similarity to reference protein, ORF identified by finding the longest stretch of uninterrupted sequence between a start codon and a stop codon in both strand orientations. The completeness of ORFs in each transcript sequence was determined by the BLASTX alignment. We considered a full-length transcript to contain a complete CDS if the ORF revealed a start codon and stop codon in agreement with the match in the database. In the context of this work, the full-length transcript was defined as a consensus sequence containing the complete CDS and at least partial 5’ and 3’ UTR sequence. The start and stop codons of CDSs were used to define the boundary between the CDSs and the 5' and 3' UTRs. If a significant match did not contain a start codon in the 5' or a stop codon in the 3' end of the coding sequence and the pairwise alignment indicated that the transcript lacked some 5' or 3' coding sequence, it was considered to be a transcript with a partial coding sequence.

Detection of catfish putative duplicated genes

Since the doubled haploid channel catfish were used, there should be no allelic variations, and the gene-derived transcripts showing signs of "SNPs" would be assembled from duplicate gene copies. In order to detect the PSVs or MSVs derived from the duplicated genes, the catfish transcripts that had unique protein hits were used as reference, and all the short reads were mapped with the similarity of 99%. The "SNPs" (actually the PSVs or MSVs) were detected as SNP detection in previous work (LIU *et al.* 2011). Briefly, at least four short reads were required for "SNP" detection at each position, the minimum number of variant alleles was required as at least two, and minor allele frequency was required as at least 10%. Putative catfish duplicated genes were assessed by aligning the transcripts with the preliminary catfish whole genome assembly (unpublished data) using BLASTN with the E-value cutoff of 1e-10 and minimum alignment length of 100 bp.

Analysis of UTRs of full-length transcripts

The catfish Kozak consensus sequences were examined in the 5' UTR analysis. Eight-base sequences spanning from position -4 to position +4 of transcripts were selected. The extracted sequences were used as input into WebLogo (CROOKS *et al.* 2004) to assess the common Kozak consensus sequence in catfish transcripts.

For 3' UTR analysis, the TEIRESIAS-based pattern discovery tool (RIGOUTSOS and FLORATOS 1998) was used to search the most frequently occurring motifs. A search for putative polyadenylation signals (PAS) in full-length transcripts was performed using 35 bp sequence immediate upstream of the polyA tail as input. Pattern discovery tool conditions in the program were: “exact discovery”, L=6 and W=6. To elucidate the sequence patterns that could affect the efficiency of translation termination, the bases around the stop codons (-6 to +12) in the catfish full-length transcripts were extracted and illustrated using WebLogo (CROOKS *et al.* 2004).

To identify evolutionarily conserved regulatory motifs in catfish transcripts, we searched the UTR database collection (UTRdb) (GRILLO *et al.* 2010) using the 5’ and 3’ UTRs as queries with the pattern match program UTRscan.

Results

Transcriptome sequencing

Illumina sequencing was conducted to generate short reads of transcripts from a doubled haploid channel catfish. The cDNA was made from pooled RNA samples isolated from 19 tissues of the catfish (See Methods for details). High throughput sequencing was conducted using the Illumina Hiseq 2000 platform to generate 100-bp paired-end reads. A total of 315,703,698 reads (157,851, 849 from each end) were generated. After removal of ambiguous nucleotides, low-quality reads (quality scores < 20) and sequences less than 15 bp, over 300 million reads totaling 27.1 billion bases were obtained for further analysis (Table 1).

Table 1. Summary of data generated for catfish transcriptome. *Sequencing was conducted using the total RNA isolated from nineteen tissues of a doubled haploid channel catfish. Tissues include head kidney, fin, pancreas, spleen, gill, brain, trunk kidney, adipose, liver, stomach, gall bladder, ovary, intestine, thymus, skin, eye, swim bladder, muscle, and heart.

Sequencing	No. of tissues*	No. of raw reads	Read length (bp)	No. of reads after trim	Avg. length after trim (bp)	No. of bases after trim (Gbp)
s_4_1_1	19	49,077,054	100	48,743,295	92.8	4.5
s_4_1	19	128,862,236	100	124,882,873	87.6	10.9
s_5_1	19	137,764,408	100	133,926,741	87.1	11.7
Total	-	315,703,698	100	307,552,909	88.0	27.1

Transcriptome assembly

In order to obtain a comprehensive and reliable assembly, three different assemblers were used including CLC Genomics Workbench (version 4.2), ABySS (version 1.2.6) and Velvet (version 1.1.02) for *de novo* assembly. Although all these three assemblers use the same de Bruijn graph algorithm, they are different in how to treat sequencing errors, resolve ambiguities and utilize read pair information (FLICEK and BIRNEY 2009). Furthermore, multiple *k*-mer approach applied in ABySS and Velvet has the capability of producing a better assembly for transcripts from both highly and lowly expressed genes. Therefore, we believe that the combinatory use of these three assemblers could improve the assembly.

As shown in Table 2, the ABySS assembly generated 192,558 contigs with minimum length of 200 bp, and generated the highest N50 and average contig lengths (1,888 bp and 1,004 bp, respectively). The CLC assembly resulted in a total of 217,114 contigs (≥ 200 bp) with N50 length of 1,120 bp, and average contig length of 658 bp. The contigs from Velvet assembly were relatively short with N50 and average contig lengths of 809 bp and 580 bp, respectively. Though a smaller number of contigs with minimum length of 200 bp were assembled in the ABySS assembly, it allowed the assembly of the largest number of contigs with length greater than 1 kb among the three assemblers (56,229 contigs ≥1 kb).

Table 2. Summary of assemblies generated using various *de novo* assemblers. *CLC denotes the assembly generated by CLC Genomics Workbench, ABySS denotes the assembly generated by ABySS with multiple *k*-mers and merged by running the first stage of Trans-ABySS pipeline, Velvet denotes the assembly generated by Velvet with multiple *k*-mers and merged by running the first stage of Trans-ABySS pipeline, and Final merged denotes the final assembly generated by combining the three sets of assembly and retain only one of longest contigs for the ones assembled by more than one assemblers.

Assemblies*	No. of contigs ≥200bp	No. of contigs with length ≥N50	No. of contigs with length ≥1kb	Avg. contig length (bp)	N50 (bp)	Total size (Mbp)
CLC	217,114	29,564	33,332	658	1,120	142.8
ABySS	192,558	28,791	56,229	1,004	1,888	193.5
Velvet	311,734	56,471	42,685	580	809	181.0
Final merged	370,798	48,730	68,569	743	1,395	275.5

The multiple *k*-mer assemblies encompass both higher *k*-mer lengths to result in a more contiguous assembly of highly expressed transcripts, and lower *k*-mer lengths to better assemble poorly expressed transcripts to increase assembly sensitivity. In order to compare the ability of each assembler to reconstruct protein-coding gene transcripts on sensitivity and contiguity, each assembly was searched against annotated protein database using BLASTX. The significant unique protein hits were identified from both zebrafish RefSeq protein and Uniprot/Swiss-Prot database with E-value cutoff of 1e-10. The largest number of unique protein hits was obtained by Velvet assembly in comparison to the assemblies generated by CLC and ABySS. The contigs generated by ABySS showed higher N50 and average lengths, but a smaller number of unique protein hits was obtained (Table 2). By comparison, the contigs generated from CLC offered larger number of unique protein hits than ABySS, while providing higher N50 than Velvet.

A total of 370,798 non-redundant transcript-derived contigs were obtained as the final assembly based on the three sets of assemblies. The combined final assembly had a N50 length of 1,395 bp and average contig length of 743 bp, including 68,569 contigs with length greater than 1 kb (Table 2). Such assembly took advantage of each assembler to achieve both highest assembly sensitivity and contiguity. In order to assess the transcriptome capture obtained by the current work, we aligned all the channel catfish ESTs currently available in NCBI (354,488) with transcripts of the final assembly reconstructed in this study. The results showed that the vast majority (97% of the total) of the ESTs were represented in our data set showing ≥ 90% identity over a length of ≥ 100 bp and E-value of ≤ 1e-10. All the 370,798 transcript-derived contigs have been submitted to the NCBI Transcriptome Shotgun Assembly (TSA) database under accession numbers: JT123347-JT494144.

Assessment of transcriptome assembly

We previously generated a total of 1,087 full-length cDNAs from channel catfish by direct sequencing of full-length cDNA clones (CHEN *et al.* 2010). This independently generated full-length cDNAs provided the material basis for the evaluation of the transcriptome assembly presented here, at least partially. The previously sequenced full-length cDNA sequences were used as queries to compare the transcript contigs assembled in this study. Based on the comparison, a general assessment can be obtained on what proportion of transcripts was successfully reconstructed with full-length in this work. The results were summarized in Table 3. Of the 1,087 full-length cDNAs, 1,011 (93%) were reconstructed completely with the same open reading frames (ORFs) in the present study. Among the remaining 76 transcripts, 45 transcripts were almost fully assembled, but lacking some sequences at either 5' end or 3' end, and 31 transcripts (< 3%) were poorly assembled in the current study which did not get any specific match with contigs in the final assembly (Table 3). These results demonstrated that the assembly quality was highly reliable; and a small fraction of transcripts that cannot be assembled were due to the absence of sequences in the current RNA-Seq sequence pool.

Table 3. Assessment of transcriptome assembly. A total of 1,087 previously identified catfish full-length cDNAs were used to compare with the assembled transcripts in this study to assess the transcriptome assembly.

	Number	Percentage
Total previous full-length cDNAs	1,087	
Completely assembled with same ORFs	1,011	93.0%
Partially assembled	45	4.1%
Poorly assembled	31	2.9%

BLASTX annotation based on homology search

To identify the putative function of catfish transcripts, all the sequences were blasted against the reference proteins available in NCBI RefSeq and Uniprot/Swiss-Prot protein databases using BLASTX with E-value ≤ 1e-10. A total of 87,931 (23.7%) catfish contigs had significant hits to zebrafish RefSeq proteins corresponding to 19,711 unique proteins. When blasted against the Uniprot database, 80,012 (21.6%) catfish contigs had significant hits, corresponding to a total of 26,369 unique accession numbers and 17,669 unique proteins. Cumulatively, a total of 94,476 (25.5%) catfish contigs had significant hits to known proteins from either NCBI RefSeq or Uniprot, corresponding to a total of 25,144 unique proteins (Table 4). The remaining contigs may represent UTRs, non-protein coding genes or additional transcripts from catfish-specific genes which are too divergent to be annotated by homology search with current E-value cutoff.

Table 4. Summary of BLASTX searches to annotated protein databases. Final assembly containing 370,798 contigs was used to search the zebrafish RefSeq protein database and the Uniprot database to identify the homologous genes represented by the catfish expressed sequences.

	No. of contigs with hits	No. of unique accessions	No. of unique protein hits
Zebrafish RefSeq	87,931	19,711	19,711
Uniprot/Swiss-Prot	80,012	26,396	17,669
Total	94,476	-	25,144

Analysis of BLAST top hit species and potential presence of non-catfish transcripts

In addition to its high capability of capturing transcripts from the target species, high throughput transcriptome sequencing using the next-generation technology had enabled capture of expressed sequences from xenobiotic species that are commensal with the species of study (HALE *et al.* 2009; VERA *et al.* 2008). In this study, total RNA of catfish samples was isolated from 19 tissues including intestine and stomach that are particularly prone to contamination of xenobiotic species. Therefore, it is possible that some of the sequences in our dataset are from xenobiotic species that are commensal with catfish. We, therefore, conducted BLASTX analysis and attempted to determine species with which the top hits were generated. Our data showed that 89% of our sequences with a significant BLAST hit had the top hits from a vertebrate species (Figure 1A), and of these 76% had top hits from fish sequences (Figure 1B). Among the sequences with top hits from fish, 99% had top hits with zebrafish, as expected with the phylogenetic relationship (Figure 1C).

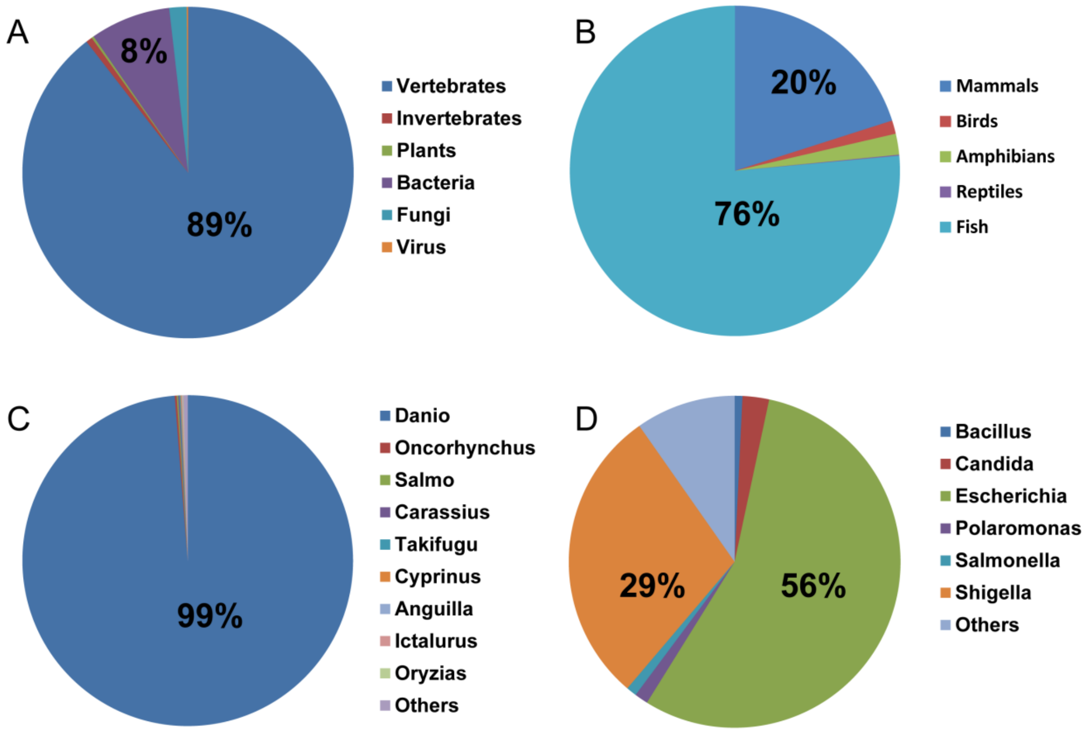

Figure 1. Distribution of taxonomic groups of BLAST top hit species. (A) The top hit species of BLAST searches are categorized into vertebrates, invertebrates, plants, bacteria, fungi and virus, and their percentages are presented. (B) The top hit species of BLAST searches within vertebrates are sub-categorized into mammals, birds, amphibians, reptiles, and fish, and their percentages are presented. (C) The top hit species of BLAST searches within fish are sub-categorized into various fish species as indicated, and their percentages are presented. (D)

The top hit species of BLAST searches within bacteria are sub-categorized into various bacterial species as indicated, and their percentages are presented.

However, it's noteworthy to mention that a good fraction (1,972 sequences, 8%) of our sequences had their top BLASTX hits with bacterial sequences, followed by those had top hits with fungi (424 sequences, 2%), and a few (33 sequences) even had top hits to viral sequences. Of these 2,429 sequences with top hits to other species than vertebrates, the top BLAST hit e-values were larger than those with top hits with vertebrates, in the range of 10^{-10} to 10^{-40}, whereas those with top hits to vertebrate species had a smaller e-value (often < 10^{-80}), suggesting the presence of xenobiotic species in the catfish samples. The top BLAST hits to bacteria were from a variety of species, of which 56% were *Escherichia coli*, and 29% were *Shigella* species (Figure 1D). Since our sequences were derived from multiple tissues, some of the bacterial sequences could belong to commensal microorganisms in digestive organs, although careful treatments had been taken to clean the tissues during collection process. Further investigation is warranted to fully understand these sequences most similar to xenobiotic species, which could provide information concerning symbionts and other microbial communities in catfish.

Conserved domain (CD) search for contigs without BLAST protein hits

After the annotation for catfish contigs based on homology search using BLASTX, a total of 276,322 catfish contigs did not get any significant protein hits from either zebrafish RefSeq or Uniprot protein database. It is reasonable to think that some protein-coding gene-derived contigs fail to get significant protein hits due to their short lengths. A comparison between the contigs with and without significant protein hits was conducted as shown in Figure 2. The majority of contigs that do not have significant protein hits from public protein database are with short length (83% less than 600 bp), while over 50% of contigs with protein hits are with length greater than 1 kb.

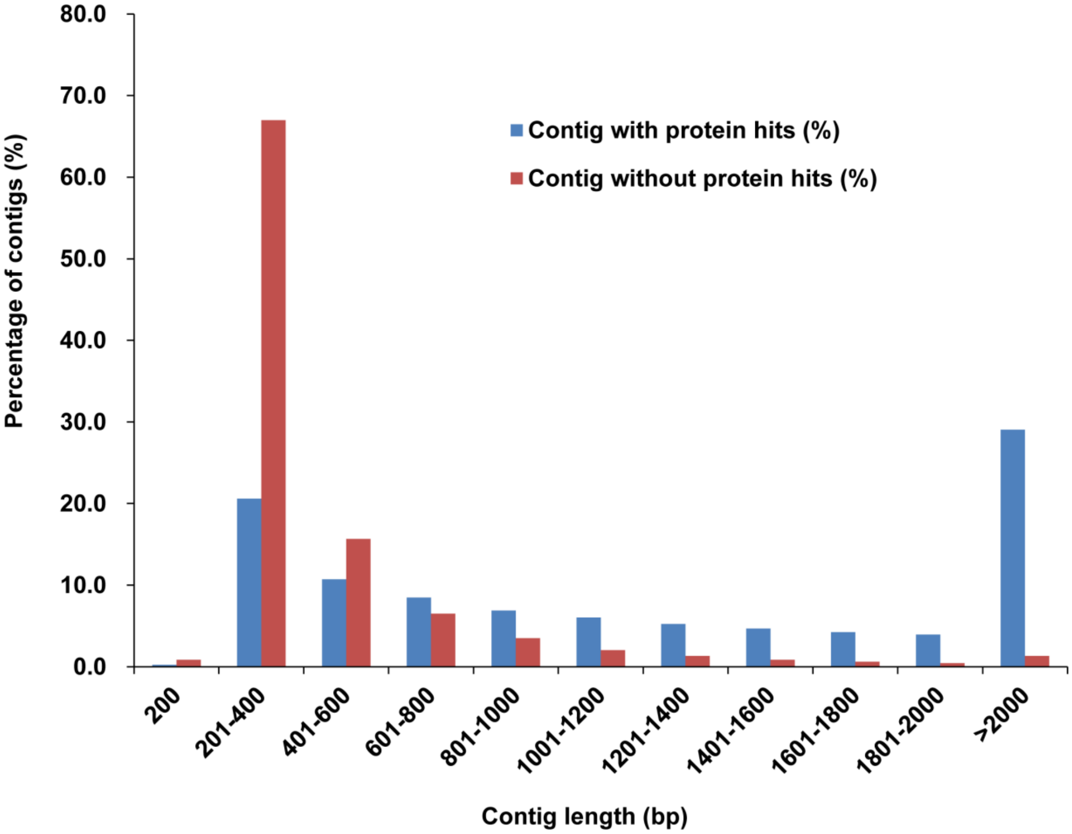

Figure 2. The length comparison between contigs with and without protein hits. The X-axis represents the contig length, and Y-axis represents the percentage of contigs. Note the high percentage of contigs that do not have significant protein hits in public protein database are in short length (83% less than 600 bp), and the high proportion of contigs with protein hits are long contigs.

In order to identify potential protein-coding genes from those contigs without significant blast hits from protein databases, we conducted *ab initio* prediction of the potential ORFs for the 276,322 catfish contigs that did not have significant hits to known proteins. A total of 260,793 contigs have putative ORFs detected with minimum length of 30 amino acids, and 16,688 of which possess putative ORFs with minimum length of 100 amino acids. To determine the putative functions of these ORFs, the CD-Search tool in NCBI was used to identify conserved domains, or functional units, within the protein query sequences. The specific hit found by a CD-Search indicates a high confident association between the protein query sequence and a conserved domain, resulting in a high confidence level for the inferred function of the protein query sequence. A total of 4,984 ORFs were identified with conserved domains, suggesting that such ORF-harboring contigs were derived from functional protein-coding genes as well.

Comparison of catfish transcripts with model fish species

In order to assess the capture of transcriptome obtained by catfish, reciprocal comparisons were conducted with the five model fish species with sequenced genome in Ensembl database including zebrafish, fugu, medaka, stickleback and *Tetraodon*. First, the protein sequences of the five sequenced fish species were mapped to the catfish transcripts using TBLASTN. From this, 39,250 (96.7%) of the 40,585 zebrafish proteins were matched, corresponding to 25,330 (96.9%) of the 26,152 Ensembl genes. The percentages of protein and gene match in other four fish were comparable to that observed in zebrafish, suggesting a high degree of transcriptome coverage obtained in catfish. Second, the catfish transcripts were mapped to the proteins of the five sequenced fish species using BLASTX (E-value ≤ 1e-10) to estimate the number of transcripts and genes represented in catfish. Based on zebrafish dataset, a total of 24,281 Ensembl zebrafish proteins were matched by catfish contigs, corresponding to 20,014 unique genes (Table 5). As indicated in Figure 3, zebrafish genes that had significant hits to catfish were relatively evenly distributed across 25 chromosomes (with the percentages ranging from 80% to 92%), suggesting the comprehensive capture of transcriptome for catfish genes on the genomic scale.

