

ASSEMBLY OF 500,000 INTER-SPECIFIC CATFISH EXPRESSED SEQUENCE
TAGS AND LARGE SCALE GENE-ASSOCIATED MARKER
DEVELOPMENT FOR GENOME SELECTION STUDIES

Except where reference is made to the work of others, the work described in this thesis is my own or was done in collaboration with my advisory committee. This thesis does not include proprietary or classified information.

Shaolin Wang

Certificate of Approval:

Joanna Diller
Associate Professor
Biology

Kevin Fielman
Assistant Professor
Biology

Zhanjiang (John) Liu, Chair
Professor
Fisheries and Allied Aquacultures

Covadonga Arias
Associate Professor
Fisheries and Allied Aquacultures

George T. Flowers
Dean
Graduate School

ASSEMBLY OF 500,000 INTER-SPECIFIC CATFISH EXPRESSED SEQUENCE
TAGS AND LARGE SCALE GENE-ASSOCIATED MARKER
DEVELOPMENT FOR GENOME SELECTION STUDIES

Shaolin Wang

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 10th, 2009

ASSEMBLY OF 500,000 INTER-SPECIFIC CATFISH EXPRESSED SEQUENCE
TAGS AND LARGE SCALE GENE-ASSOCIATED MARKER
DEVELOPMENT FOR GENOME SELECTION STUDIES

Shaolin Wang

Permission is granted to Auburn University to make copies of this dissertation at its discretion, upon request of individuals or institutions and at their expense.
The author reserves all publication rights.

Signature of Author

Date of Graduation

VITA

Shaolin Wang, son of Jundong Wang and Jianhua Hong, was born Jan 24th, 1982, in Tiangu, China. He graduated with a Bachelor of Science degree in 2003 from China Agriculture University, China majoring in Biology. He entered Graduate School of Auburn University in the Department of Fisheries and Allied Aquacultures in August 2004, to pursue a Doctor of Philosophy degree in August 2009.

DISSERTATION ABSTRACT

ASSEMBLY OF 500,000 INTER-SPECIFIC CATFISH EXPRESSED SEQUENCE
TAGS AND LARGE SCALE GENE-ASSOCIATED MARKER
DEVELOPMENT FOR GENOME SELECTION STUDIES

Shaolin Wang

Doctor of Philosophy, August 10th, 2009
(B.Sc., China Agriculture University, 2003)

110 Typed Pages

Directed by Zhanjiang (John) Liu

Expressed Sequence Tag (EST) sequencing is one of the most efficient means for gene discovery and gene expression profiling. With a good resource of ESTs, a large number of molecular markers can be identified, and issues related to alternative splicing and differential poly adenylation can be addressed at the genome-wide scale. Through the Community Sequencing Program, a catfish EST sequencing project was selected by the DOE's Joint Genome Institute (JGI). In this project, a total of 12 cDNA libraries were constructed including eight from channel catfish (*Ictalurus punctatus*) and four from blue catfish (*I. furcatus*). A total of 600,000 sequencing attempts were made, generating a total of 438,321 quality ESTs. With previously existing ESTs in GenBank, this project brings the total of ESTs to nearly 500,000 in the catfish. The JGI EST sequencing had an overall sequencing success rate of 73% with an average length of 576 bp. All the ESTs were

assembled using CAP3, resulting in 111,578 unique sequences, including 45,306 contigs and 66,272 singletons. Of these unique sequences, over 35% had significant similarities to known genes by BLASTX searches, which allowed the identification of 14,776 unique genes in the catfish. A total of 1,350 and 849 full length cDNAs have been identified from channel catfish and blue catfish, respectively. The ESTs are an enormous resource for SNP identification. The quality assessment parameters for EST-derived were established based on a pilot study with 384 SNPs. In order to select reliable SNPs, contigs containing four or more ESTs should be used and the minor allele sequence should be represented at least twice. Genotyping primers should be designed from a single exon, completely avoiding introns. Application of such quality assessment measures, along with large resources of ESTs, should provide effective means for SNP identification in species where genome sequence resources are lacking. Over 300,000 putative SNPs have been identified, of which over 48,000 are high quality SNPs as defined by contig size of at least four sequences and the minor allele presence of at least twice in the contig. The EST resource should also be valuable for identification of microsatellites, comparative genome analysis. This large scale EST sequencing project would allow the identification of majority of catfish transcriptome. The parallel analysis of ESTs from the two closely related ictalurid catfishes should also provide powerful means for the evaluation of ancient and recent gene duplications, and for the development of high-density microarrays in catfish. The inter- and intra- specific SNPs identified from all catfish EST dataset assembly will greatly benefit the catfish introgression breeding selection and whole genome association studies. All ESTs have been deposited in GenBank.