Table 5. Reciprocal BLAST comparison between catfish and five model fish. First, protein sequences of five model fish in Ensembl database were searched against catfish contigs using TBLASTN to assess the extent of reference sequences that were observed in catfish; reciprocally, all catfish contigs were searched against protein sequences of five model fish using BLASTX to

assess the number of genes represented by the catfish sequences when compared to those in each of the five species.

Species	#Total proteins	#protein coding genes	#proteins have hits in catfish	#genes have hits in catfish	#Unique protein hits by catfish	#Unique gene hits by catfish
zebrafish	40,585	26,152	39,250	25,330	24,281	20,014
fugu	47,841	18,523	47,583	18,336	26,230	15,834
medaka	24,661	19,686	23,311	18,465	18,200	16,007
stickleback	27,576	20,787	26,375	19,740	19,503	16,835
Tetraodon	23,118	19,602	22,648	19,208	17,636	15,996

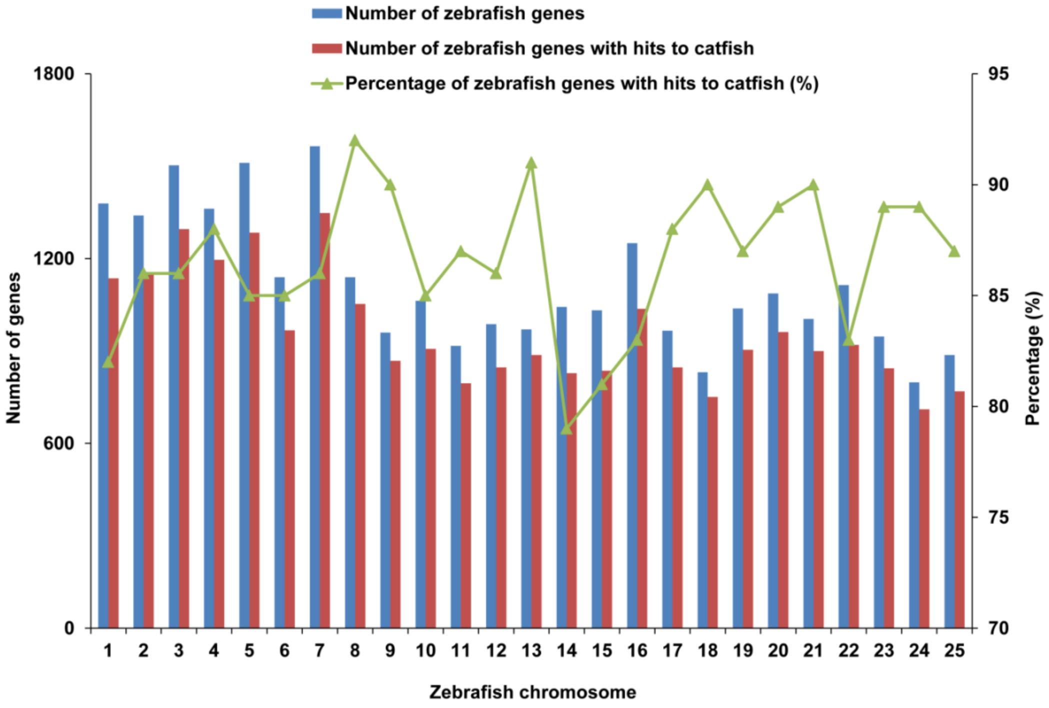

Figure 3. Distribution of identified catfish genes on zebrafish chromosomes. X-axis represents 25 zebrafish chromosomes. The left Y-axis represents the number of genes, and right Y-axis is the percentage of zebrafish genes on each chromosome identified in catfish.

Identification of gene duplicates

The highly diversified and duplicated genome hindered sequence assembly and complicated the genome annotation as well as SNP identification. Duplicated regions contain paralogous sequence variants (PSVs) or multisite variants (MSVs) which are readily mistaken for SNPs

(Fredman *et al.* 2004; Gidskehaug *et al.* 2011). In this work, a doubled haploid channel catfish individual was used, which provided the opportunity to examine the gene duplication in catfish at genome-scale. Such analysis was based on a simple principle that the doubled haploid channel catfish, which had two sets of identical chromosomes, should not contain allelic variations. Therefore, the transcripts should be derived from duplicated gene copies once there was any 'SNP' detected (Figure 4).

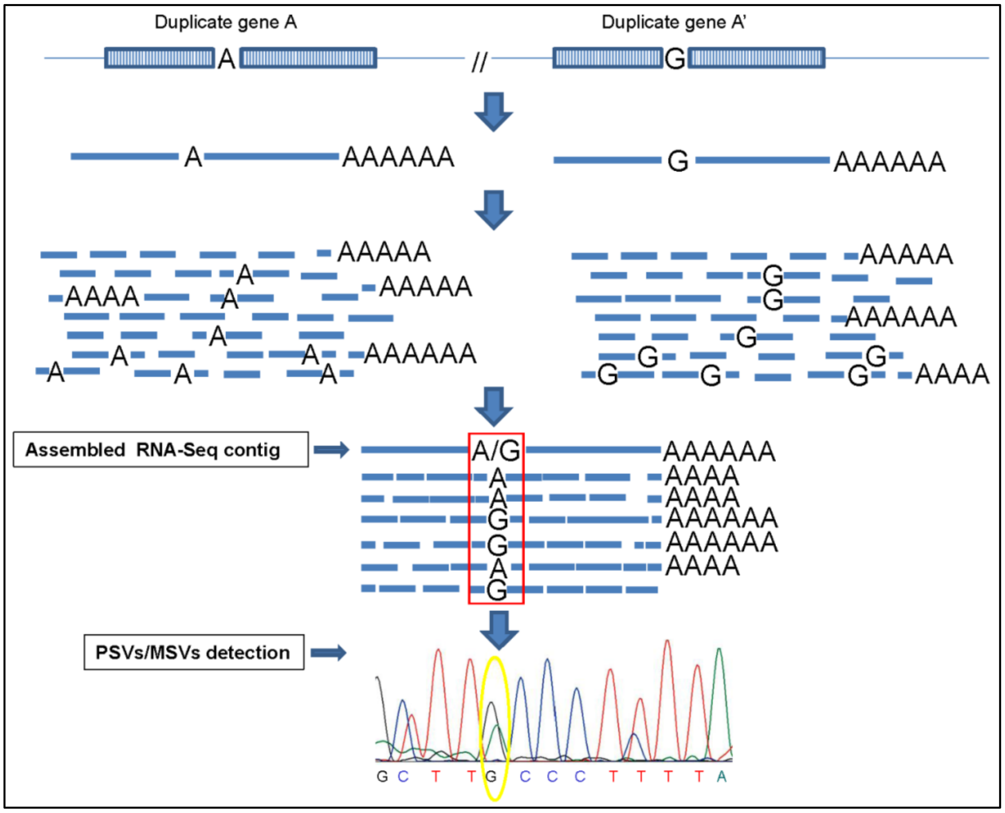

Figure 4. Detection of putative catfish gene duplicates and false SNPs. The reconstructed transcripts from protein-coding genes that show signs of "SNPs" (PSVs/MSVs) can be assembled by short reads from duplicated genes.

All the 25,144 contigs with unique protein hits were used for detection of "SNPs" as in previous study (Liu *et al.* 2011). In this analysis, a total of 4,878 PSVs or MSVs were detected from 2,692 transcripts accounting for 2,659 unique genes. Among the genes with PSVs or MSVs detected, the majority (67.3%, 1,789/2,659) contained only one PSV or MSV, while

113 (4.2%) genes contained greater than five PSVs or MSVs (Figure 5). These genes are putative gene duplicates in the catfish genome.

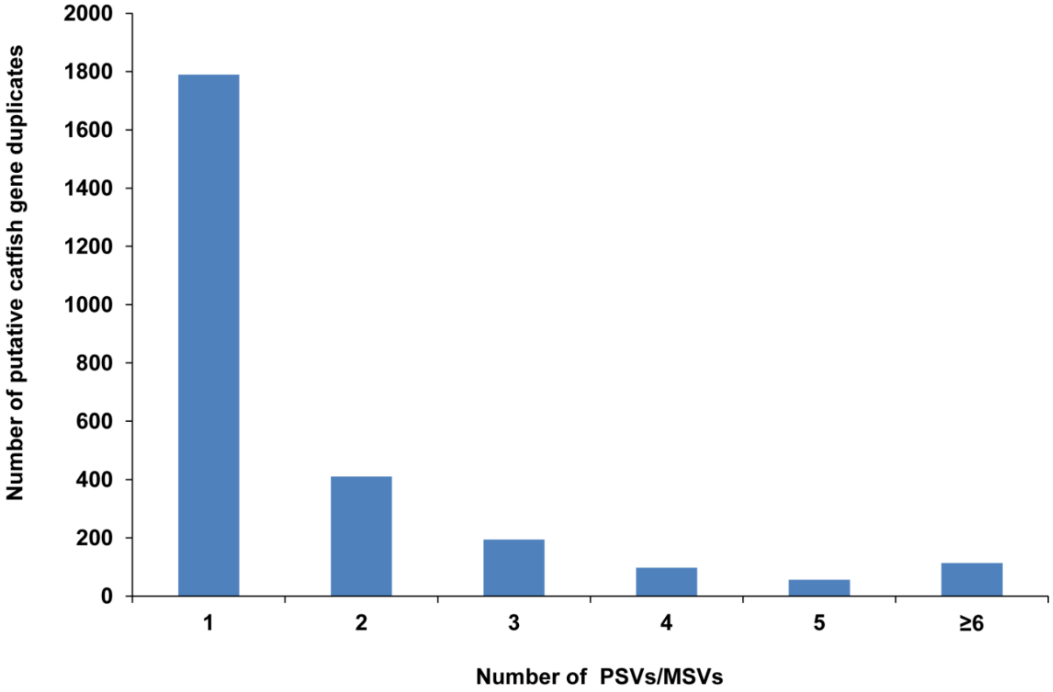

Figure 5. Detection of putative catfish duplicated genes. X-axis represents the number of PSVs or MSVs detected, while the Y-axis is the number of putative duplicated genes in catfish that contained the PSVs or MSVs.

To evaluate the detected gene duplicates, we aligned the transcripts to the preliminary assembly of the catfish genome (255,858 contigs with mean length of 2,996 bp and N50 of 6,027 bp, unpublished data). Duplicated genes were expected to be present in different genome locations (generally, i.e., genome contigs), and thus on different genomic contigs. The vast majority (92%, 2,446/2,659) of transcript-derived contigs hit more than one genomic contigs, suggesting their potential involvement in duplication. Of the remaining 213 transcript contigs, 196 contigs had only one catfish genome contig hits, indicating their uniqueness in the catfish genome. However, the possibility that it was because of the lack of corresponding genome contigs in the preliminary catfish genome assembly cannot be excluded. A total of 17 gene-derived transcripts did not get any significant match with genome contigs, suggesting the incompleteness of the current catfish preliminary assembly. Additional analysis of paralogous relationships may be strengthened by examination of physical locations in the genome for tandom duplications and flanking regions for segmental duplications when catfish genome scaffolds become available in the near future. It is understood that the same gene can still be placed into two or more genomic contigs with the preliminary assembly, but this problem should be overcome soon when the catfish genome assembly is completed.

Identification of catfish full-length transcripts

It has been shown that transcriptome sequencing using RNA-Seq can be a cost-effective approach for reconstruction of full-length transcripts in species without a reference genome (GRABHERR *et al.* 2011). In the context of this work, we took advantage of an individual homozygous catfish which provided a haploid transcriptome to facilitate the transcriptome assembly. A large fraction of transcripts were reconstructed and a large portion of these reconstructed transcripts were expected to be full-length transcripts. To identify full-length transcripts in our transcript collection, we utilized the functional annotation results with all the contigs in the final assembly being searched against zebrafish RefSeq and Uniprot database. The ORFs were predicted by using BLASTX-aided method which detects ORFs by finding the starting methionine and stop codon in catfish transcripts relative to the same features in the most closely related reference proteins identified by BLASTX.

The lengths of protein sequence, translated from the catfish transcripts which had significant protein hits from protein database, ranged from 21 to 12,863 amino acids (aa). The relationship between catfish proteins and homologous reference proteins is shown in Figure 6. As shown in Figure 6A, the most common occurrence is when the catfish protein and reference protein lengths are identical, which occurs with 17% (4,282/25,144) of the catfish transcripts with unique protein hits. The majority of catfish transcripts (66%) have translated proteins within ± 20% of their reference protein lengths (Figure 6A).

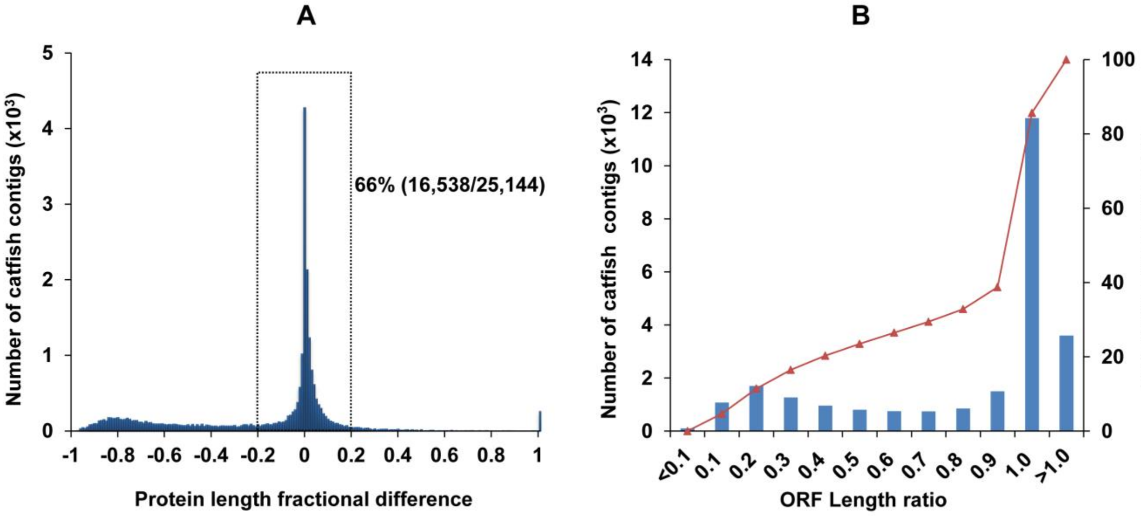

Figure 6. Comparison of catfish proteins with reference proteins from databases. (A) Fractional distribution of catfish proteins with lengths falling within various fractions, e.g., 0 indicates identical lengths of catfish proteins with reference proteins; -0.2 and 0.2 indicates the lengths of catfish proteins are 20% shorter or longer than those of reference proteins, respectively, and so on. Overall 66% or 16,538 transcripts out of a total of 25,144 identified unique catfish genes are within 80% bracket as compared with the lengths of reference protein counterparts. (B) Ratio of catfish predicted protein length versus length of reference protein was indicated in histograms (left Y axis), and the curved line denotes the cumulative percentage (right Y-axis). X-axis is the ratio of predicted catfish protein length to corresponding reference orthologous protein length, i.e. catfish protein length/reference protein length. Note that the unit of X-axis is ten times of that in (A), i.e. the ratio of 1.0 represents the transcripts of length within 95% bracket as compared with the lengths of reference protein counterparts, etc. The left Y-axis represents the number of occurrence of catfish protein lengths in thousand, and right Y-axis is the cumulative percentage.

In the present study, we defined the full-length transcript as a consensus sequence containing the complete CDS including the translational start codon ATG and the termination codon TAA, TAG, or TGA. To get an estimate of the number of transcripts containing full-length coding sequences, we used the ORF length ratio (catfish predicted protein length/reference protein length) to assess the completeness of CDS in each transcript. As indicated in Figure 6B, over 60%

of catfish contigs with unique protein hits contain predicted CDS with comparable length to reference protein (ORF ratio ≥ 1.0, 15,397/25,144). A total of 16,869 contigs with ORF ratio ≥ 0.9 potentially harboring full-length coding sequences were selected for further manual inspection (Table 6).

Table 6. Summary of full-length transcript identification. A total of 16,869 potential full-length transcripts that matched 90% length of reference proteins were selected for manual inspection. Full-length transcripts were defined as transcripts contain a complete CDS if the ORF revealed a start codon and stop codon in agreement with the match in the database.

	Number
Potential full-length transcripts	16,869
Full-length transcripts	14,240
Transcripts lack at 5' end	1,725
Transcripts lack at 3' end	711
Transcripts with incorrect ORFs	193

Manual inspection of completeness of CDS in each transcript sequence was determined by the BLASTX alignment. In this procedure, we consider a full-length transcript to contain a complete CDS if the ORF revealed a start codon and stop codon in agreement with the matched protein sequence in the database. As summarized in Table 6, a total of 14,240 were identified as full-length transcripts with complete coding sequences, 1,725 contigs and 711 contigs were partial transcripts lacking of 5’ end or 3’ end, respectively, and 193 contigs had incorrect open reading frames predicted due to either partial sequences or incorrect bases in start codon regions.

As shown in Figure 7, the majority of full-length transcripts were with lengths ranging from 1 kb to 4 kb with the average of 3,006 bp. The lengths of ORFs ranged from 132 bp to 15,684 bp, with an average length of 1,654 bp. The lengths of 5’ UTRs were relatively short, with the average of 254 bp, while the 3’ UTRs were longer with the average length of 1,096 bp. It is noteworthy that a large proportion of full-length transcripts contained “very” long 3’ UTRs (35% transcripts with length greater than 1.2 kb). This is quite different from the results obtained in our previous work where most of the 3’ UTRs had lengths less than 400 bp (CHEN *et al.* 2010). As indicated before, the bias of cDNA library creation and the selection process towards smaller transcripts could be the major reason for the short 3’ UTRs in previous study. The results shown here indicated that RNA-Seq for the full-length transcripts used in present work was capable to reconstruct most of the transcripts including the ones with long ORFs and 3’ UTRs that are difficult for full-length cDNA clone sequencing to get the full-length.

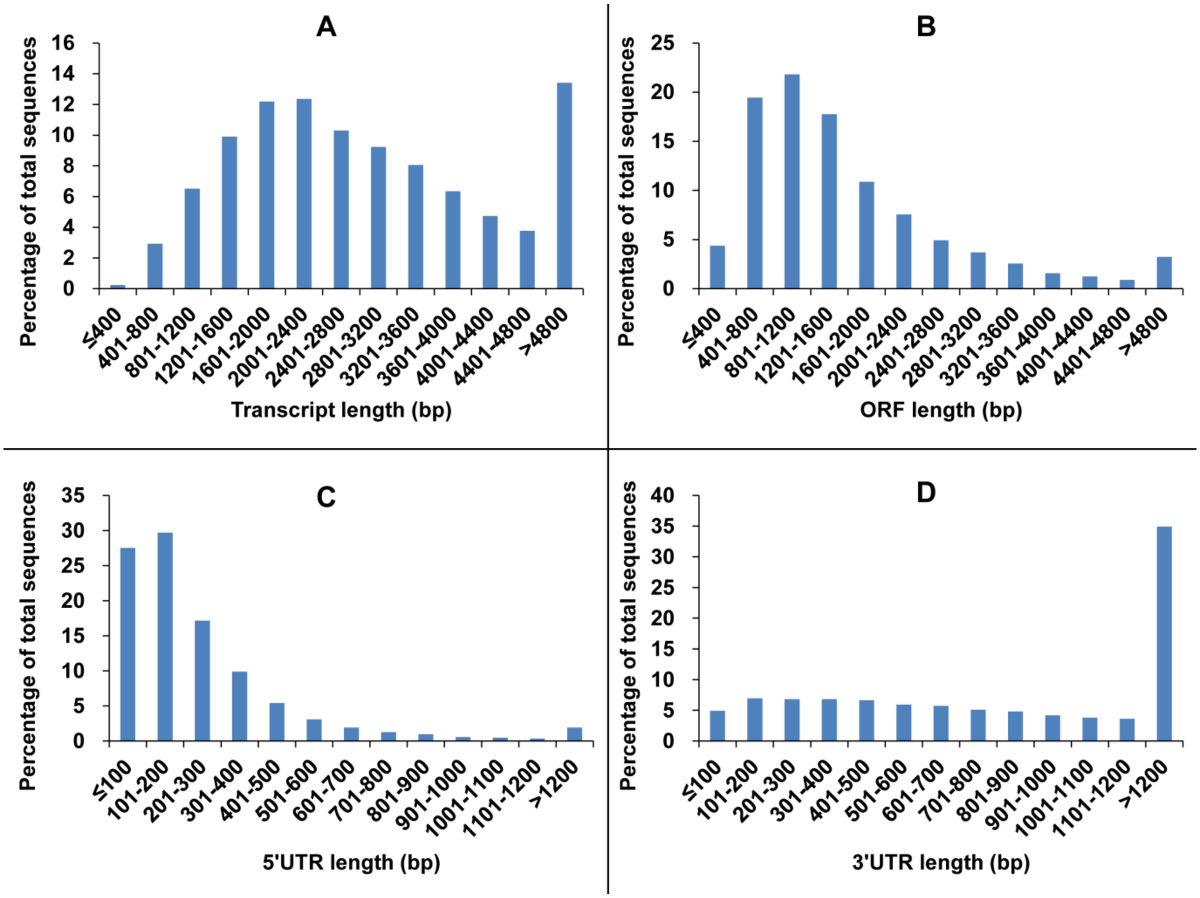

Figure 7. Length distributions of putative catfish full-length transcripts. Whole transcript (A), ORF (B), 5’-UTR (C), and 3’-UTR (D).

The protein sequences translated from catfish full-length transcripts had lengths ranging from 44 to 5,228 amino acids (aa). The majority of catfish proteins lengths were similar to that of their homologs. The relationship between catfish proteins and reference proteins from related species is illustrated in Figure 8, where the catfish protein lengths are plotted against the corresponding reference protein lengths. The majority of catfish proteins had same lengths as their homologous proteins in other species, but some are longer while others are shorter (Figure 8).

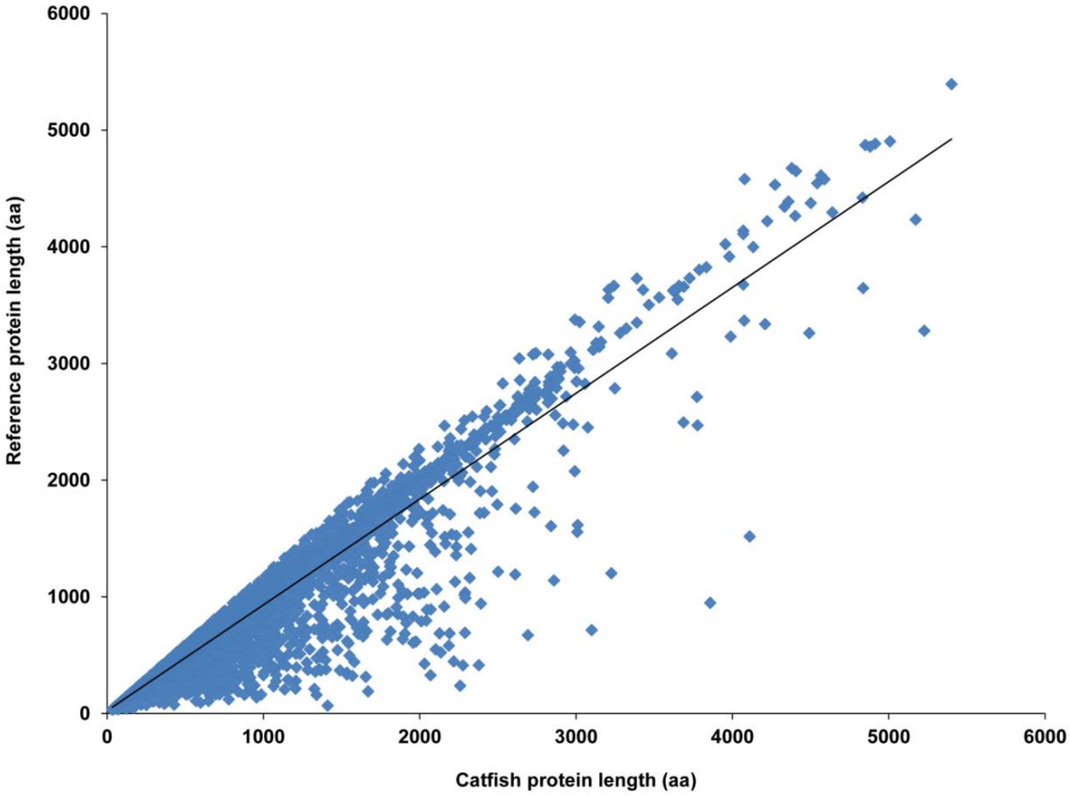

Figure 8. Comparison of catfish full-length transcripts with reference proteins. X-axis: catfish predicted protein length (amino acids), and Y-axis: reference protein length (amino acids).

Examination of sequences surrounding the start codon and stop codon of the catfish full-length transcripts

The AUG start codon context, also referred to as the Kozak motif, was reported as a consensus sequence for initiation of translation in vertebrates (KOZAK 1987). The catfish Kozak motif, spanning the position -4 to +4 was illustrated using WebLogo (CROOKS *et al.* 2004) as shown in Figure 9. The bases most frequently observed in the catfish Kozak motif are AAAC<u>ATG</u>G with the start ATG codon underlined, which is same as the results obtained in our previous work (CHEN *et al.* 2010). The most conserved bases are, as in other species, position -3 (A/G), the start codon (ATG) and position +4 (G). The consensus sequence of Kozak motif is reported as CACC<u>ATG</u>G in mammals (HARHAY *et al.* 2005), with the start ATG codon underlined. The most frequently observed bases in the Kozak motif of Atlantic salmon, *Salmo salar*, were CAAC<u>ATG</u>G (ANDREASSEN *et al.* 2009). Catfish Kozak consensus sequence appears to be highly similar to *S. salar* except that an adenine base instead of cytosine base was observed as the -4 position. The conservation of the catfish Kozak consensus sequence, as with the 5' UTR analysis, provided additional support for the proper identification of the start codon ATG in this work.

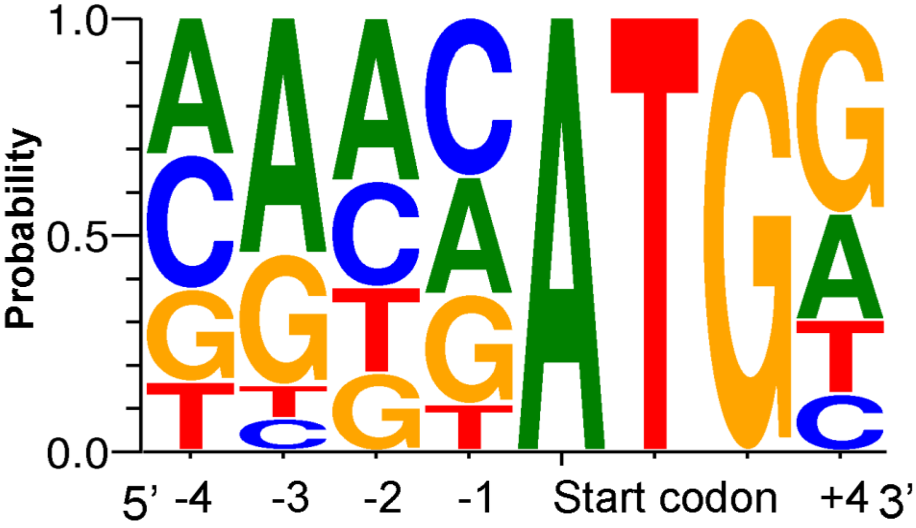

Figure 9. Analysis of Kozak consensus sequences surrounding the start codon AUG. Kozak consensus sequences were illustrated by WebLogo using stacks of symbols, one stack for each position in the sequence. The size of symbols within the stack indicates the relative frequency of each base at that position.

The translation termination was a very complex process, which involved stop codon recognition, peptidyl-tRNA hydrolysis and release of ribosome from the mRNA (FROLOVA *et al.*

2000; Kisselev and Buckingham 2000). The stop codon recognition was known as the first important step, while the contexts surrounding the stop codons were well known as a crucial determinant of the translation termination efficiency (Bonetti *et al.* 1995; Cassan and Rousset 2001; McCaughan *et al.* 1995; Tate *et al.* 1995). To elucidate the sequence patterns that could affect the efficiency of translation termination, the bases around the stop codons (-6 to +12) in the catfish full-length transcripts were examined. As illustrated in Figure 9, the bases around stop codon were biased. In particular, the -2 positions were biased toward A/U and the +4 positions were preferred for purines (A/G). The results we found in catfish were in agreement with previous studies in other eukaryotes, such as human, mouse, fruit fly and worm (Cavener and Ray 1991; Liu 2005). Numerous studies indicated that the nucleotide immediately following the stop codon (defined as +4) was crucial for termination and was biased toward purines (Bonetti *et al.* 1995; Cavener and Ray 1991). Translation termination was also influenced by the sequence elements immediately upstream of the stop codons (-2 and -1 bases) as indicated in many studies (e.g., (Mottagui-Tabar *et al.* 1998)), and the -2 positions were biased toward A and/or U in several eukaryotes previously examined (Liu 2005). Of the three stop codons, the usage frequency of the UGA is much higher than that of UAA and UAG in catfish (UGA 48.5%, UAA 32.5% and UAG 19%), which was consistent with general frequency of use of stop codons in eukaryotes (Liu 2005).

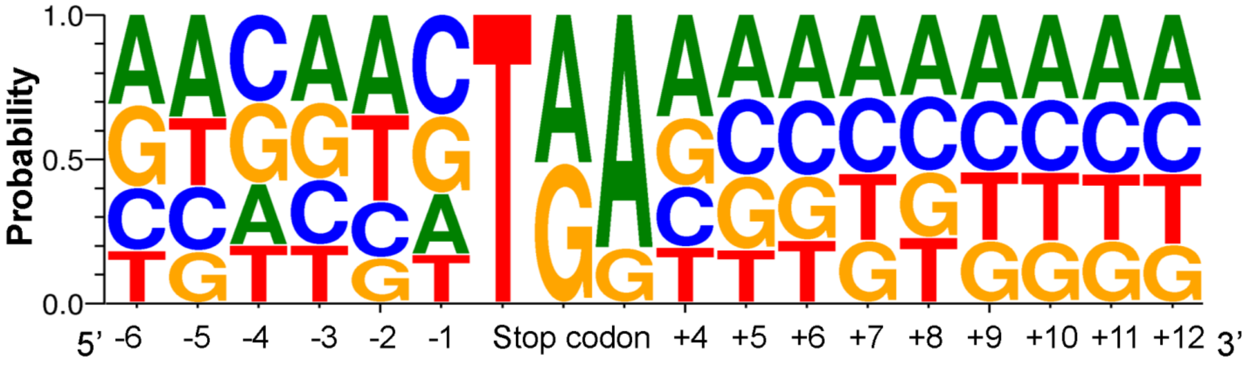

Figure 10. Analysis of consensus sequences surrounding the stop codon. The consensus sequences were illustrated by WebLogo using stacks of symbols, one stack for each position in the sequence. The size of the symbols within the stack indicates the relative frequency of each base at that position.

To our best knowledge, there were no systematic studies on the sequence contexts surrounding the stop codons in fish species. However, we reexamined our previously identified 1,087 full-length cDNAs generated from clone sequencing to reconfirm the results we provided here. The contexts around stop codons were highly consistent with the present results in Figure 9. The frequency of usage of the three stop codons were 45.4%, 38.6% and 16% for UGA, UAA and UAG, respectively, also similar as frequencies found from this study. This also provided additional support for the proper identification of the ORFs in this study.

Polyadenylation signal (PAS) in the catfish transcripts

The polyadenylation signal (PAS) plays an important role in polyadenylation by determining the site for addition of a polyA tail to pre-mRNA. The PAS in mammals has been widely investigated and has been identified as the canonical hexamer AAUAAA which is located 10-30 nucleotide upstream of the cleavage site (BEAUDOING *et al.* 2000; GRABER *et al.* 1999). Although the AAUAAA signal is considered to be highly conserved, studies have shown that different variants of the PAS are certainly present in the 3' ends of transcripts, and that the frequency distribution of the most common PAS versus alternate signals is species- and tissue-dependent (GRABER *et al.* 1999; MACDONALD and REDONDO 2002). The high diversity of gene expression of limited number of genes can be due to the 3' end polyadenylation of pre-mRNA (TUPLER *et al.* 2001). The choice of which element is used as PAS might play important roles in gene regulation, possibly by including or excluding down-stream regulatory motifs situated between such alternate PAS, as well as mRNA stability (BEAUDOING *et al.* 2000; GRABER *et al.* 1999; MACDONALD and REDONDO 2002).

For the identification of PAS in catfish transcripts, 2,047 transcripts with at least seven adenosines at the end were selected for this analysis. Different variants of the PAS were observed in catfish. The most common observed immediate upstream of the polyA tail (within 35 bp) was the canonical AAUAAA (1,264 transcripts, 61%). The second most common variant was AUUAAA, present in 435 transcripts and accounting for 21% (Table 7), similar results as obtained in previous work (CHEN *et al.* 2010). These findings are in agreement with PAS motif frequency distribution in other species. Previous studies in mammals for the incidences of AAUAAA and AUUAAA polyadenylation signals in 3' ends of cDNAs from EST sources have revealed that AAUAAA and AUUAAA are the two most common polyadenylation signals with the percentages of 50-60% and 10-15%, respectively (MACDONALD and REDONDO 2002). In order to find the most frequently occurring hexamers in the remaining catfish transcripts, the TEIRESIAS algorithm (RIGOUTSOS and FLORATOS 1998) was used. The results showed that no dominant hexamer present in the remaining transcripts, rather several additional hexamers that share sequence similarity with canonical PAS sequence were identified. Table 7 lists the twelve most frequently occurring hexamers (single-base variants of AAUAAA) present in 10 or more transcripts. Of these putative PASs, five hexamers have been reported as PAS variants in salmon or catfish transcripts previously (ANDREASSEN *et al.* 2009; CHEN *et al.* 2010), eight of them have been identified in human genes based on EST resources (BEAUDOING *et al.* 2000).

Table 7. Identification and analysis of the catfish polyadenylation signal (PAS). A subset of the full-length transcripts with polyA tails generated from sequencing were selected for analysis. Most common PAS alternatives (hexamers with single-base variant of AAUAAA) were identified within 35 nucleotides upstream of the polyA tail from full-length transcripts. [1]Alternative PAS identified in Chen et al. (2010) (CHEN *et al.* 2010); [2] Alternative PAS identified

in Andreassen et al. (2009) (Andreassen *et al.* 2009); and [3]Alternative PAS identified in Macdonald et al. (2002) (MacDonald and Redondo 2002).

Hexamer	Number of transcripts	Found in catfish[1]	Found in salmon[2]	Found in Human[3]
AAUAAA	1,264	+	+	+
AUUAAA	435	+	+	+
AAAAAA	47	+		
UAUAAA	25	+	+	+
AAGAAA	23		+	+
AAUATA	20	+	+	+
CAUAAA	19			+
AACAAA	18			
AAUACA	18			+
AAUGAA	16			+
AGUAAA	15	+		+
AAUAAU	10			
ACUAAA	10			+
AAUUAA	10			

Regulatory motifs in UTR regions of catfish full-length transcripts

Regulatory motifs are short nucleotide sequences that regulate gene expression. Most of the regulatory motifs are thought to be embedded in the non-coding part of the genomes. Among non-coding regions, the 5' and 3' UTRs of eukaryotic mRNAs have often been experimentally demonstrated to contain sequence elements crucial for many aspects of gene regulation and expression (Grillo *et al.* 2010). The large set of full-length transcripts generated in this study provided the opportunity for the identification of conserved regulatory motifs in the UTR regions of catfish transcripts. All 5’ and 3’ UTRs from catfish full-length transcripts were searched against UTRsite collection by using the pattern match program UTRscan. The UTRsite is a collection of functional sequence patterns located in 5' or 3' UTR sequences whose function and structure have been experimentally determined and published (Grillo *et al.* 2010).

Table 8. Identification of conserved regulatory motifs from untranslated regions. Catfish conserved regulatory motifs were identified by comparison with experimentally validated

regulatory motifs deposited in UTRsite database using pattern search program UTRscan. Number of UTRs denotes the number of catfish UTRs from which the corresponding regulatory motifs were identified.

Regulatory Motifs	Standard Name	Location (UTR)	Number of UTRs
IRE	Iron Responsive Element	5' and 3’	13
IRES	Internal Ribosome Entry Site	5'	2,491
SXL_BS	SXL binding site	5' and 3’	1,469
TOP	Terminal Oligopyrimidine Tract	5'	265
UNR-bs	UNR binding site	5' and 3’	907
uORF	Upstream Open Reading Frame	5'	5,492
15-LOX-DICE	15-Lipoxygenase Differentiation Control Element	3'	2
ADH_DRE	Alcohol dehydrogenase 3'UTR downregulation control element	3'	91
ARE2	AU-rich class-2 Element	3'	82
BRD-BOX	Brd-Box	3'	578
BRE	Bruno 3'UTR responsive element	3'	28
CPE	Cytoplasmic polyadenylation element	3'	180
G3A	Elastin G3A 3'UTR stability motif	3'	2
GLUT1	Glusose transporter type-1 3'UTR cis-acting element	3'	2
GY-BOX	GY-Box	3'	232
INS_SCE	Insulin 3'UTR stability element	3'	1
K-BOX	K-Box	3'	919
MBE	Musashi binding element	3'	3,546
SECIS1	Selenocysteine Insertion Sequence - type 1	3'	42
SECIS2	Selenocysteine Insertion Sequence - type 2	3'	36
TGE	TGE translational regulation element	3'	5

As summarized in Table 8, a total of 21 regulatory motifs were identified from catfish UTRs. For instance, the analysis of the 5’ and 3’ UTRs revealed transcripts with a motif that matched the iron responsive element (IRE). IRE, a particular hairpin structure located in the 5' UTR or 3' UTR of various mRNAs coding for proteins involved in cellular iron metabolism, is recognized by trans-acting proteins known as iron regulatory proteins that regulate mRNA translation rate and stability (KALDY *et al.* 1999). This evolutionary conserved motif was known to be present in the ferritin genes of vertebrates (THOMSON *et al.* 1999). Our observation that presence of an IRE located in the 5’ UTR of catfish ferritin supported the idea that the motif identified was a true functional IRE. The analysis of the 3’ UTRs of the full-length cDNAs also revealed transcripts with motifs that matched Selenocysteine Insertion Sequences (SECIS). The SECIS element is a specific 60 bp stem-loop structure located in 3’ UTRs of mRNAs, and required for decoding UGA selenocysteine instead of termination of translation (WALCZAK *et al.* 1996). Catfish transcripts with matches to the SECIS element encoded selenium-related genes, such as glutathione peroxidase (FAGEGALTIER *et al.* 2000), defender against cell death protein (FISCHER *et al.* 2001), selenoprotein and Glutaredoxin (BJORNSTEDT *et al.* 1997), also suggesting the correct identification of this functional element.

Discussion

The transcriptome sequencing enables various structural and functional genomic studies of an organism. Although a lot of Sanger-based EST sequencing projects had been carried out for comprehensive characterization of transcriptomes, expressed sequence data are still limited resources, specifically in non-model species. The next-generation sequencing technologies provide a low cost, labor-saving and rapid means for transcriptome sequencing and characterization. However, the *de novo* assembly of short reads without a known reference is still difficult (SCHUSTER 2008). High throughput 454 sequencing which generates longer reads has been widely used in many transcriptome sequencing studies in non-model species previously (MEYER *et al.* 2009; VERA *et al.* 2008). Recently, more and more studies have shown the feasibility of transcriptome assembly by using Illumina short reads (GIBBONS *et al.* 2009), especially with the combination of paired-end reads sequencing technology to facilitate the assembly. However, the differential gene expression results in variable coverage in transcriptome sequencing, the choice of a single *k*-mer value usually used in genome assembly cannot generate an assembly with emphasis on both transcript diversity and contiguity. Performing multiple assemblies with various *k*-mer lengths and to retain the best part of each one to form the final assembly has been shown effective for *de novo* transcriptome assembly (SURGET-GROBA 2010 and MONTOYA-BURGOS 2010). Furthermore, as each assembler utilizes different approaches to deal with sequencing errors and paired-end information, the assemblers may differ in their abilities to capture different portions of the transcriptome with accuracy. It is reasonable that merging assemblies from multiple assemblers might yield a combined assembly with higher accuracy. More comprehensive transcripts would be obtained with combinatory use of several *de novo* assemblers.

In this study, we reported an efficient assembly and annotation of the catfish transcriptome by applying several combined strategies. Firstly, we took advantage of a doubled haploid channel catfish to reduce the complexity of transcriptome. Most importantly, the doubled haploid allows biologically meaningful assembly of transcripts without artificially creating “haplotypes” that do not exist in nature.   The sequence assembly was facilitated since there are no allelic variations. Various tissues were collected with the aim to cover a comprehensive transcriptome. The paired-end reads were generated to resolve the assembly problem caused by repetitive regions. Secondly, we generated a final assembly with a combinatory use of three different widely used *de novo* assemblers. Multiple *k*-mer method was also used to enable the assembly sensitivity and contiguity. A higher N50 length and average length are considered as a benchmark for better assembly on contiguity. Our results showed that N50 length and average length of contigs varied greatly as a function of *k*-mer length, and also varied greatly between different assemblers. To get the optimum results, the validation of different assembly programs was conducted by comparing sequence similarity with closely related species. The ABySS generated contigs with higher N50 length and average length indicating its strength in generating assembly with better contiguity, while the Velvet generated sequences with more number and percentage of contigs showed significant similarity with zebrafish proteins. The CLC Genomics Workbench performed as intermediate between ABySS and Velvet according to both contiguity and sensitivity. The final assembly obtained by merging all these three sets of assembly provided a more comprehensive and accurate assembly.

We believe that the sequencing depth was sufficient to cover the vast majority of transcripts. To assess the depth of sequencing obtained for transcriptome assembly in this work, three lanes of sequencing was conducted with one of which being sequenced for around one fourth yield of a whole lane (Table 1) and the sequence datasets were resampled into several sub-datasets with various read depths. The *de novo* assemblies of these sub-datasets were generated using CLC Genomics Workbench to determine the effects of read depths on the transcriptome assembly. The number of contigs with minimum length of 200 bp and 1 kb were collected as two benchmarks for the assembly sensitivity and continuity. As shown in Figure 11A, with the increase of sequencing depth, the number of contigs with minimum length of 200 bp increased. A significant increase of transcriptome coverage (assembly sensitivity) was observed from 48M to 124M. Relatively slight increase was observed when the number of reads increased from 258M to 308M, but nonetheless the number of assembled contigs with sizes greater than 200 bp or 1 kb continues to grow as the sequence depth increased. This was somewhat outside of our expectation, perhaps due to the segmented assemblies. We therefore decided to determine if the percentage of gene hits continue to increase with the increase of sequencing depth.