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to my major professor, Dr. Zhanjiang Liu, for his guidance throughout my study. I would like to express my gratitude to my committee: Dr. Joanna Diller, Dr. Covadonga Arias, and Dr. Kevin Fielman for their advice and critical reading of my dissertation. My thanks also go to all the colleagues in the laboratory for their help, collaboration, and friendship. I am grateful to my parents and my beloved wife for their constant support.

Style manual used Genome Research

Computer software used Microsoft Word 2003, Microsoft Excel 2003, Adobe Photoshop CS2, BeadStudio, JoinMap, R, Phred, Lucy, CAP3, RepeatMasker, Msatfinder, ESTscan, Pfam, AutoSNP, NCBI-blast, Blast2go, APACHE, PHP, MySQL.

TABLE OF CONTENTS

LIST OF TABLES	x
LIST OF FIGURES	xii
I. INTRODUCTION	1
II. RESEARCH OBJECTIVES	27
III. QUALITY ASSESSMENT OF EST-DERIVED SNPS FROM CATFISH.....	28
Abstract	29
Background	31
Results	33
Discussion	40
Materials and methods	47
Acknowledgements.....	50
References	51
IV. ASSEMBLY OF 500,000 INTER-SPECIFIC CATFISH EXPRESSED SEQUENCE TAGS AND LARGE SCALE GENE-ASSOCIATED MARKER DEVELOPMENT FOR GENOME SELECTION STUDIES	55
Abstract	56
Background	58
Results	59
Discussion	77
Materials and methods	82
Acknowledgements.....	85
References	86
V. CONCLUSIONS	92
APPENDIX.....	94

LIST OF TABLES

Table 1 Summary of the EST assembly	33
Table 2 Initial identification of SNPs as detected by AutoSNP software	34
Table 3 Overall summary of the EST-derived SNP genotyping using the Illumina Bead Array technology	35
Table 4 SNP polymorphic rates vs contig size and minor sequence allele frequency	38
Table 5 Effect of low sequence quality (as defined by the presence of hot spots of SNP occurrence) and the presence of predicted intron on success rate of SNP genotyping	40
Table 6 cDNA library information and sequencing summary	60
Table 7 EST assembly statistics	61
Table 8 Inter-specific similarity comparison of blue catfish and channel catfish unique sequences	64
Table 9 Summary of BLASTX searches analysis of catfish ESTs	67
Table 10 Full-length cDNA identification	70
Table 11 Summary of microsatellite marker identification from catfish ESTs.....	71
Table 12 Summary of SNP identification from the catfish ESTs.....	72

Table 13	Quality assessment of the filtered putative SNPs identified from the catfish ESTs based on the number of sequences per contig and the sequence frequencies of the minor alleles.....	73
Table 14	Estimation of proportions of inter-specific and intra-specific SNPs from the set of filtered SNPs identified from the inter-specific all catfish EST assembly	76