The transcriptome coverage and completeness achieved by these sub-assemblies were then evaluated by matching with annotated known genes. Due to the limited number of catfish genes in the public database, we used the zebrafish RefSeq protein sequences in the NCBI database to conduct this evaluation. There are a total of 27,239 zebrafish annotated protein sequences available which were searched against the assembled catfish contigs for homologous match using TBLASTN. There were 25,795 zebrafish proteins observed in the catfish RNA-Seq assembly accounting for 94.7% of all annotated proteins in the database. As shown in Figure 11B, the number of observed genes increased almost linearly with lower sequencing depth. However, the number of observed genes started to plateau when the sequencing depth reached 124 million reads. From 124 million reads to 308 million reads, the percent of gene hits increased from 94% to 95%; while the percentage of gene hits with greater than 90% length homology stayed essentially unchanged, suggesting that the sequencing depth was sufficient to provide a good coverage of the transcriptome.

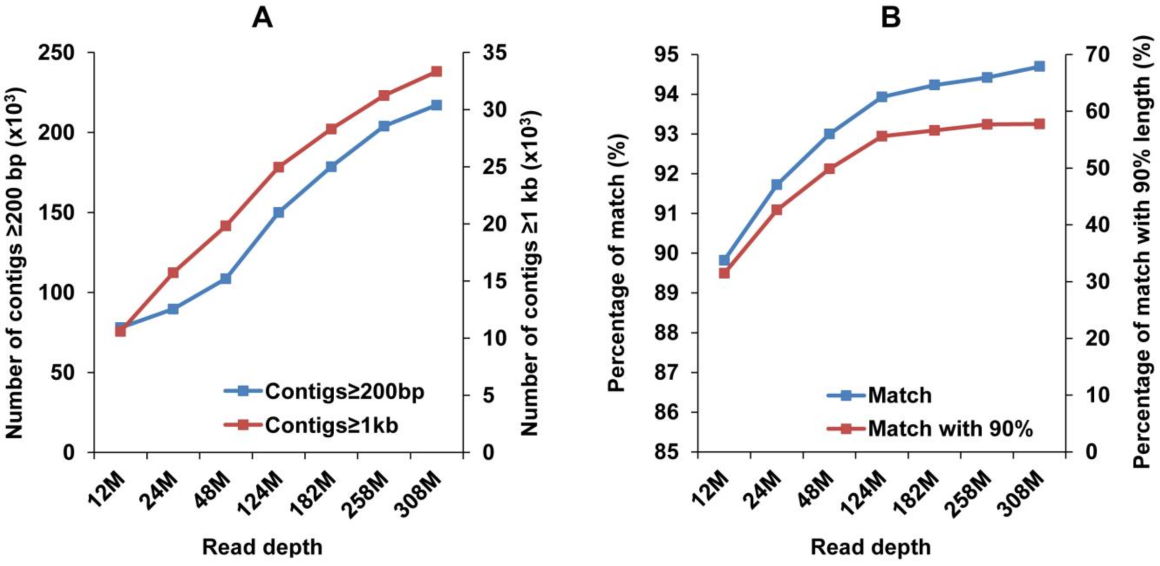

Figure 11. Evaluation of sequencing depth for the catfish transcriptome assembly. (A) Assemblies were evaluated based on the number of assembled contigs with length ≥ 200 bp and 1 kb. The X-axis represents assemblies with various sequencing depths generated by CLC Genomics Workbench, left Y-axis represents the number of contigs with length ≥ 200 bp in thousand, and right Y-axis represents the number of contigs with length ≥ 1 kb in thousand. (B) Assemblies were evaluated based on the number of zebrafish proteins that were identified in the assembled catfish contigs. The X-axis represents assemblies with various sequencing depths assembled by CLC Genomics Workbench, left Y-axis represents the percentage of zebrafish

proteins that can be detected in catfish and right Y-axis represents the percentage of zebrafish proteins that can be detected in catfish with match length ≥ 90%.

The analysis of sequence conservation comparing with other model fish species helps in transfer of knowledge from model species to catfish for both structural and functional genomic studies. A large number of catfish transcripts showed significant similarity with model fish at protein level as expected, suggesting that their function might also be conserved. Interestingly, a large number of the transcripts did not show significant homology with any other reference sequences, which may be novel and transcribed from catfish-specific genes. The study of these genes will be very important to dissect the species-specific cellular process, ‘catfish-specific’ gene duplication and divergence and study evolutionary processes of speciation and adaptation.

The doubled haploid fish used in this work provided an opportunity to evaluate genome-scale gene duplication in catfish. There were a total of 2,569 unique genes detected as putative duplicated genes. Ultimately, duplicated genes will have different genome coordinates as to their locations. However, the catfish genome sequencing is still in progress. Nonetheless, the evaluation based on the catfish preliminary genome assembly (unpublished) supported that the majority of these genes had duplicate copies. Clearly, the use of doubled haploid catfish was not only important for the transcript assembly, but also important for the initial identification of putative gene duplications in catfish. Additional analysis and validation are needed to demonstrate that the putative duplicated genes are indeed duplicated in the future.

In conclusion, we have demonstrated the use of short-reads sequence data to efficiently and comprehensively characterize a draft transcriptome of an organism without sequenced genome. The strategy of *de novo* assembly described here can be potentially used for other species. The advantages offered by the use of a homozygote are applicable to most teleost species where doubled haploid can be made. Our study contributed a significant non-redundant set of 370,798 transcripts including 14,240 full-length transcripts in catfish. The detailed analyses of these sequences has provided several important features of catfish transcriptome such as length distribution, sequence patterns around translation initiation and termination codons, conserved regulatory motifs, conserved genes across fishes, and functional annotation. It is anticipated that the results from this study will contribute significantly towards assembly and annotation of the catfish genome. Such resources will likely be important for structural and functional genomic studies in other teleosts and related species as well.

Acknowledgements

This project was supported by Agriculture and Food Research Initiative Competitive Grant no. 2009-35205-05101 and 2010-65205-20356 from the USDA National Institute of Food and Agriculture (NIFA). Yu Zhang, Zunchun Zhou, Geoff Waldbieser, Fanyue Sun, Jianguo Lu, Jiaren Zhang, Yanliang Jiang, Hao Zhang, Xiuli Wang, K.V. Rajendran, Lester Khoo, Huseyin Kucuktas, Eric Peatman and Zhanjiang Liu are coauthors on this manuscript. The authors want to thank Ludmilla Kaltenboeck for lab assistance during RNA preparation. Thanks are also given to Alabama Supercomputer Center for providing the computer capacity for the bioinformatics analysis.

Chapter 4

Development of the high-density catfish 250K SNP array for genome-wide association studies

Abstract

Quantitative traits, such as disease resistance, are most often controlled by a set of genes involving a complex array of regulation. The dissection of genetic basis of these traits requires large numbers of genetic markers. In recent years, next-generation sequencing technologies have enabled discovery of genome-wide SNPs. In catfish, over eight million SNPs have been identified, but the challenge is how to efficiently and economically use such SNP resources for genetic analysis. One of the most efficient ways to genotype large numbers of SNPs is to design a high-density array that includes hundreds of thousands of SNPs distributed throughout the genome. In this work, we developed a catfish 250K SNP array using Affymetrix Axiom Genotyping technology. Catfish SNPs were obtained from multiple sources including gene-associated SNPs, anonymous genomic SNPs, and interspecific SNPs. A set of 640K high-quality SNPs were obtained following specific requirements of array design, and were submitted for evaluation. A panel of 250,113 SNPs was finalized for inclusion on the array. The performance evaluated by genotyping individuals from wild populations and backcross families suggested the good utility of the catfish 250K SNP array. This is the first high density SNP array for catfish. The array should be valuable in both industry and research such as in whole genome-based selection, genome-wide association studies, fine mapping, high density linkage mapping, and haplotype analysis.

Introduction

Catfish is the most important aquaculture species in the United States. In recent years, catfish industry encounters great challenges including devastating diseases which cause the largest economic loss. Improved brood stocks with a high level of disease resistance are desperately needed. In addition to traditional selective breeding programs, efforts have been made to develop resources for marker-assisted selection based on genomic information (GODDARD and HAYES 2009; MEUWISSEN *et al.* 2001).

Quantitative traits, such as disease resistance, are most often controlled by a set of genes involving a complex array of regulation. The dissection of genetic basis of these traits requires large numbers of genetic markers. Single nucleotide polymorphisms (SNPs) are now the markers of choice because they are the most abundant type of genetic variations and widely distributed in the genome. In addition, SNPs are generally bi-allelic that is amenable to automated genotyping (KRUGLYAK 1997). SNPs are also efficient for genome-wide association studies because linkage disequilibrium (LD) can be detected with high density SNPs covering the genome when dealing with complex traits. For instance, simultaneous analysis of thousands of SNPs have enabled genome-wide association studies for performance and production traits in chicken (ABASHT and LAMONT 2007; WOLC *et al.* 2012), pig (BECKER *et al.* 2013; SAHANA *et al.* 2013), cattle (KHATKAR *et al.* 2008; KIM *et al.* 2009; MEUWISSEN *et al.* 2001), horse (BROOKS *et al.* 2010) and sheep (BECKER *et al.* 2010; KIJAS *et al.* 2009). However, such studies have not been possible with most aquaculture species including catfish due to lack of genome-wide SNP markers and high-throughput SNP genotyping platform.

Until recently genome-scale SNP identification was a great challenge for most non-model species (ALTSHULER *et al.* 2000; LINDBLAD-TOH *et al.* 2000; WANG *et al.* 1998). The next-generation sequencing technologies enabled efficient identification of SNPs from genomes of various organisms (DAVEY *et al.* 2011). With the availability of a large number of SNPs, the challenge then is how to genotype these SNPs efficiently and economically.

A variety of SNP array platforms are available, of which, the Sequenom MassArray (Sequenom, San Diego, CA), Illumina iSelect HD Custom BeadChip (Illumina, San Diego, CA), and Affymetrix GeneChip Custom Array (Illumina, Santa Clara, CA) are widely used. More recently, Affymetrix adopted the Axiom genotyping technology that allows development of full- or semi-customized arrays containing 1,500 to 2.6 million SNPs (HOFFMANN *et al.* 2011). These platforms differ in their requirements for SNP marker number, sample size, cost and automation.

Up to date, SNP arrays have been developed in several livestock species, including cattle (MATUKUMALLI *et al.* 2009), horse (MCCUE *et al.* 2012), pig (RAMOS *et al.* 2009), sheep (MILLER *et al.* 2011), dog (MOGENSEN *et al.* 2011) and chicken (GROENEN *et al.* 2011). The Illumina BovineSNP50 Beadchip featuring 54K informative SNP probes was first developed for detecting variations in cattle breeds (MATUKUMALLI *et al.* 2009). Recently, two cattle SNP arrays with higher-density were developed: the 777K BovineHD array by Illumina and the 648K Axiom BOS 1 by Affymetrix (RINCON *et al.* 2011). The Illumina PorcineSNP60 was designed to include over 64K pig SNPs identified by next generation sequencing (RAMOS *et al.* 2009). The equine SNP array (EquineSNP50 BeadChip) has been developed and evaluated on a panel of samples representing 14 domestic horse breeds and 18 evolutionarily related species (MCCUE *et al.* 2012b). The Illumina OvineSNP50 that included 42,469 SNPs was recently developed by the International Sheep Genomics Consortium (ISGC). In dog, Illumina developed the CanineSNP20 BeadChip with 20K SNPs, and the recent CanineHD BeadChip containing over 170,000 SNPs (MOGENSEN *et al.* 2011). A 60K chicken SNP array powered by Illumina iSelect BeadChip was designed to contain 57,636 SNPs (GROENEN *et al.* 2011). More recently, a high density 600K chicken SNP array was developed with Affymetrix Axiom array technology (KRANIS *et al.* 2013). Apparently, there are no technology barriers for the development of high density SNP arrays. However, the economic challenge is still tremendous because even though the unit cost per genotype is currently extremely low, the total costs for the high density SNP arrays with very high numbers of SNPs can be still beyond the economic powers of researchers working with a "minor" species.

No high density SNP arrays have been developed for aquaculture species. The Illumina GoldenGate Assay was used to evaluate 384 rainbow trout SNPs, resulting in a validation rate of 48% for the tested SNPs (SANCHEZ *et al.* 2009). The GoldenGate Assay was also used to genotype 384 catfish EST-derived SNPs to assess the factors affecting SNP validation rates (WANG *et al.* 2008). The Sequenom MassArray platform was used in several studies in the context of QTL analysis, linkage mapping and map integration. A custom Illumina iSelect SNP array containing approximately 6K SNP markers from Atlantic salmon was developed and used for linkage mapping and QTL analysis (GUTIERREZ *et al.* 2012; LIEN *et al.* 2011).

In catfish, over two million gene-associated SNPs were identified in channel catfish and blue catfish, respectively, using the next-generation sequencing (LIU *et al.* 2011). In a recent study, over eight million SNPs were identified in channel catfish by whole genome sequencing of one wild and four aquaculture populations (SUN *et al.*, in review). With the availability of these SNP resources, here we report the development and performance evaluation of the 250K catfish SNP array using the Affymetrix Axiom genotyping technology.

Materials and Methods

SNP selection and array design

Gene-associated SNPs were generated from Liu et al (LIU *et al.* 2011). Anonymous SNPs from non-coding genomic regions were from SUN *et al.* (in review). SNPs were filtered following the specific requirements by Affymetrix SNP array design. Flanking sequences of 35 bp for each SNP were required where no other variations (SNPs and/or Indels) were allowed within 30 bp of SNPs. The balanced A/T/G/C of flanking sequences was required with GC content of 30%-70%. No repetitive elements were allowed in flanking sequences, in addition that single simple repeats of G or C greater than 4 and A or T greater than 6 were not allowed.

All catfish SNPs passed the in-house filter were submitted to Affymetrix for design score assessment, where each of the two probes flanking SNP was assigned with a p-convert value, respectively. SNPs with probes of high p-convert values were highly likely convertible. A p-convert value threshold was determined by excluding the tail of lowest performing probes to facilitate selection of final SNP list. For the SNPs with both probes pass the p-convert value threshold, one of the probes with the greater value was selected. For the SNPs with both probes having low p-convert values, both probes were selected to ensure conversion with a high probability.

To select well-spaced SNPs, at least one but no more than two SNPs per transcript contig were selected for gene-associated SNPs. For anonymous SNPs, the preliminary catfish genome assembly was used (255,858 contigs with mean length of 2,996 bp and N50 of 6,027 bp, unpublished data) to facilitate SNP selection according to contig length. One SNP per contig was selected from the contigs with length less than 4 kb. Two SNPs per contig were selected from the contigs with length greater than 4 kb. For the two SNPs selected from one contig, the SNPs with largest distance were selected to ensure the good spacing in the genome.

In addition, A/T and C/G SNPs were not selected because the two alleles of these SNPs match the same dye and additional distinct probes in different physical locations on the array are required to distinguish them. Non-polymorphic 31-mers were randomly picked from non-repetitive regions of the genome for data quality control (QC) probes. The QC probes along with final list of SNPs were submitted to Affymetrix for production with Axiom genotyping array.

Assessment of SNP spacing

To assess the genome distribution of SNPs on the array, all the 250,113 SNPs with 35-bp up- and down-stream flanking sequences (71 bp in total) were aligned with the latest draft genome assembly (62,461 scaffolds with N50 of 3Mb, unpublished data) using BLAST to determine SNP positions. The inter-SNP spacing was determined based on SNP positions in the scaffolds. The intervals for SNPs at the ends of sequences were not able to determine. Similarly, SNPs with flanking sequences were aligned with currently available catfish BAC end sequences (BES) to identify SNPs associated with BES.

SNP array performance evaluation

Fish sources

Three different sample sources were used for genotyping to assess the SNP array performance: 1) 192 unrelated channel catfish from wild populations; 2) 192 catfish hybrids from 1st generation of backcrossing and 3) 192 catfish hybrids from 3rd generation of backcrossing. Samples from wild populations were channel catfish collected for a previously study (SIMMONS *et al.* 2006). The 1st generation of backcrossing hybrids was produced by backcrossing the interspecific F_1 hybrids (channel x blue) with a male channel catfish, and the 3rd generation of backcrossing hybrids was produced by backcrossing the 2nd generation of backcross hybrids with a male channel catfish.

DNA isolation

Blood samples (300-500 μl) were collected in a 1-ml syringe and immediately expelled into a 15-ml tube containing 5 ml of DNA extraction buffer (100 mM NaCl, 10 mM Tris, pH 8, 25 mM EDTA, 0.5% SDS, and 0.1 mg/ml freshly made proteinase K) for DNA isolation. DNA was isolated as previously described (KUCUKTAS *et al.* 2009; NINWICHIAN *et al.* 2012). Picogreen dye (Quant-iT Pico Green, Invitrogen) were used in order to quantify double-stranded DNA according to the manufacturer's protocol. The integrity of DNA samples were checked by 1% agarose gel electrophoresis stained with ethidium bromide.

SNP genotyping

Genomic DNA samples were arranged in a 96-well microtiter plate, and normalized to a final concentration of 50 ng/μl with the final volume of 10 μl. The DNA samples were genotyped by GeneSeek (Lincoln, NE, USA) with the catfish 250K SNP array.

SNP analysis

Raw data in the form of CEL files were imported into the Affymetrix Genotyping Console software (v4.1) for quality control analysis and genotype calling using Axiom GT1 algorithm (Affymetrix). Samples with dish quality control (QC) value of 0.82 or better and call rate >0.97

were considered to have passed. Following genotyping, SNPolisher (Affymetrix), an R package, was used to post-process genotyping results. The package can calculate the QC metrics for each SNP/probeset to determine its quality and classify SNPs into six categories (Figure 1): (i), "PolyHighResolution" where three clusters are formed with good resolution; (ii), "NoMinorHom" where two clusters are formed with no examples of the minor homozygous genotypes; (iii), "MonoHighResolution" in which only one cluster is formed; (iv), "OTV", off-target variants, where three good clusters are formed, but with one extra OTV cluster that is caused by sequence dissimilarity between probes and target genome regions (DIDION *et al.* 2012); (v), "CallRateBelowThreshold" where SNP call rate is below threshold, but other cluster properties are above threshold; and (vi), "Other" where one or more cluster properties are below threshold. In this study, SNPs of categories (i) to (iv) were considered as convertible SNPs, and SNPs of categories (i) and (ii) were considered as polymorphic SNPs.

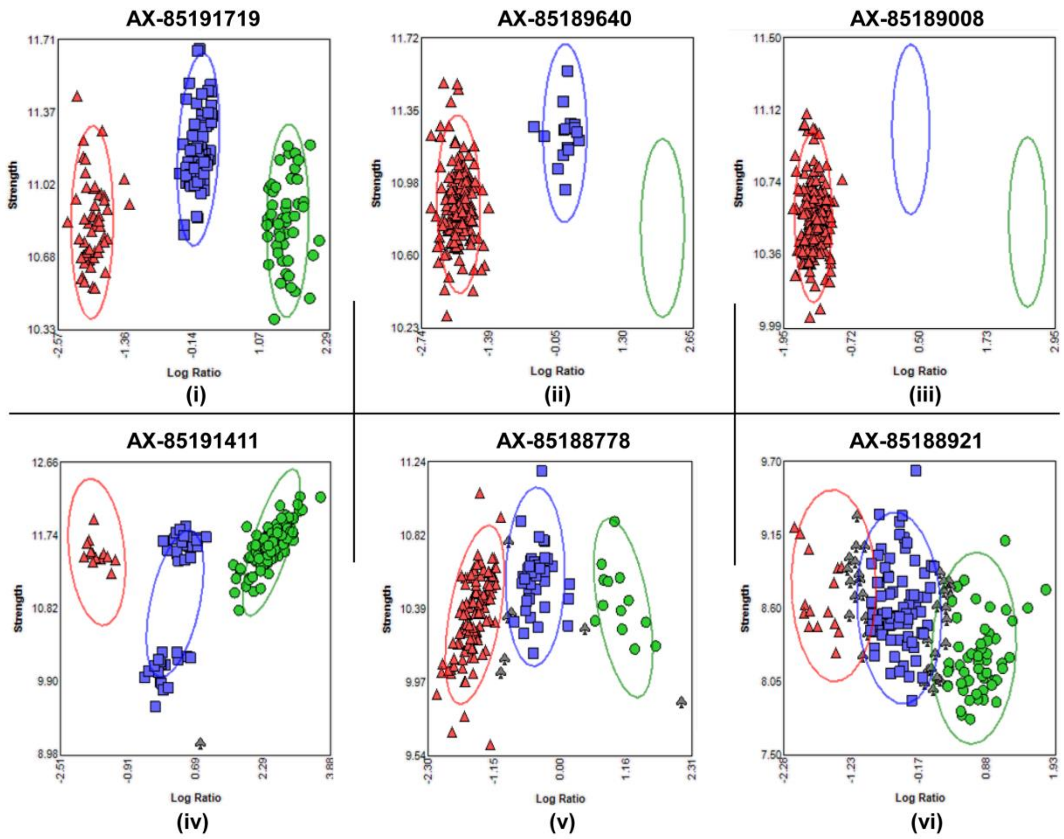

Figure 1. Illustration of six SNP/probeset categories based on cluster properties.

Results and Discussion

Selection of SNPs for the SNP array

One of the most important goals of the SNP array development is to have a good coverage of the genome. The first task was to select a subset of SNPs from all identified SNPs, gene-associated as well as anonymous SNPs. From a large pool of the previously identified SNPs (LIU *et al.* 2011; SUN *et al.*, in review), the following selection criteria were used for the initial selection of SNPs: 1) For gene associated SNPs, at least one but no more than two SNPs per transcript contig were selected; 2) For anonymous SNPs, one SNP was selected for small contigs of less than 4 kb, but two SNPs were selected for contigs larger than 4 kb, with their spacing being the largest within the contig. In addition, a set of sequence features were also considered for the selection of the initial SNPs (see Materials and Methods).

Table 1. Summary of SNPs used for the catfish 250K SNP array design.

SNP sources	SNPs selected for array	SNPs passed Affymetrix evaluation	SNPs included on the array
Gene-associated SNPs			
Channel catfish	93,699	72,202	32,188 (12.9%)
Blue catfish	59,464	48,900	31,392 (12.6%)
Inter-specific	83,549	72,260	39,605 (15.8%)
Anonymous SNPs			
Channel catfish	404,777	302,309	146,928 (58.7%)
Total SNPs	641,489	495,671	250,113 (100%)

The initial in house selection resulted in 641,489 SNPs that were submitted to Affymetrix for *in-silico* analysis to assess the predicted performance on Axiom arrays. Both forward and reverse probe of each SNP were assigned with p-convert values, which were derived from a random forest model to predict the probability that the SNP will convert on the array. The model considers factors including probe sequence, binding energy and the expected degree of non-specific hybridization to multiple genomic regions (personal communication with Affymetrix). SNP probes with high p-convert values are expected to convert on the SNP array with high probability.

A total of 495,671 SNPs passed the Affymetrix *in-silico* evaluation with various p-convert values, but the vast majority of SNPs had p-convert values greater than 0.5 (Figure 2). Because many more than needed SNPs passed the p-convert evaluation, only SNPs with p-convert values greater than 0.5 were further considered for inclusion on the array. For the SNPs with both probes passing the p-convert threshold level, the probes with the higher p-covert values were selected. For the SNPs with both probes having relatively low p-convert values, both probes were selected to increase the conversion rate for the SNP. In the final SNP list, A/T and C/G SNPs were removed because such SNPs require twice the number of probes.

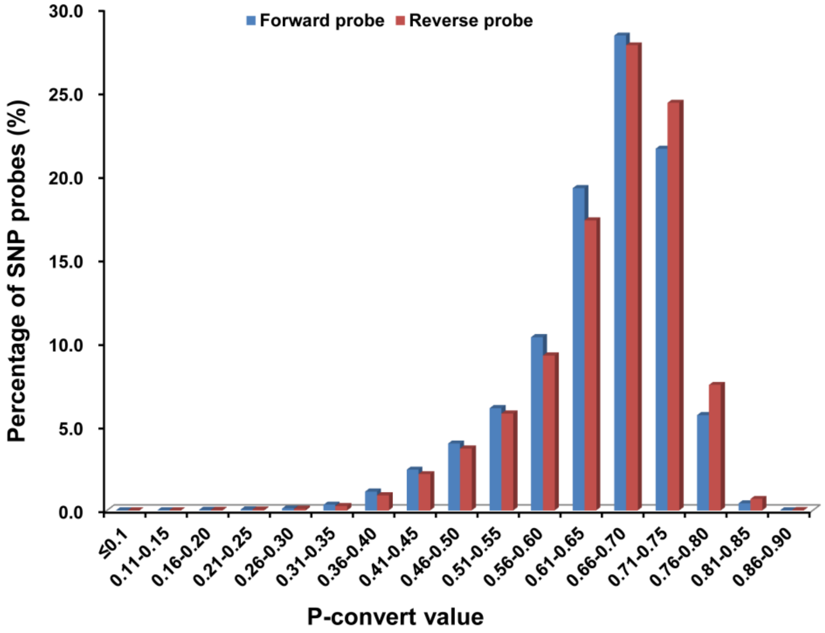

Figure 2. Distribution of SNP probes based on p-convert values.

SNPs included on the 250K array

The SNPs used for the development of the catfish 250K SNP array are summarized in Table 1. A total of 250,113 SNPs were included in the 250K array including 103,185 (41.3%) gene-associated SNPs and 146,928 (58.7%) anonymous SNPs. Of the gene-associated SNPs, 32,188 were from SNPs identified from channel catfish, 31,392 were from SNPs identified from blue catfish, and 39,605 were interspecific SNPs identified between channel catfish and blue catfish (Table 1).

A total of 316,706 SNP probes were synthesized for interrogation of these 250,113 SNPs with 66,593 SNPs of which tiled with two probes (Table 2). In addition to SNP probes, 2,000 data quality control (QC) probes were included on the SNP array serving as negative controls.

The QC probes are non-polymorphic 31-mers randomly picked from non-repetitive regions of catfish genome. Of which, 1,000 QC probes were selected with A or T at the 31st base, and 1,000 QC probes were selected with G or C at the 31st base.

Table 2. Summary of the catfish 250K SNP array.

SNP array	Number
Total number of SNPs	250,113
Number of SNPs tiled with single probe	183,520
Number of SNPs tiled with two probes	66,593
Total number of probes	316,706
Number of data quality control probes	2,000

Inclusion of gene-associated SNPs should enhance the conversion rate because genes and their associated sequences are more unique in the genome than the non-coding genomic sequences. In addition, as genes are broadly distributed across all 29 pairs of chromosomes of the catfish genome (LIU *et al.* 2011), SNPs derived from genes should reflect the distribution of genes within the genome. However, as genes are not entirely evenly distributed, inter-marker spacing is not equal. Genomic SNPs from anonymous regions would fill the gaps. A subset of gene-associated SNPs are interspecific SNPs that are useful for genetic analysis of the interspecific hybrid system.

Distribution of SNP spacing

It was unable to determine the absolute SNP coordinates and thereby their distribution in the genome because a fully assembled catfish genome is not yet available. After the completion of SNP array development, a draft catfish genome assembly was generated. To assess the SNP distribution, the inter-SNP spacing was evaluated using this draft genome assembly. A total of 248,308 SNPs (99.3%) with flanking sequences were uniquely mapped to 11,017 genome scaffolds which span a total of 785.6 Mb, approximately 80% of the genome. As shown in

Figure 3, a total of 237,291 SNPs with inter-SNP spacing were examined. Of which, 49,718 SNPs had a small inter-SNP spacing of less than 500 bp, and 31,811 SNPs had an inter-SNP spacing of 500-1000 bp. The largest number of SNPs (46,804) had an inter-SNP spacing of 1000-2000 bp. A total of 31,184 had a marker spacing of 2000-3000 bp, 21,100 had a marker spacing of 3000-4000 bp, 14,538 had a marker spacing of 4000-5000 bp, 10,316 had a marker spacing of 5000-6000 bp, and 31,820 had a marker spacing of more than 6000 bp. Cumulatively, approximately 34.4% SNPs had a marker spacing of less than 1 kb, and approximately 13.4% had a marker spacing of greater than 6 kb. Approximately 50% SNPs had marker spacing of 1-6 kb (Figure 3).

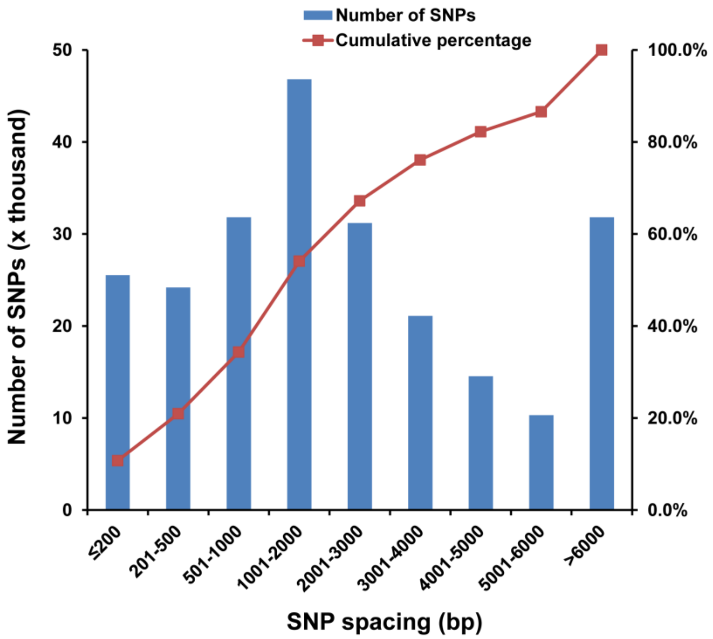

Figure 3. Distribution of inter-SNP spacing for SNPs included on the array. SNP intervals were determined based on current catfish assembly.

Due to the lack of fully assembled genome sequence, the inter-marker spacing is probably underestimated. The current draft genome assembly had a total genome size of 830.5 Mb, but the catfish genome is approximately 950 Mb. In addition, the spacing from the most external SNPs of each scaffold to the next marker is not included in the assessment. Therefore, the overall marker spacing could have been slightly larger.

The SNP distribution was also evaluated with regard to association with bacterial artificial chromosome (BAC) end sequences (BES). A total of 22,051 SNPs on the array are associated with 16,832 BES derived from 14,849 BAC clones. Accordingly, such SNPs are associated with 2,853 (86.3% of 3,307) contigs from the catfish physical map developed by Xu et al (XU *et al.* 2007), 1,141 of which were not able to integrate with linkage map constructed mainly using microsatellite markers (NINWICHIAN *et al.* 2012). Such BAC associated SNPs are useful because they are separated by a known distance in the genome (BAC insert size of 161 Kb on average (XU *et al.* 2007)), and their use can facilitate full integration of genetic linkage and physical maps.

Performance of the catfish 250K SNP array

Genotyping performance of the SNP array

Performance of the SNP array was examined by genotyping unrelated catfish samples from wild populations and catfish backcross families. As summarized in Table 3, 182 of 192 wild catfish samples were successfully genotyped after sample quality control. A total of 204,437 SNPs (81.7%) were converted, and 137,459 (55.0%) of which were polymorphic in these samples. The average call rate of converted SNPs was greater than 99.4%.

The catfish 250K SNP array was also examined by genotyping in catfish backcross hybrids: the 1st generation of backcross produced by crossing female hybrid with male channel catfish (BC1) and the 3rd generation of backcross produced by crossing 2nd generation of backcross female hybrid with male channel catfish (BC3). As summarized in Table 3, percentages of SNPs converted in BC1 catfish were relatively lower than that in unrelated wild catfish, while much higher percentages of SNPs were converted in BC3 catfish than in BC1 catfish as well as in unrelated wild catfish (Table 3). The BC3 catfish, in theory, possess higher fraction of "channel catfish" genome materials than BC1 catfish due to two more generations of backcrossing with channel catfish. Therefore, higher proportions of intra-specific SNPs from channel catfish were expected to convert in BC3 catfish than in BC1 catfish. Accordingly, higher proportions of inter-specific SNPs were expected to convert in BC1 than in BC3 and wild catfish.

Table 3. Performance assessment of the catfish 250K SNP array. *BC1 denotes the catfish from 1st generation of backcross, and BC3 denotes the catfish from 3rd generation of backcross. **SNPs on the array that work ***Average SNP call rate, for given converted SNPs, the average percentage of samples whose genotypes were successfully measured.

Samples*	**Samples processed**	**Samples passed**	**SNPs converted****	**Polymorphic SNPs**	**Avg. SNP call rate*****
Wild catfish	192	182 (94.8%)	204,437 (81.7%)	137,459 (55.0%)	99.4%
BC1	192	179 (93.2%)	198,583 (79.4%)	130,685 (52.3%)	99.7%
BC3	192	192 (100%)	218,440 (87.3%)	156,357 (62.5%)	99.8%

Comparisons of SNP polymorphic rates among wild catfish, BC1 and BC3 catfish indicated that a large number of SNPs (70,599) were polymorphic in all examined groups of fish (Figure 4A). A number of 57,766, 66,282 and 60,459 SNPs were non-polymorphic observed in wild catfish, BC1 and BC2 catfish (Figure 4B). These SNPs were most likely pseudo-SNPs or they were polymorphic in the fish where they were identified, but they were not polymorphic among these tested fish. The comparisons of non-polymorphic SNPs resulted in a total of 22,714 SNPs that are monomorphic in all examined samples, suggesting that the remainders of SNPs are most likely true SNPs that are not polymorphic in present tested samples.

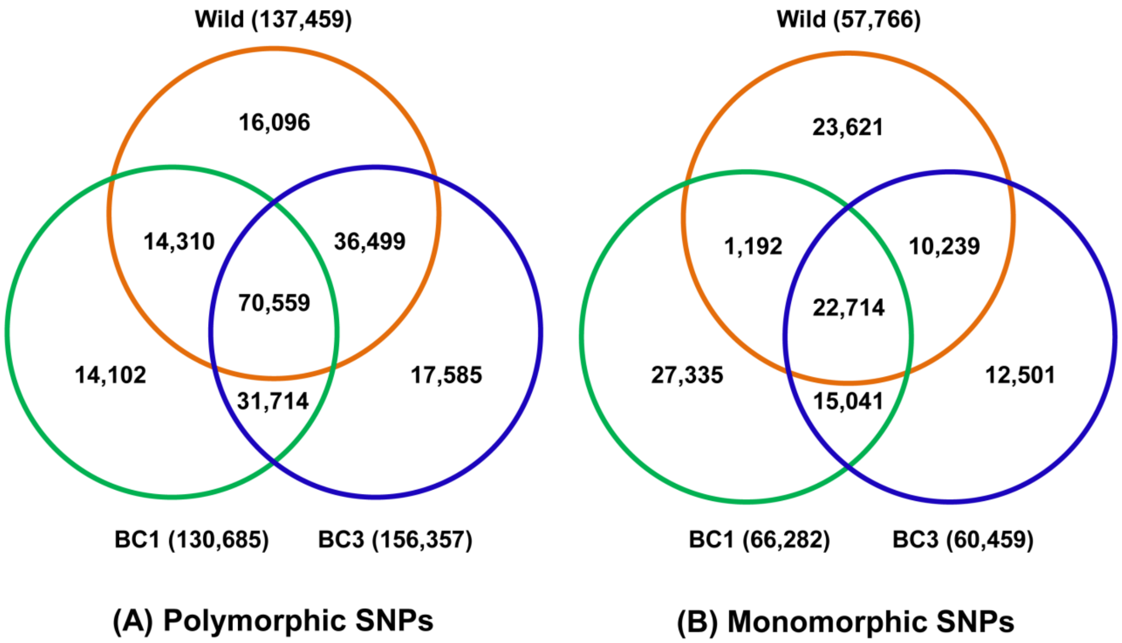

Figure 4. Comparisons of polymorphic SNPs and monomorphic SNPs. (A) Polymorphic SNPs, (B) Monomorphic SNPs. Wild, unrelated wild channel catfish, BC1, 1st generation of backcross progeny, and BC3, 3rd generation of backcross progeny.

The hybrid catfish produced by crossing female channel catfish and male blue catfish possess superior performance traits, therefore, the catfish hybrid system is not only important to the catfish industry, but also interesting for the genetics studies to dissect genetic basis underlying performance traits. The high percentages of converted SNPs and large numbers of polymorphic SNPs achieved in genotyping fish from backcross families as well as wild populations suggested the good performance of this SNP array.

Assessment of informativeness of SNPs on the array

Marker informativeness is reflected in minor allele frequencies as SNPs are most often bi-allelic markers. In order to assess the informativeness of the SNPs on the array, the minor allele frequencies of the SNPs were determined in wild population and two families of first (BC1) and third (BC3) backcross progenies. The genotypes of 130,685, 156,357 and 137,459 polymorphic SNPs in BC1, BC3 and Wild catfish samples were used for the analysis. Distribution of minor allele frequencies with intervals of 0.1 was shown in Figure 5. Overall, most polymorphic SNPs had a MAF of greater than 0.1, with 36,011 between 0.1-0.2, 28,628 between 0.2-0.3, 19,340 between 0.3-0.4, and 16,716 between 0.4-0.5 in wild catfish samples. A notable fraction (23%) were non-polymorphic, suggesting these SNPs were most likely pseudo SNPs or they were polymorphic in the fish where they were identified, but they were not polymorphic among these tested fish. Because a large number of wild fish (192) was tested and they are not genetically related, it is more likely that these SNPs represented pseudo SNPs. However, from the practical perspective, the catfish 250K SNP array will be quite informative as over 50% were polymorphic even within two families of backcrosses.

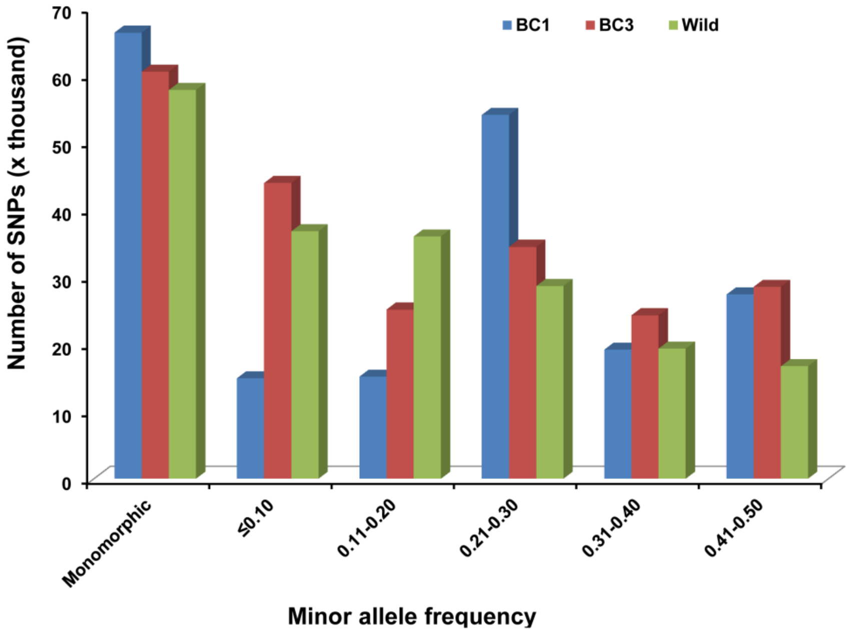

Figure 5. Distribution of minor allele frequencies in three catfish groups.

The minor allele frequency (MAF) directly relates to the polymorphism information content that measures marker informativeness (PETKOV *et al.* 2004). Obviously, the higher MAF the SNP has, the greater informative it will be. However, SNPs with low MAFs (rarer variants) are possibly with larger effects therefore are essential in genome-wide association analysis (MANOLIO *et al.* 2009).

Relationships between design score and SNP performance

As the p-convert values are important factors for selection of SNPs, it is interesting to analyze its relationships with SNP performance. As shown in Figure 6, the p-covert values were positively correlated with the performance of the SNP probes, the higher the p-convert values, the better performance the probes. However, once the p-convert values reached 0.7 or higher, further increase in p-convert values did not have significant effect on probe performance any longer (Figure 6). This relationship holds not only for percent of probes that worked, but also for percent of converted SNPs. The spike in percent of converted SNPs with probes of lower p-convert values is an artifact due to the inclusion of two probes per SNP for SNPs with relatively lower p-convert values (Figure 6). Apparently, the p-convert values served well as a parameter for the prediction of SNP probe performance.

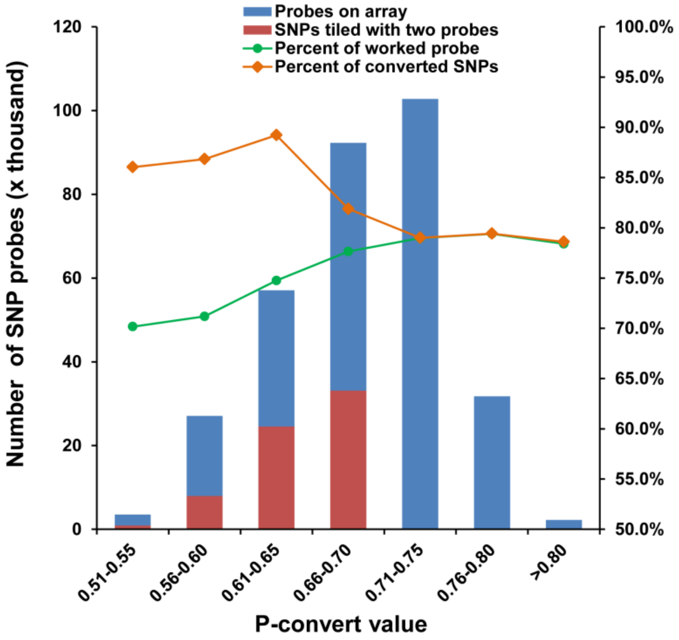

Figure 6. Relationships between Affymetrix design scores and SNP probe performance.

Performance of gene-associated SNPs versus anonymous SNPs

On the catfish 250K SNP array, 179,116 SNPs were identified from channel catfish, of which 146,928 were anonymous SNPs while 32,188 were gene-associated SNPs. To compare the performance gene-associated SNPs and anonymous SNPs, the conversion rates and percentages of polymorphic SNPs were analyzed. As shown in Figure 7, there is no significant difference in performance between gene-associated SNPs and anonymous SNPs, while the conversion rates and polymorphic SNP percentages of gene-associated SNPs are slightly higher, by a couple of percentage, than those of anonymous SNPs in all three examined populations.

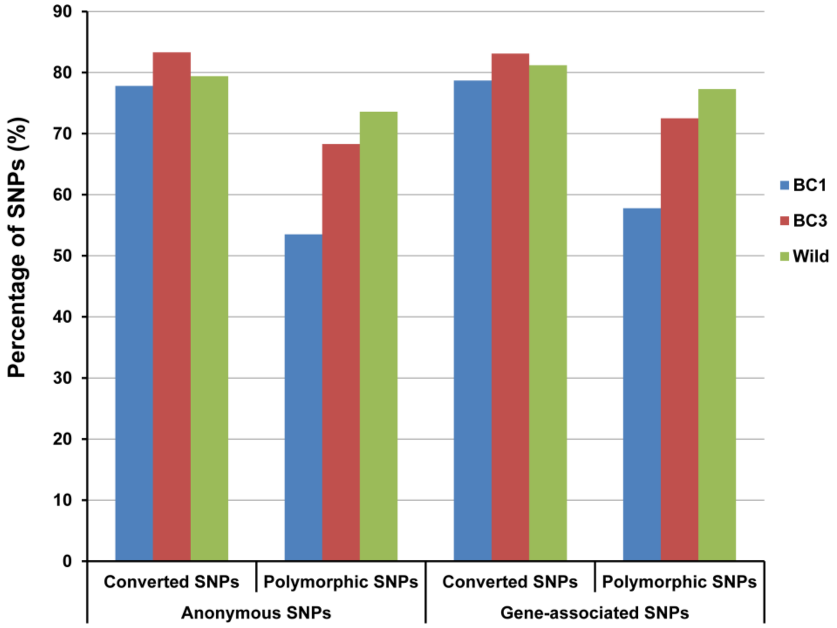

Figure 7. Performance between gene-associated SNPs and anonymous SNPs.

Performance of intra-specific and inter-specific SNPs

The performances of 32,188 intra-specific SNPs in channel catfish, 31,392 intra-specific SNPs in blue catfish and 39,605 inter-specific SNPs between the two species were examined as shown in Figure 8. As expected, the highest percentage of polymorphic SNPs was converted from SNPs in the channel catfish when being genotyped in channel catfish samples from the wild population. In contrast, the intra-specific SNPs identified from blue catfish had a very low polymorphic rate in wild channel catfish population. Similarly, only 8% inter-specific SNPs were polymorphic among wild channel catfish. However, such inter-specific SNPs performed really well in interspecific backcross families, as expected (Figure 8).

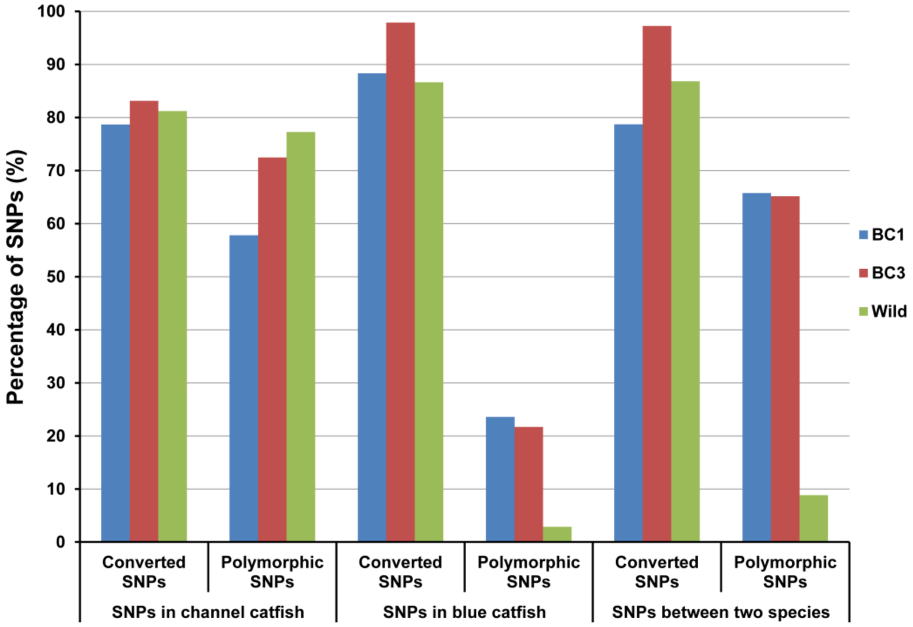

Figure 8. Performance between intra-specific SNPs and inter-specific SNPs.

Performance of transition and transversion SNPs

Of the 250,113 SNPs on the array, 75.9% are transitions and 24.1% are transversions. Transition SNPs are more abundant than transversion SNPs, with an estimated ratio of 1.8-1.9 in catfish among gene-associated SNPs (LIU *et al.* 2011; WANG *et al.* 2010). The exclusion of A/T and G/C SNPs in design stage of SNP array reduced the fraction of transversion SNPs. It's interesting to examine the performance of these two types of SNPs. As shown in Figure 9, the two types of SNPs had nearly identical conversion rates and polymorphic rates, suggesting that transition and transversion SNPs made no difference in their performance for genotyping.

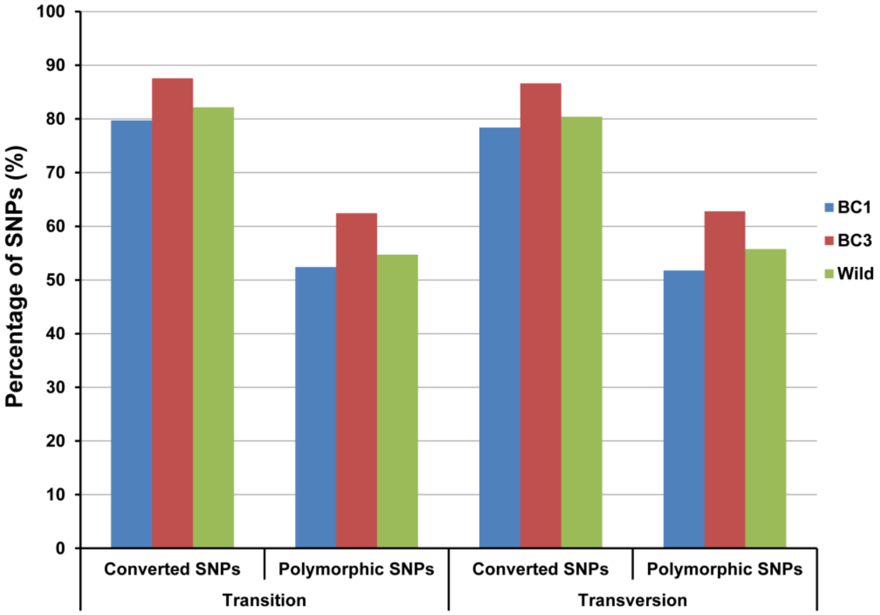

Figure 9. Performance between transition SNPs and transversion SNPs.

<u>SNP Transferability to other related catfish species</u>

To assess the utility of the catfish 250K SNP array in the related catfish taxa, a set of DNA samples were tested from blue catfish (*Ictalurus furcatus*), brown bullhead catfish (*Ameiurus nebulosus*) and white catfish (*A. catus*). As summarized in Table 4, overall SNP conversion rate across these species were actually quite high, with a minimal rate of 50.4% with bullhead catfish and as high as 77.2% with D&B strain of blue catfish. However, the polymorphic rates were much lower, more than 10 times lower than the conversion rates in most cases. For instance, the polymorphic rates of all SNPs on the array had a polymorphic rate of 7.8% and 3.9% with Rio Grande and D&B strain of blue catfish, respectively, as compared to 43.5% and 18.7% polymorphic rates of SNPs identified from blue catfish when tested in the same two strains. The SNPs on the array had low polymorphic rates for bullhead catfish and white catfish as well, ranging from 4.4-5.1%. Taken together, these results suggested that the probes designed from channel catfish and blue catfish sequences could actually hybridize to the genomic DNA of bullhead catfish and whitre catfish, but the bases at the SNP sites were not polymorphic with bullhead catfish and white catfish.

In spite of the low polymorphic rates, the number of SNPs that were polymorphic was still notable with various species of catfishes. For instance, over 12,000 SNPs were polymorphic for bullhead catfish and white catfish, suggesting its applicability for genetic analysis of related catfish taxa. The polymorphic rates estimated here are probably an underestimate because only 10 individuals were genotyped. Polymorphic rates would increase if more fish had been genotyped. Although these estimate are quite preliminary, the very low polymorphic rates of D&B strain of blue catfish suggested that this strain could have been inbred and may have a small number of founders, at least twice less diverse than the Rio Grande strain (Table 4).

Table 4. Transferability of SNPs to other catfish species. *All 250,113 SNPs on the array, **SNPs from blue catfish (31,392).

Species	**Latin name**	**Converted SNPs***	**Polymorphic SNPs***	**Converted SNPs****	**Polymorphic SNPs****
Blue catfish (Rio Grande)	*Ictalurus furcatus*	190,867 (76.3%)	19,549 (7.8%)	25,722 (81.9%)	13,667 (43.5%)
Blue catfish (D&B)	*I. furcatus*	193,039 (77.2%)	9,684 (3.9%)	25,109 (80.0%)	5,859 (18.7%)
Bullhead catfish	*Ameiurus nebulosus*	126,076 (50.4%)	12,649 (5.1%)	17,739 (56.5%)	1,376 (4.4%)
White catfish	*A. catus*	129,716 (51.9%)	12,833 (5.1%)	18,286 (58.3%)	1,452 (4.6%)

Conclusion

In this study, we report the development of the catfish 250K SNP array using Affymetrix Axiom technology. The SNPs were well-spaced in the genome. Distribution of minor allele frequency indicated that uniform SNPs were included on the array. The evaluation of SNP array performance by genotyping samples from pedigree families and unrelated wild populations suggested high SNP conversion rates (~80%) and high polymorphic rates (over 50%) can be obtained in all the examined samples. Technically, we showed that the Affymetrix design score (p-convert value) adequately predict SNP probe performance and the inclusion of alternative probes greatly increased the SNP conversion rates, especially for SNPs with probes that had low design scores. The catfish 250K SNP array should be valuable in both industry and research such as in whole genome-based selection, genome-wide association studies, fine QTL mapping, high-density linkage mapping, and haplotype analysis.

Acknowledgements

This project was supported by Agriculture and Food Research Initiative Competitive Grant no. 2009-35205-05101, 2010-65205-20356 and 2012-67015-19410 from the USDA National Institute of Food and Agriculture (NIFA). Luyang Sun, Yun Li, Fanyue Sun, Yanliang Jiang, Yu Zhang, Jiaren Zhang, Jianbin Feng, Ludmilla Kaltenboeck, Huseyin Kucuktas and Zhanjiang Liu are coauthors on the manuscript. The authors are grateful for the support from Affymetrix Bioinformatics Services and custom genotyping program, especially Lucy Reynolds and Ali Pirani. Thanks go to the center for Applied Aquatic Genomics, Chinese Academy of Fishery science for the collaboration with development of SNP array, specifically Jian Xu and Peng Xu.

We thank Dr. Geoff Waldbieser for the collaboration to generate the preliminary catfish whole genome assembly. We thank all the lab colleagues who provided assistance with sample collection.

Chapter 5

Conclusion

The development of catfish breeds with superior performance traits such as high level of resistance to major diseases and fast-growth would represent a significant stimulus to the catfish industry in the United States. Progress toward this goal will be made more rapidly as genomic tools and resources are developed and implemented in catfish breeding. These tools will automate and simplify once overwhelming tasks and enable the dissection of genetic basis underlying phenotypic variations. High density SNP array is one of the most important tools. Future work warrants the effectiveness of utilization of this tool for genome research. In fact, its application in genome-wide association studies for complex traits has been presented technically feasible in several livestock animals. It's expected that the identification of genome level trait-marker associations will be achievable in the very near future to enable the whole genome-based selection in catfish for genetic improvements.

In this study, with the ultimate goal to develop the catfish SNP array for genome-wide association studies, and towards genomic selection in catfish. We report the generation of genome-scale SNP markers, the selection of representative SNPs and the development of catfish 250K SNP array. To facilitate the selection of representative SNPs from genes, we generated a comprehensive catfish transcriptome assembly by RNA-Seq of a doubled haploid channel catfish with deep sequencing depth. The SNP identification and transcriptome assembly were achieved *de novo* with the next-generation sequencing technologies in an efficient and economical way. Well-spaced SNPs in the genome were selected to enable the genome coverage. The Affymetrix design score assigned to probe sets were also evaluated when finalizing SNPs to include on the catfish SNP array. A total of 250,113 well-spaced SNPs were used for developing the catfish 250K SNP array. The performance of the catfish SNP array was assessed by genotyping samples from pedigree families and unrelated wild populations. The results indicated that high SNP conversion rates (~80%) and polymorphic rates (over 50%) were achieved. Distribution of minor allele frequency indicates that SNPs with uniform allele frequencies are included on the array, indicating the good utility for genome-wide association studies. The catfish 250K SNP array should be valuable in both industry and research such as in whole genome-based selection, genome-wide association studies, fine QTL mapping, high-density linkage mapping, and haplotype analysis

With this SNP array, numerous GWAS for production and performance traits in catfish are readily feasible to be conducted. As a pilot study, a GWAS study is being on the schedule for deciphering the genetic basis of ESC disease resistance in catfish. Catfish showing susceptibility and resistance to ESC disease will be collected for genotyping. Samples from both pedigree families and unrelated wild populations will be used for the cross-validation. It's anticipated that significant SNPs associated with genomic regions responsible for disease resistance can be identified and can be timely used in assisting selection for breeding superior catfish brood stocks.

Besides its implementation in various GWAS for other major diseases and for economically important traits such as growth rate, feed conversion efficiency, low oxygen level tolerance and high temperature tolerance, the catfish 250K SNP array can be utilized for other genomics and genetics analysis. A high-density linkage map can be constructed by genotyping multiple mapping resource families each with hundreds of individuals with the SNP array. The resulted linkage map should have much higher resolution than any previous catfish linkage map. Most importantly, this should be a more accurate map due to the use of multiple families and much larger number of individuals providing stronger statistical power. The high-resolution, accurate map would further allow for the fine QTL mapping and the integration with the whole genome scaffolds to build up the chromosomal level genome assembly. In addition, linkage disequilibrium analysis can be conducted for detection of haplotype blocks as well. Tag SNP from each block can be selected to represent the haplotype block because SNPs located in same haplotype block are generally not segregated. With Tag SNPs, a small-scale SNP array can be designed which will be flexible and useful for genome analysis.

References

ABASHT, B., and S. J. LAMONT, 2007 Genome-wide association analysis reveals cryptic alleles as an important factor in heterosis for fatness in chicken F-2 population. Animal Genetics **38:** 491-498.

ADAMIDI, C., Y. WANG, D. GRUEN, G. MASTROBUONI, X. YOU *et al.*, 2011 *De novo* assembly and validation of planaria transcriptome by massive parallel sequencing and shotgun proteomics. Genome Research **21:** 1193-1200.

AHMAD, R., D. E. PARFITT, J. FASS, E. OGUNDIWIN, A. DHINGRA *et al.*, 2011 Whole genome sequencing of peach (*Prunus persica* L.) for SNP identification and selection. BMC Genomics **12:** 569.

ALTSHULER, D., V. J. POLLARA, C. R. COWLES, W. J. VAN ETTEN, J. BALDWIN *et al.*, 2000 An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature **407:** 513-516.

ANDREASSEN, R., S. LUNNER and B. HOYHEIM, 2009 Characterization of full-length sequenced cDNA inserts (FLIcs) from Atlantic salmon (*Salmo salar*). BMC Genomics **10:** 502.

ANTTILA, V., H. STEFANSSON, M. KALLELA, U. TODT, G. M. TERWINDT *et al.*, 2010 Genome-wide association study of migraine implicates a common susceptibility variant on 8q22.1. Nature Genetics **42:** 869-873.

ARGUE, B. J., Z. J. LIU and R. A. DUNHAM, 2003 Dress-out and fillet yields of channel catfish, *Ictalurus punctatus*, blue catfish, *Ictalurus furcatus*, and their F-1, F-2 and backcross hybrids. Aquaculture **228:** 81-90.

ASLAM, M. L., J. W. BASTIAANSEN, M. G. ELFERINK, H. J. MEGENS, R. P. CROOIJMANS *et al.*, 2012 Whole genome SNP discovery and analysis of genetic diversity in Turkey (*Meleagris gallopavo*). BMC Genomics **13:** 391.

AWANO, T., G. S. JOHNSON, C. M. WADE, M. L. KATZ, G. C. JOHNSON *et al.*, 2009 Genome-wide association analysis reveals a SOD1 mutation in canine degenerative myelopathy that resembles amyotrophic lateral sclerosis. Proceedings of the National Academy of Sciences of the United States of America **106:** 2794-2799.

BARRETT, J. C., J. C. LEE, C. W. LEES, N. J. PRESCOTT, C. A. ANDERSON *et al.*, 2009 Genome-wide association study of ulcerative colitis identifies three new susceptibility loci, including the HNF4A region. Nature Genetics **41:** 1330-U1399.

BEAUDOING, E., S. FREIER, J. R. WYATT, J. M. CLAVERIE and D. GAUTHERET, 2000 Patterns of variant polyadenylation signal usage in human genes. Genome Research **10:** 1001-1010.

BECKER, D., J. TETENS, A. BRUNNER, D. BURSTEL, M. GANTER *et al.*, 2010 Microphthalmia in Texel Sheep Is Associated with a Missense Mutation in the Paired-Like Homeodomain 3 (PITX3) Gene. PLoS ONE **5:** e8689.

BECKER, D., K. WIMMERS, H. LUTHER, A. HOFER and T. LEEB, 2013 A Genome-Wide Association Study to Detect QTL for Commercially Important Traits in Swiss Large White Boars. PLoS ONE **8:** e55951.

BJORNSTEDT, M., S. KUMAR, L. BJORKHEM, G. SPYROU and A. HOLMGREN, 1997 Selenium and the thioredoxin and glutaredoxin systems. Biomed Environ Sci **10:** 271-279.

BOLORMAA, S., B. J. HAYES, K. SAVIN, R. HAWKEN, W. BARENDSE *et al.*, 2011a Genome-wide association studies for feedlot and growth traits in cattle. Journal of Animal Science **89:** 1684-1697.

BOLORMAA, S., L. R. NETO, Y. D. ZHANG, R. J. BUNCH, B. E. HARRISON *et al.*, 2011b A genome-wide association study of meat and carcass traits in Australian cattle. Journal of Animal Science **89:** 2297-2309.

BONETTI, B., L. FU, J. MOON and D. M. BEDWELL, 1995 The efficiency of translation termination is determined by a synergistic interplay between upstream and downstream sequences in Saccharomyces cerevisiae. J Mol Biol **251:** 334-345.

BOULDING, E. G., M. CULLING, B. GLEBE, P. R. BERG, S. LIEN *et al.*, 2008 Conservation genomics of Atlantic salmon: SNPs associated with QTLs for adaptive traits in parr from four trans-Atlantic backcrosses. Heredity **101:** 381-391.

BROOKS, S. A., N. GABRESKI, D. MILLER, A. BRISBIN, H. E. BROWN *et al.*, 2010 Whole-Genome SNP Association in the Horse: Identification of a Deletion in Myosin Va Responsible for Lavender Foal Syndrome. Plos Genetics **6:** e1000909.

BRUNO, V. M., Z. WANG, S. L. MARJANI, G. M. EUSKIRCHEN, J. MARTIN *et al.*, 2010 Comprehensive annotation of the transcriptome of the human fungal pathogen Candida albicans using RNA-seq. Genome Research **20:** 1451-1458.

BURTON, P. R., D. G. CLAYTON, L. R. CARDON, N. CRADDOCK, P. DELOUKAS *et al.*, 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature **447:** 661-678.

CAI, Q. Y., J. R. LONG, W. LU, S. M. A. QU, W. Q. WEN *et al.*, 2011 Genome-wide association study identifies breast cancer risk variant at 10q21.2: results from the Asia Breast Cancer Consortium. Human Molecular Genetics **20:** 4991-4999.

CANOVAS, A., G. RINCON, A. ISLAS-TREJO, S. WICKRAMASINGHE and J. F. MEDRANO, 2010 SNP discovery in the bovine milk transcriptome using RNA-Seq technology. Mamm Genome **21:** 592-598.

CASSAN, M., and J. P. ROUSSET, 2001 UAG readthrough in mammalian cells: effect of upstream and downstream stop codon contexts reveal different signals. BMC Mol Biol **2:** 3.

CAVENER, D. R., and S. C. RAY, 1991 Eukaryotic start and stop translation sites. Nucleic Acids Res **19:** 3185-3192.

CHAMBERS, J. C., W. H. ZHANG, Y. LI, J. SEHMI, M. N. WASS *et al.*, 2009 Genome-wide association study identifies variants in TMPRSS6 associated with hemoglobin levels. Nature Genetics **41:** 1170-1172.

CHEN, F., Y. LEE, Y. L. JIANG, S. L. WANG, E. PEATMAN *et al.*, 2010 Identification and Characterization of Full-Length cDNAs in Channel Catfish (*Ictalurus punctatus*) and Blue Catfish (*Ictalurus furcatus*). PLoS ONE **5:** e11546.

CHIO, A., J. C. SCHYMICK, G. RESTAGNO, S. W. SCHOLZ, F. LOMBARDO *et al.*, 2009 A two-stage genome-wide association study of sporadic amyotrophic lateral sclerosis. Human Molecular Genetics **18:** 1524-1532.

CIRULLI, E. T., A. SINGH, K. V. SHIANNA, D. L. GE, J. P. SMITH *et al.*, 2010 Screening the human exome: a comparison of whole genome and whole transcriptome sequencing. Genome Biology **11:** R57.

CONESA, A., S. GOTZ, J. M. GARCIA-GOMEZ, J. TEROL, M. TALON *et al.*, 2005 Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics **21:** 3674-3676.

CROOKS, G. E., G. HON, J. M. CHANDONIA and S. E. BRENNER, 2004 WebLogo: a sequence logo generator. Genome Research **14:** 1188-1190.

DAVEY, J. W., P. A. HOHENLOHE, P. D. ETTER, J. Q. BOONE, J. M. CATCHEN *et al.*, 2011 Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics **12:** 499-510.

DENOEUD, F., J. M. AURY, C. DA SILVA, B. NOEL, O. ROGIER *et al.*, 2008 Annotating genomes with massive-scale RNA sequencing. Genome Biology **9:** R175.

DIDION, J. P., H. N. YANG, K. SHEPPARD, C. P. FU, L. MCMILLAN *et al.*, 2012 Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias. BMC Genomics **13:** 34.

DU, Z. Q., D. C. CIOBANU, S. K. ONTERU, D. GORBACH, A. J. MILEHAM *et al.*, 2010 A gene-based SNP linkage map for pacific white shrimp, *Litopenaeus vannamei*. Animal Genetics **41:** 286-294.

DU, Z. Q., X. ZHAO, N. VUKASINOVIC, F. RODRIGUEZ, A. C. CLUTTER *et al.*, 2009 Association and Haplotype Analyses of Positional Candidate Genes in Five Genomic Regions Linked to Scrotal Hernia in Commercial Pig Lines. PLoS ONE **4:** e4837.

DUIJVESTEIJN, N., E. F. KNOL, J. W. M. MERKS, R. P. M. A. CROOIJMANS, M. A. M. GROENEN *et al.*, 2010 A genome-wide association study on androstenone levels in pigs reveals a cluster of candidate genes on chromosome 6. BMC Genetics **11:** 42.

DUNHAM, R. A., 2007 Comparison of six generations of selection, interspecific hybridization, intraspecific crossbreeding and gene transfer for growth improvement in ictalurid catfish. Aquaculture **272:** S252-S253.

DUNHAM, R. A., and B. J. ARGUE, 1998 Seinability of channel catfish, blue catfish, and their F-1, F-2, F-3, and backcross hybrids in earthen ponds. Progressive Fish-Culturist **60:** 214-220.

DUNHAM, R. A., and B. J. ARGUE, 2000 Reproduction among channel catfish, blue catfish, and their F-1 and F-2 hybrids. Transactions of the American Fisheries Society **129:** 222-231.

DUNHAM, R. A., A. N. BART and H. KUCUKTAS, 1999 Effects of fertilization method and of selection for body weight and species on fertilization efficiency of channel catfish eggs with blue or channel catfish sperm. North American Journal of Aquaculture **61:** 156-161.

DUNHAM, R. A., R. E. BRUMMETT, M. O. ELLA and R. O. SMITHERMAN, 1990 Genotype Environment Interactions for Growth of Blue, Channel and Hybrid Catfish in Ponds and Cages at Varying Densities. Aquaculture **85:** 143-151.

DUNHAM, R. A., J. A. JOYCE, K. BONDARI and S. P. MALVESTUTO, 1985 Evaluation of Body Conformation, Composition, and Density as Traits for Indirect Selection for Dress-out Percentage of Channel Catfish. Progressive Fish-Culturist **47:** 169-175.

DUNHAM, R. A., and R. O. SMITHERMAN, 1983a Crossbreeding Channel Catfish for Improvement of Body-Weight in Earthen Ponds. Growth **47:** 97-103.

DUNHAM, R. A., and R. O. SMITHERMAN, 1983b Response to Selection and Realized Heritability for Body-Weight in 3 Strains of Channel Catfish, Ictalurus-Punctatus, Grown in Earthen Ponds. Aquaculture **33:** 89-96.

DUNHAM, R. A., R. O. SMITHERMAN and R. K. GOODMAN, 1987 Comparison of Mass Selection, Crossbreeding, and Hybridization for Improving Growth of Channel Catfish. Progressive Fish-Culturist **49:** 293-296.

DUNHAM, R. A., R. O. SMITHERMAN, R. K. GOODMAN and P. KEMP, 1986 Comparison of Strains, Crossbreeds and Hybrids of Channel Catfish for Vulnerability to Angling. Aquaculture **57:** 193-201.

DUNHAM, R. A., R. O. SMITHERMAN and C. WEBBER, 1983 Relative Tolerance of Channel X Blue Hybrid and Channel Catfish to Low Oxygen Concentrations. Progressive Fish-Culturist **45:** 55-57.

DUPUIS, M. C., Z. Y. ZHANG, T. DRUET, J. M. DENOIX, C. CHARLIER *et al.*, 2011 Results of a haplotype-based GWAS for recurrent laryngeal neuropathy in the horse. Mammalian Genome **22:** 613-620.

ECK, S. H., A. BENET-PAGES, K. FLISIKOWSKI, T. MEITINGER, R. FRIES *et al.*, 2009 Whole genome sequencing of a single Bos taurus animal for single nucleotide polymorphism discovery. Genome Biology **10:** R82.

EELES, R. A., Z. KOTE-JARAI, A. A. AL OLAMA, G. G. GILES, M. GUY *et al.*, 2009 Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. Nature Genetics **41:** 1116-U1197.

EGGEN, A., 2012 The development and application of genomic selection as a new breeding paradigm. Animal Frontiers **2:** 10-15.

FAGEGALTIER, D., N. HUBERT, P. CARBON and A. KROL, 2000 The selenocysteine insertion sequence binding protein SBP is different from the Y-box protein dbpB. Biochimie **82:** 117-122.

FAMOSO, A. N., K. ZHAO, R. T. CLARK, C. W. TUNG, M. H. WRIGHT *et al.*, 2011 Genetic architecture of aluminum tolerance in rice (*Oryza sativa*) determined through genome-wide association analysis and QTL mapping. PLoS Genetics **7:** e1002221.

FINLAY, E. K., D. P. BERRY, B. WICKHAM, E. P. GORMLEY and D. G. BRADLEY, 2012 A genome wide association scan of bovine tuberculosis susceptibility in Holstein-Friesian dairy cattle. PLoS ONE **7:** e30545.

FISCHER, A., J. PALLAUF, K. GOHIL, S. U. WEBER, L. PACKER *et al.*, 2001 Effect of selenium and vitamin E deficiency on differential gene expression in rat liver. Biochem Biophys Res Commun **285:** 470-475.

FLETCHER, O., N. JOHNSON, N. ORR, F. J. HOSKING, L. J. GIBSON *et al.*, 2011 Novel breast cancer susceptibility locus at 9q31.2: results of a genome-wide association study. J Natl Cancer Inst **103:** 425-435.

FLICEK, P., and E. BIRNEY, 2009 Sense from sequence reads: methods for alignment and assembly. Nature Methods **6:** S6-S12.

FLINT, J., and E. ESKIN, 2012 Genome-wide association studies in mice. Nature Reviews Genetics **13:** 807-817.

FRAZER, K. A., S. S. MURRAY, N. J. SCHORK and E. J. TOPOL, 2009 Human genetic variation and its contribution to complex traits. Nature Reviews Genetics **10:** 241-251.

FREDMAN, D., S. J. WHITE, S. POTTER, E. E. EICHLER, J. T. DEN DUNNEN *et al.*, 2004 Complex SNP-related sequence variation in segmental genome duplications. Nature Genetics **36:** 861-866.

FROLOVA, L. Y., T. I. MERKULOVA and L. L. KISSELEV, 2000 Translation termination in eukaryotes: polypeptide release factor eRF1 is composed of functionally and structurally distinct domains. RNA **6:** 381-390.

FURUNO, M., T. KASUKAWA, R. SAITO, J. ADACHI, H. SUZUKI *et al.*, 2003 CDS annotation in full-length cDNA sequence. Genome Research **13:** 1478-1487.

GIBBONS, J. G., E. M. JANSON, C. T. HITTINGER, M. JOHNSTON, P. ABBOT *et al.*, 2009 Benchmarking next-generation transcriptome sequencing for functional and evolutionary genomics. Mol Biol Evol **26:** 2731-2744.

GIDSKEHAUG, L., M. KENT, B. J. HAYES and S. LIEN, 2011 Genotype calling and mapping of multisite variants using an Atlantic salmon iSelect SNP array. Bioinformatics **27:** 303-310.

GODDARD, M. E., and B. J. HAYES, 2009 Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nature Reviews Genetics **10:** 381-391.

GRABER, J. H., C. R. CANTOR, S. C. MOHR and T. F. SMITH, 1999 In silico detection of control signals: mRNA 3'-end-processing sequences in diverse species. Proc Natl Acad Sci U S A **96:** 14055-14060.

GRABHERR, M. G., B. J. HAAS, M. YASSOUR, J. Z. LEVIN, D. A. THOMPSON *et al.*, 2011 Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology **29:** 644-U130.

GRILLO, G., A. TURI, F. LICCIULLI, F. MIGNONE, S. LIUNI *et al.*, 2010 UTRdb and UTRsite (RELEASE 2010): a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. Nucleic Acids Res **38:** D75-80.

GRINDFLEK, E., S. LIEN, H. HAMLAND, M. H. HANSEN, M. KENT *et al.*, 2011 Large scale genome-wide association and LDLA mapping study identifies QTLs for boar taint and related sex steroids. BMC Genomics **12:** 362.

GROENEN, M. A. M., H. J. MEGENS, Y. ZARE, W. C. WARREN, L. W. HILLIER *et al.*, 2011 The development and characterization of a 60K SNP chip for chicken. BMC Genomics **12:** 274.

GU, X. R., C. G. FENG, L. MA, C. SONG, Y. Q. WANG *et al.*, 2011 Genome-Wide Association Study of Body Weight in Chicken F2 Resource Population. PLoS ONE **6:** e21872.

GUDMUNDSSON, J., P. SULEM, D. F. GUDBJARTSSON, T. BLONDAL, A. GYLFASON *et al.*, 2009 Genome-wide association and replication studies identify four variants associated with prostate cancer susceptibility. Nature Genetics **41:** 1122-U1104.

GURYEV, V., E. BEREZIKOV, R. MALIK, R. H. A. PLASTERK and E. CUPPEN, 2004 Single nucleotide polymorphisms associated with rat expressed sequences. Genome Research **14:** 1438-1443.

GUT, I. G., and G. M. LATHROP, 2004 Duplicating SNPs. Nature Genetics **36:** 789-790.

GUTIERREZ, A. P., K. P. LUBIENIECKI, E. A. DAVIDSON, S. LIEN, M. P. KENT *et al.*, 2012 Genetic mapping of quantitative trait loci (QTL) for body-weight in Atlantic salmon (*Salmo salar*) using a 6.5 K SNP array. Aquaculture **358:** 61-70.

HALE, M. C., C. R. MCCORMICK, J. R. JACKSON and J. A. DEWOODY, 2009 Next-generation pyrosequencing of gonad transcriptomes in the polyploid lake sturgeon (*Acipenser fulvescens*): the relative merits of normalization and rarefaction in gene discovery. BMC Genomics **10:** 203.

HALLERMAN, E. M., R. A. DUNHAM and R. O. SMITHERMAN, 1986 Selection or Drift Isozyme Allele Frequency Changes among Channel Catfish Selected for Rapid Growth. Transactions of the American Fisheries Society **115:** 60-68.

HAN, J. W., H. F. ZHENG, Y. CUI, L. D. SUN, D. Q. YE *et al.*, 2009 Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. Nature Genetics **41:** 1234-U1298.

HARDY, J., and A. SINGLETON, 2009 Genomewide association studies and human disease. N Engl J Med **360:** 1759-1768.

HARHAY, G. P., T. S. SONSTEGARD, J. W. KEELE, M. P. HEATON, M. L. CLAWSON *et al.*, 2005 Characterization of 954 bovine full-CDS cDNA sequences. BMC Genomics **6:** 166.

HAROLD, D., R. ABRAHAM, P. HOLLINGWORTH, R. SIMS, A. GERRISH *et al.*, 2009 Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nature Genetics **41:** 1088-1093.

HE, C., L. CHEN, M. SIMMONS, P. LI, S. KIM *et al.*, 2003 Putative SNP discovery in interspecific hybrids of catfish by comparative EST analysis. Animal Genetics **34:** 445-448.

HELYAR, S. J., M. T. LIMBORG, D. BEKKEVOLD, M. BABBUCCI, J. VAN HOUDT *et al.*, 2012 SNP Discovery Using Next Generation Transcriptomic Sequencing in Atlantic Herring (*Clupea harengus*). PLoS ONE **7:** e42089.

HILL, E. W., B. A. MCGIVNEY, J. J. GU, R. WHISTON and D. E. MACHUGH, 2010 A genome-wide SNP-association study confirms a sequence variant (g.66493737C > T) in the equine myostatin (MSTN) gene as the most powerful predictor of optimum racing distance for Thoroughbred racehorses. BMC Genomics **11:** 552.

HILLIER, L. W., G. T. MARTH, A. R. QUINLAN, D. DOOLING, G. FEWELL *et al.*, 2008 Whole-genome sequencing and variant discovery in C-elegans. Nature Methods **5:** 183-188.

HIRSCHHORN, J. N., and M. J. DALY, 2005 Genome-wide association studies for common diseases and complex traits. Nature Reviews Genetics **6:** 95-108.

HOFFMANN, T. J., M. N. KVALE, S. E. HESSELSON, Y. P. ZHAN, C. AQUINO *et al.*, 2011 Next generation genome-wide association tool: Design and coverage of a high-throughput European-optimized SNP array. Genomics **98:** 79-89.

HOU, R., Z. M. BAO, S. WANG, H. L. SU, Y. LI *et al.*, 2011 Transcriptome Sequencing and De Novo Analysis for Yesso Scallop (*Patinopecten yessoensis*) Using 454 GS FLX. PLoS ONE **6:** e21560.

HU, Z., C. WU, Y. SHI, H. GUO, X. ZHAO *et al.*, 2011 A genome-wide association study identifies two new lung cancer susceptibility loci at 13q12.12 and 22q12.2 in Han Chinese. Nature Genetics **43:** 792-796.

HUANG, X., X. WEI, T. SANG, Q. ZHAO, Q. FENG *et al.*, 2010 Genome-wide association studies of 14 agronomic traits in rice landraces. Nature Genetics **42:** 961-967.

HUANG, X., Y. ZHAO, X. WEI, C. LI, A. WANG *et al.*, 2012 Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nature Genetics **44:** 32-39.

HUBERT, S., B. HIGGINS, T. BORZA and S. BOWMAN, 2010 Development of a SNP resource and a genetic linkage map for Atlantic cod (*Gadus morhua*). BMC Genomics **11:** 191.

JIANG, L., J. F. LIU, D. X. SUN, P. P. MA, X. D. DING *et al.*, 2010 Genome Wide Association Studies for Milk Production Traits in Chinese Holstein Population. PLoS ONE **5:** e13661.

JOHNSTON, S. E., J. C. MCEWAN, N. K. PICKERING, J. W. KIJAS, D. BERALDI *et al.*, 2011 Genome-wide association mapping identifies the genetic basis of discrete and quantitative variation in sexual weaponry in a wild sheep population. Molecular Ecology **20:** 2555-2566.

KALDY, P., E. MENOTTI, R. MORET and L. C. KUHN, 1999 Identification of RNA-binding surfaces in iron regulatory protein-1. EMBO J **18:** 6073-6083.

KARLSSON, S., T. MOEN, S. LIEN, K. A. GLOVER and K. HINDAR, 2011 Generic genetic differences between farmed and wild Atlantic salmon identified from a 7K SNP-chip. Molecular Ecology Resources **11:** 247-253.

KENNEDY, G. C., H. MATSUZAKI, S. L. DONG, W. M. LIU, J. HUANG *et al.*, 2003 Large-scale genotyping of complex DNA. Nature Biotechnology **21:** 1233-1237.

KERSTENS, H. H. D., S. KOLLERS, A. KOMMADATH, M. DEL ROSARIO, B. DIBBITS *et al.*, 2009 Mining for single nucleotide polymorphisms in pig genome sequence data. BMC Genomics **10:** 4.

KHATKAR, M. S., F. W. NICHOLAS, A. R. COLLINS, K. R. ZENGER, J. A. CAVANAGH *et al.*, 2008 Extent of genome-wide linkage disequilibrium in Australian Holstein-Friesian cattle based on a high-density SNP panel. BMC Genomics **9:** 187.

KHOR, C. C., S. DAVILA, W. B. BREUNIS, Y. C. LEE, C. SHIMIZU *et al.*, 2011 Genome-wide association study identifies FCGR2A as a susceptibility locus for Kawasaki disease. Nature Genetics **43:** 1241-U1104.

KIJAS, J. W., D. TOWNLEY, B. P. DALRYMPLE, M. P. HEATON, J. F. MADDOX *et al.*, 2009 A genome wide survey of SNP variation reveals the genetic structure of sheep breeds. PLoS ONE **4:** e4668.

KIM, E. S., P. J. BERGER and B. W. KIRKPATRICK, 2009 Genome-wide scan for bovine twinning rate QTL using linkage disequilibrium. Animal Genetics **40:** 300-307.

KIM, J. J., Y. M. HONG, S. SOHN, G. Y. JANG, K. S. HA *et al.*, 2011 A genome-wide association analysis reveals 1p31 and 2p13.3 as susceptibility loci for Kawasaki disease. Human Genetics **129:** 487-495.

KISSELEV, L. L., and R. H. BUCKINGHAM, 2000 Translational termination comes of age. Trends Biochem Sci **25:** 561-566.

KLEIN, R. J., C. ZEISS, E. Y. CHEW, J. Y. TSAI, R. S. SACKLER *et al.*, 2005 Complement factor H polymorphism in age-related macular degeneration. Science **308:** 385-389.

KOZAK, M., 1987 An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs. Nucleic Acids Res **15:** 8125-8148.

KRANIS, A., A. A. GHEYAS, C. BOSCHIERO, F. TURNER, L. YU *et al.*, 2013 Development of a high density 600K SNP genotyping array for chicken. BMC Genomics **14:** 59.

KRUGLYAK, L., 1997 The use of a genetic map of biallelic markers in linkage studies. Nature Genetics **17:** 21-24.

KRUGLYAK, L., 2008 The road to genome-wide association studies. Nature Reviews Genetics **9:** 314-318.

KUCUKTAS, H., S. WANG, P. LI, C. HE, P. XU *et al.*, 2009 Construction of genetic linkage maps and comparative genome analysis of catfish using gene-associated markers. Genetics **181:** 1649-1660.

LE, S. Q., and R. DURBIN, 2011 SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. Genome Research **21:** 952-960.

LI, D. F., L. LIAN, L. J. QU, Y. M. CHEN, W. B. LIU *et al.*, 2012 A genome-wide SNP scan reveals two loci associated with the chicken resistance to Marek's disease. Animal Genetics **44:** 217-222.

LI, R. Q., Y. R. LI, X. D. FANG, H. M. YANG, J. WANG *et al.*, 2009 SNP detection for massively parallel whole-genome resequencing. Genome Research **19:** 1124-1132.

LI, W., and A. GODZIK, 2006 Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics **22:** 1658-1659.

LIEN, S., L. GIDSKEHAUG, T. MOEN, B. J. HAYES, P. R. BERG *et al.*, 2011 A dense SNP-based linkage map for Atlantic salmon (*Salmo salar*) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns. BMC Genomics **12:** 615.

LIN, X., D. LU, Y. GAO, S. TAO, X. YANG *et al.*, 2012a Genome-wide association study identifies novel loci associated with serum level of vitamin B12 in Chinese men. Human Molecular Genetics **21:** 2610-2617.

LIN, Z., J. X. BEI, M. SHEN, Q. LI, Z. LIAO *et al.*, 2012b A genome-wide association study in Han Chinese identifies new susceptibility loci for ankylosing spondylitis. Nature Genetics **44:** 73-77.

LINDBLAD-TOH, K., E. WINCHESTER, M. J. DALY, D. G. WANG, J. N. HIRSCHHORN *et al.*, 2000 Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. Nature Genetics **24:** 381-386.

LIU, Q., 2005 Comparative analysis of base biases around the stop codons in six eukaryotes. Biosystems **81:** 281-289.

LIU, S., Z. ZHOU, J. LU, F. SUN, S. WANG *et al.*, 2011a Generation of genome-scale gene-associated SNPs in catfish for the construction of a high-density SNP array. BMC Genomics **12:** 53.

LOMAX, J., 2005 Get ready to GO! A biologist's guide to the Gene Ontology. Brief Bioinform **6:** 298-304.

LORENZ, S., S. BRENNA-HANSEN, T. MOEN, A. ROSETH, W. S. DAVIDSON *et al.*, 2010 BAC-based upgrading and physical integration of a genetic SNP map in Atlantic salmon. Animal Genetics **41:** 48-54.

LU, X., L. WANG, S. CHEN, L. HE, X. YANG *et al.*, 2012 Genome-wide association study in Han Chinese identifies four new susceptibility loci for coronary artery disease. Nature Genetics **44:** 890-894.

MACDONALD, C. C., and J. L. REDONDO, 2002 Reexamining the polyadenylation signal: were we wrong about AAUAAA? Mol Cell Endocrinol **190:** 1-8.

MAI, M. D., G. SAHANA, F. B. CHRISTIANSEN and B. GULDBRANDTSEN, 2010 A genome-wide association study for milk production traits in Danish Jersey cattle using a 50K single nucleotide polymorphism chip. Journal of Animal Science **88:** 3522-3528.

MANOLIO, T. A., 2010 Genomewide association studies and assessment of the risk of disease. N Engl J Med **363:** 166-176.

MANOLIO, T. A., F. S. COLLINS, N. J. COX, D. B. GOLDSTEIN, L. A. HINDORFF *et al.*, 2009 Finding the missing heritability of complex diseases. Nature **461:** 747-753.

MARDIS, E. R., 2008a The impact of next-generation sequencing technology on genetics. Trends in Genetics **24:** 133-141.

MARDIS, E. R., 2008b Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics **9:** 387-402.

MATSUZAKI, H., S. L. DONG, H. LOI, X. J. DI, G. Y. LIU *et al.*, 2004 Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. Nature Methods **1:** 109-111.

MATUKUMALLI, L. K., C. T. LAWLEY, R. D. SCHNABEL, J. F. TAYLOR, M. F. ALLAN *et al.*, 2009 Development and characterization of a high density SNP genotyping assay for cattle. PLoS ONE **4:** e5350.

MCCARROLL, S. A., F. G. KURUVILLA, J. M. KORN, S. CAWLEY, J. NEMESH *et al.*, 2008 Integrated detection and population-genetic analysis of SNPs and copy number variation. Nature Genetics **40:** 1166-1174.

McCarthy, M. I., G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little *et al.*, 2008 Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Reviews Genetics **9:** 356-369.

McCaughan, K. K., C. M. Brown, M. E. Dalphin, M. J. Berry and W. P. Tate, 1995 Translational termination efficiency in mammals is influenced by the base following the stop codon. Proc Natl Acad Sci U S A **92:** 5431-5435.

McCue, M. E., D. L. Bannasch, J. L. Petersen, J. Gurr, E. Bailey *et al.*, 2012 A High Density SNP Array for the Domestic Horse and Extant Perissodactyla: Utility for Association Mapping, Genetic Diversity, and Phylogeny Studies. Plos Genetics **8:** e1002451.

Melum, E., A. Franke, C. Schramm, T. J. Weismuller, D. N. Gotthardt *et al.*, 2011 Genome-wide association analysis in primary sclerosing cholangitis identifies two non-HLA susceptibility loci. Nature Genetics **43:** 17-19.

Meredith, B. K., F. J. Kearney, E. K. Finlay, D. G. Bradley, A. G. Fahey *et al.*, 2012 Genome-wide associations for milk production and somatic cell score in Holstein-Friesian cattle in Ireland. BMC Genetics **13:** 21.

Meurs, K. M., E. Mauceli, S. Lahmers, G. M. Acland, S. N. White *et al.*, 2010 Genome-wide association identifies a deletion in the 3' untranslated region of Striatin in a canine model of arrhythmogenic right ventricular cardiomyopathy. Human Genetics **128:** 315-324.

Meuwissen, T. H. E., B. J. Hayes and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics **157:** 1819-1829.

Meyer, E., G. V. Aglyamova, S. Wang, J. Buchanan-Carter, D. Abrego *et al.*, 2009 Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. BMC Genomics **10:** 219.

Miller, J. M., J. Poissant, J. W. Kijas, D. W. Coltman and I. S. G. Consortium, 2011 A genome-wide set of SNPs detects population substructure and long range linkage disequilibrium in wild sheep. Molecular Ecology Resources **11:** 314-322.

Min, X. J., G. Butler, R. Storms and A. Tsang, 2005 OrfPredictor: predicting protein-coding regions in EST-derived sequences. Nucleic Acids Res **33:** W677-680.

Mizrachi, E., C. A. Hefer, M. Ranik, F. Joubert and A. A. Myburg, 2010 *De novo* assembled expressed gene catalog of a fast-growing Eucalyptus tree produced by Illumina mRNA-Seq. BMC Genomics **11:** 681.

Moen, T., M. Delghandi, M. S. Wesmajervi, J. I. Westgaard and K. T. Fjalestad, 2009 A SNP/microsatellite genetic linkage map of the Atlantic cod (*Gadus morhua*). Animal Genetics **40:** 993-996.

Moen, T., B. Hayes, M. Baranski, P. R. Berg, S. Kjoglum *et al.*, 2008 A linkage map of the Atlantic salmon (*Salmo salar*) based on EST-derived SNP markers. BMC Genomics **9**.

Mogensen, M. S., P. Karlskov-Mortensen, H. F. Proschowsky, F. Lingaas, A. Lappalainen *et al.*, 2011a Genome-Wide Association Study in Dachshund: Identification of a Major Locus Affecting Intervertebral Disc Calcification. Journal of Heredity **102:** S81-S86.

Morin, P. A., G. Luikart, R. K. Wayne and S. W. Grp, 2004 SNPs in ecology, evolution and conservation. Trends in Ecology & Evolution **19:** 208-216.

Mottagui-Tabar, S., M. F. Tuite and L. A. Isaksson, 1998 The influence of 5' codon context on translation termination in Saccharomyces cerevisiae. Eur J Biochem **257:** 249-254.

MUCHERO, W., N. N. DIOP, P. R. BHAT, R. D. FENTON, S. WANAMAKER *et al.*, 2009 A consensus genetic map of cowpea [*Vigna unguiculata* (L) Walp.] and synteny based on EST-derived SNPs. Proceedings of the National Academy of Sciences of the United States of America **106:** 18159-18164.

NEWTON-CHEH, C., T. JOHNSON, V. GATEVA, M. D. TOBIN, M. BOCHUD *et al.*, 2009 Genome-wide association study identifies eight loci associated with blood pressure. Nature Genetics **41:** 666-676.

NIELSEN, R., J. S. PAUL, A. ALBRECHTSEN and Y. S. SONG, 2011 Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics **12:** 443-451.

NINWICHIAN, P., E. PEATMAN, H. LIU, H. KUCUKTAS, B. SOMRIDHIVEJ *et al.*, 2012 Second-Generation Genetic Linkage Map of Catfish and Its Integration with the BAC-Based Physical Map. G3-Genes Genomes Genetics **2:** 1233-1241.

NOVAES, E., D. R. DROST, W. G. FARMERIE, G. J. PAPPAS, JR., D. GRATTAPAGLIA *et al.*, 2008 High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. BMC Genomics **9:** 312.

ONTERU, S. K., B. FAN, Z. Q. DU, D. J. GARRICK, K. J. STALDER *et al.*, 2012 A whole-genome association study for pig reproductive traits. Animal Genetics **43:** 18-26.

ONTERU, S. K., B. FAN, M. T. NIKKILA, D. J. GARRICK, K. J. STALDER *et al.*, 2011 Whole-genome association analyses for lifetime reproductive traits in the pig. Journal of Animal Science **89:** 988-995.

ORR, N., W. BACK, J. GU, P. LEEGWATER, P. GOVINDARAJAN *et al.*, 2010 Genome-wide SNP association-based localization of a dwarfism gene in Friesian dwarf horses. Animal Genetics **41:** 2-7.

PAINTER, J. N., C. A. ANDERSON, D. R. NYHOLT, S. MACGREGOR, J. LIN *et al.*, 2011 Genome-wide association study identifies a locus at 7p15.2 associated with endometriosis. Nature Genetics **43:** 51-54.

PANT, S. D., F. S. SCHENKEL, C. P. VERSCHOOR, Q. M. YOU, D. F. KELTON *et al.*, 2010 A principal component regression based genome wide analysis approach reveals the presence of a novel QTL on BTA7 for MAP resistance in holstein cattle. Genomics **95:** 176-182.

PARCHMAN, T. L., K. S. GEIST, J. A. GRAHNEN, C. W. BENKMAN and C. A. BUERKLE, 2010 Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. BMC Genomics **11:** 180.

PETKOV, P. M., Y. M. DING, M. A. CASSELL, W. D. ZHANG, G. WAGNER *et al.*, 2004 An efficient SNP system for mouse genome scanning and elucidating strain relationships. Genome Research **14:** 1806-1811.

PHILIPP, U., B. LUPP, S. MOMKE, V. STEIN, A. TIPOLD *et al.*, 2011 A MITF mutation associated with a dominant white phenotype and bilateral deafness in German Fleckvieh cattle. PLoS ONE **6:** e28857.

PONSUKSILI, S., E. MURANI, B. BRAND, M. SCHWERIN and K. WIMMERS, 2011 Integrating expression profiling and whole-genome association for dissection of fat traits in a porcine model. Journal of Lipid Research **52:** 668-678.

QI, L., M. C. CORNELIS, P. KRAFT, K. J. STANYA, W. H. LINDA KAO *et al.*, 2010 Genetic variants at 2q24 are associated with susceptibility to type 2 diabetes. Human Molecular Genetics **19:** 2706-2715.

RAMOS, A. M., R. P. M. A. CROOIJMANS, N. A. AFFARA, A. J. AMARAL, A. L. ARCHIBALD *et al.*, 2009 Design of a High Density SNP Genotyping Assay in the Pig Using SNPs Identified and Characterized by Next Generation Sequencing Technology. PLoS ONE **4:** e6524.

RAMOS, A. M., H. J. MEGENS, R. P. CROOIJMANS, L. B. SCHOOK and M. A. GROENEN, 2011 Identification of high utility SNPs for population assignment and traceability purposes in the pig using high-throughput sequencing. Animal Genetics **42:** 613-620.

RAPLEY, E. A., C. TURNBULL, A. A. AL OLAMA, E. T. DERMITZAKIS, R. LINGER *et al.*, 2009 A genome-wide association study of testicular germ cell tumor. Nature Genetics **41:** 807-U859.

REZK, M. A., R. O. SMITHERMAN, J. C. WILLIAMS, A. NICHOLS, H. KUCUKTAS *et al.*, 2003 Response to three generations of selection for increased body weight in channel catfish, Ictalurus punctatus, grown in earthen ponds. Aquaculture **228:** 69-79.

RIGOUTSOS, I., and A. FLORATOS, 1998 Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. Bioinformatics **14:** 55-67.

RINCON, G., K. L. WEBER, A. L. VAN EENENNAAM, B. L. GOLDEN and J. F. MEDRANO, 2011 Hot topic: Performance of bovine high-density genotyping platforms in Holsteins and Jerseys. Journal of Dairy Science **94:** 6116-6121.

ROBERTSON, G., J. SCHEIN, R. CHIU, R. CORBETT, M. FIELD *et al.*, 2010 De novo assembly and analysis of RNA-seq data. Nature Methods **7:** 909-912.

ROZEN, S., and H. SKALETSKY, 2000 Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol **132:** 365-386.

SAHANA, G., B. GULDBRANDTSEN, C. BENDIXEN and M. S. LUND, 2010 Genome-wide association mapping for female fertility traits in Danish and Swedish Holstein cattle. Animal Genetics **41:** 579-588.

SAHANA, G., B. GULDBRANDTSEN and M. S. LUND, 2011 Genome-wide association study for calving traits in Danish and Swedish Holstein cattle. Journal of Dairy Science **94:** 479-486.

SAHANA, G., V. KADLECOVA, H. HORNSHOJ, B. NIELSEN and O. F. CHRISTENSEN, 2013 A genome-wide association scan in pig identifies novel regions associated with feed efficiency trait. Journal of Animal Science **91:** 1041-1050.

SANCHEZ, C. C., T. P. SMITH, R. T. WIEDMANN, R. L. VALLEJO, M. SALEM *et al.*, 2009 Single nucleotide polymorphism discovery in rainbow trout by deep sequencing of a reduced representation library. BMC Genomics **10:** 559.

SANDMANN, T., M. C. VOGG, S. OWLARN, M. BOUTROS and K. BARTSCHERER, 2011 The head-regeneration transcriptome of the planarian Schmidtea mediterranea. Genome Biology **12:** R76.

SCHAEFFER, L. R., 2006 Strategy for applying genome-wide selection in dairy cattle. Journal of Animal Breeding and Genetics **123:** 218-223.

SCHMID, K. J., T. R. SORENSEN, R. STRACKE, O. TORJEK, T. ALTMANN *et al.*, 2003 Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in Arabidopsis thaliana. Genome Research **13:** 1250-1257.

SCHOPEN, G. C. B., M. H. P. W. VISKER, P. D. KOKS, E. MULLAART, J. A. M. VAN ARENDONK *et al.*, 2011 Whole-genome association study for milk protein composition in dairy cattle. Journal of Dairy Science **94:** 3148-3158.

SCHUSTER, S. C., 2008 Next-generation sequencing transforms today's biology. Nature Methods **5:** 16-18.

SETTLES, M., R. ZANELLA, S. D. MCKAY, R. D. SCHNABEL, J. F. TAYLOR *et al.*, 2009 A whole genome association analysis identifies loci associated with Mycobacterium avium subsp paratuberculosis infection status in US holstein cattle. Animal Genetics **40:** 655-662.

SHI, Y. Y., Z. B. HU, C. WU, J. C. DAI, H. Z. LI *et al.*, 2011 A genome-wide association study identifies new susceptibility loci for non-cardia gastric cancer at 3q13.31 and 5p13.1. Nature Genetics **43:** 1215-U1266.

SIMMONS, M., K. MICKETT, H. KUCUKTAS, P. LI, R. DUNHAM *et al.*, 2006 Comparison of domestic and wild channel catfish (Ictalurus punctatus) populations provides no evidence for genetic impact. Aquaculture **252:** 133-146.

SIMON-SANCHEZ, J., C. SCHULTE, J. M. BRAS, M. SHARMA, J. R. GIBBS *et al.*, 2009 Genome-wide association study reveals genetic risk underlying Parkinson's disease. Nature Genetics **41:** 1308-1312.

SIMPSON, J. T., K. WONG, S. D. JACKMAN, J. E. SCHEIN, S. J. JONES *et al.*, 2009 ABySS: a parallel assembler for short read sequence data. Genome Research **19:** 1117-1123.

SLADEK, R., G. ROCHELEAU, J. RUNG, C. DINA, L. SHEN *et al.*, 2007 A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature **445:** 881-885.

SMIT A, Hubley R., GREEN P: **REPEATMASKER OPEN-3.2.2.** http://www.repeatmasker.org/.

SNELLING, W. M., E. CASAS, R. T. STONE, J. W. KEELE, G. P. HARHAY *et al.*, 2005 Linkage mapping bovine EST-based SNP. BMC Genomics **6**.

SOMRIDHIVEJ, B., S. L. WANG, Z. X. SHA, H. LIU, J. QUILANG *et al.*, 2008 Characterization, polymorphism assessment, and database construction for microsatellites from BAC end sequences of channel catfish (Ictalurus punctatus): A resource for integration of linkage and physical maps. Aquaculture **275:** 76-80.

SONESSON, A. K., and T. H. E. MEUWISSEN, 2009 Testing strategies for genomic selection in aquaculture breeding programs. Genetics Selection Evolution **41:** 37.

STOTHARD, P., J. W. CHOI, U. BASU, J. M. SUMNER-THOMSON, Y. MENG *et al.*, 2011 Whole genome resequencing of Black Angus and Holstein cattle for SNP and CNV discovery. BMC Genomics **12:** 559.

SUN, L. D., F. L. XIAO, Y. LI, W. M. ZHOU, H. Y. TANG *et al.*, 2011 Genome-wide association study identifies two new susceptibility loci for atopic dermatitis in the Chinese Han population. Nature Genetics **43:** 690-694.

SURGET-GROBA, Y., and J. I. MONTOYA-BURGOS, 2010 Optimization of de novo transcriptome assembly from next-generation sequencing data. Genome Research **20:** 1432-1440.

TATE, W. P., E. S. POOLE, J. A. HORSFIELD, S. A. MANNERING, C. M. BROWN *et al.*, 1995 Translational termination efficiency in both bacteria and mammals is regulated by the base following the stop codon. Biochem Cell Biol **73:** 1095-1103.

THOMSON, A. M., J. T. ROGERS and P. J. LEEDMAN, 1999 Iron-regulatory proteins, iron-responsive elements and ferritin mRNA translation. Int J Biochem Cell Biol **31:** 1139-1152.

TORRES, T. T., M. METTA, B. OTTENWALDER and C. SCHLOTTERER, 2008 Gene expression profiling by massively parallel sequencing. Genome Research **18:** 172-177.

TRICK, M., Y. LONG, J. L. MENG and I. BANCROFT, 2009 Single nucleotide polymorphism (SNP) discovery in the polyploid Brassica napus using Solexa transcriptome sequencing. Plant Biotechnology Journal **7:** 334-346.

TUPLER, R., G. PERINI and M. R. GREEN, 2001 Expressing the human genome. Nature **409:** 832-833.

UEMOTO, Y., T. ABE, N. TAMEOKA, H. HASEBE, K. INOUE *et al.*, 2010 Whole-genome association study for fatty acid composition of oleic acid in Japanese Black cattle. Animal Genetics **42:** 141-148.

VAN HULZEN, K. J., G. C. SCHOPEN, J. A. VAN ARENDONK, M. NIELEN, A. P. KOETS *et al.*, 2012 Genome-wide association study to identify chromosomal regions associated with antibody response to Mycobacterium avium subspecies paratuberculosis in milk of Dutch Holstein-Friesians. Journal of Dairy Science **95:** 2740-2748.

VERA, J. C., C. W. WHEAT, H. W. FESCEMYER, M. J. FRILANDER, D. L. CRAWFORD *et al.*, 2008 Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Molecular Ecology **17:** 1636-1647.

VITHANA, E. N., C. C. KHOR, C. QIAO, M. E. NONGPIUR, R. GEORGE *et al.*, 2012 Genome-wide association analyses identify three new susceptibility loci for primary angle closure glaucoma. Nature Genetics **44:** 1142-1146.

WALCZAK, R., E. WESTHOF, P. CARBON and A. KROL, 1996 A novel RNA structural motif in the selenocysteine insertion element of eukaryotic selenoprotein mRNAs. RNA **2:** 367-379.

WALDBIESER, G. C., B. G. BOSWORTH and S. M. QUINIOU, 2010 Production of viable homozygous, doubled haploid channel catfish (Ictalurus punctatus). Mar Biotechnol (NY) **12:** 380-385.

WANG, D. G., J. B. FAN, C. J. SIAO, A. BERNO, P. YOUNG *et al.*, 1998 Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science **280:** 1077-1082.

WANG, F., C. Q. XU, Q. HE, J. P. CAI, X. C. LI *et al.*, 2011 Genome-wide association identifies a susceptibility locus for coronary artery disease in the Chinese Han population. Nature Genetics **43:** 345-U396.

WANG, S., E. PEATMAN, J. ABERNATHY, G. WALDBIESER, E. LINDQUIST *et al.*, 2010 Assembly of 500,000 inter-specific catfish expressed sequence tags and large scale gene-associated marker development for whole genome association studies. Genome Biology **11:** R8.

WANG, S., Z. SHA, T. S. SONSTEGARD, H. LIU, P. XU *et al.*, 2008 Quality assessment parameters for EST-derived SNPs from catfish. Bmc Genomics **9:** 450.

WANG, Z., M. GERSTEIN and M. SNYDER, 2009 RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics **10:** 57-63.

WELLS, K. L., Y. HADAD, D. BEN-AVRAHAM, J. HILLEL, A. CAHANER *et al.*, 2012 Genome-wide SNP scan of pooled DNA reveals nonsense mutation in FGF20 in the scaleless line of featherless chickens. BMC Genomics **13:** 257.

WOLC, A., J. ARANGO, P. SETTAR, J. E. FULTON, N. P. O'SULLIVAN *et al.*, 2012 Genome-wide association analysis and genetic architecture of egg weight and egg uniformity in layer chickens. Animal Genetics **43:** 87-96.

WOLTERS, W. R., and M. R. JOHNSON, 1995 Analysis of a diallel cross to estimate effects of crossing on resistance to enteric septicemia in channel catfish, Ictalurus punctatus. Aquaculture **137:** 263-269.

WU, C., Z. HU, Z. HE, W. JIA, F. WANG *et al.*, 2011 Genome-wide association study identifies three new susceptibility loci for esophageal squamous-cell carcinoma in Chinese populations. Nature Genetics **43:** 679-684.

WU, C., X. MIAO, L. HUANG, X. CHE, G. JIANG *et al.*, 2012a Genome-wide association study identifies five loci associated with susceptibility to pancreatic cancer in Chinese populations. Nature Genetics **44:** 62-66.

WU, X., G. SCELO, M. P. PURDUE, N. ROTHMAN, M. JOHANSSON *et al.*, 2012b A genome-wide association study identifies a novel susceptibility locus for renal cell carcinoma on 12p11.23. Human Molecular Genetics **21:** 456-462.

XIE, L., C. L. LUO, C. G. ZHANG, R. ZHANG, J. TANG *et al.*, 2012 Genome-Wide Association Study Identified a Narrow Chromosome 1 Region Associated with Chicken Growth Traits. PLoS ONE **7:** e30910.

XU, J., Z. MO, D. YE, M. WANG, F. LIU *et al.*, 2012 Genome-wide association study in Chinese men identifies two new prostate cancer risk loci at 9q31.2 and 19q13.4. Nature Genetics.

XU, P., S. L. WANG, L. LIU, J. THORSEN, H. KUCUKTAS *et al.*, 2007 A BAC-based physical map of the channel catfish genome. Genomics **90:** 380-388.

YOU, F. M., N. X. HUO, K. R. DEAL, Y. Q. GU, M. C. LUO *et al.*, 2011 Annotation-based genome-wide SNP discovery in the large and complex Aegilops tauschii genome using next-generation sequencing without a reference genome sequence. BMC Genomics **12**.

ZERBINO, D. R., and E. BIRNEY, 2008 Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research **18:** 821-829.

ZHANG, H., Z. WANG, S. WANG and H. LI, 2012 Progress of genome wide association study in domestic animals. J Anim Sci Biotechnol **3:** 26.

ZHANG, K., and F. SUN, 2005 Assessing the power of tag SNPs in the mapping of quantitative trait loci (QTL) with extremal and random samples. BMC Genetics **6:** 51.

ZHANG, X. J., W. HUANG, S. YANG, L. D. SUN, F. Y. ZHANG *et al.*, 2009 Psoriasis genome-wide association study identifies susceptibility variants within LCE gene cluster at 1q21. Nature Genetics **41:** 205-210.

ZHAO, X., K. E. DITTMER, H. T. BLAIR, K. G. THOMPSON, M. F. ROTHSCHILD *et al.*, 2011 A novel nonsense mutation in the DMP1 gene identified by a genome-wide association study is responsible for inherited rickets in Corriedale sheep. PLoS ONE **6:** e21739.