LIST OF FIGURES

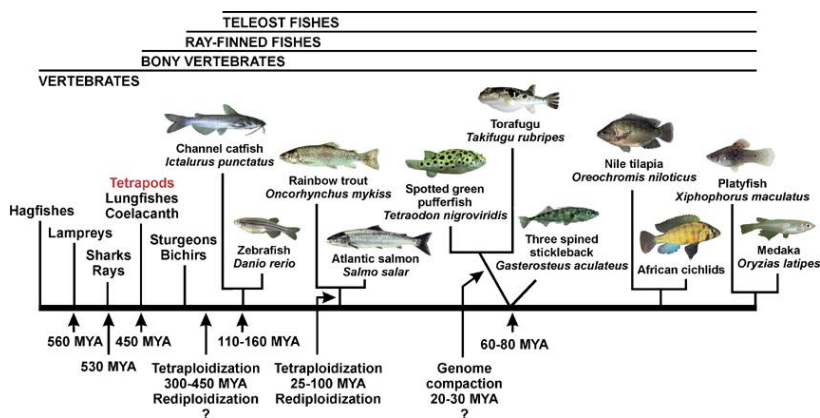
Figure 1 History of fish evolution.....	1
Figure 2 U.S. catfish production	3
Figure 3 Distribution of minor allele frequency in domestic and wild channel catfish strains.....	36
Figure 4 SNP quality assessment based on EST contig size and sequence frequency of the alleles	43
Figure 5 Schematic illustration of the effect of introns involved in SNP genotyping	45
Figure 6 Length distribution of JGI EST sequences	61
Figure 7 Distribution of contig sizes.....	63
Figure 8 Open reading frame (ORF) length distribution from unique sequences of all catfish assembly	65
Figure 9 Analysis of open reading frames	66
Figure 10 Number of catfish homologous genes identified from other species using BLASTX searches	68
Figure 11 Categorization of four different types of SNPs (a-d) that can be identified from the all catfish EST assembly, and examples of SNPs whose categories could not be determined due to the minor allele sequence from a given species is fewer than two (e)	74
Figure 12 JGI EST analysis pipeline	81

I. INTRODUCTION

Overview

Fishes are an extremely diverse group of vertebrates, including jawless fishes (hagfishes, lampreys), cartilaginous fishes (sharks, rays) and bony fishes (coelacanth, lungfishes and ray-finned fishes (Figure 1) [1]. Ray-finned fishes (actinopterygians) accounts for 95% of all existing fish species and more than 99.8% of ray-finned fishes are teleosts. Teleost fish accounts for half of the existing vertebrate species. Several teleost fish species are subjected the genetics and genomics studies with or approaching to have the whole genome sequence, including zebrafish, *Tetraodon*, fugu, medaka, and stickleback [2-3]. These species are widely used in the development, evolution, biomedical studies, which are considered as model species. For the species like Atlantic salmon, Rainbow trout, tilapia, along with channel catfish, are also subjected the genetics and genomics studies partially driven by economic motivation.

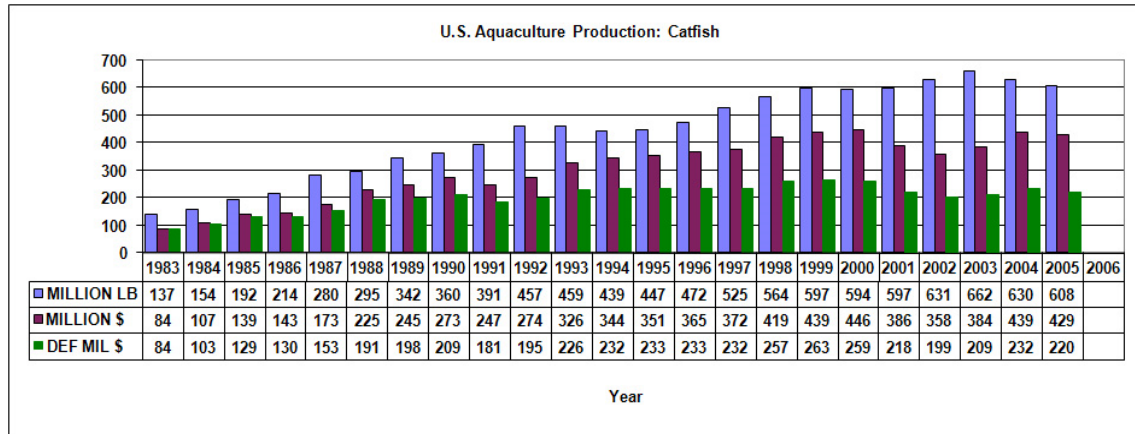
Figure 1 The fish lineage



*Volf 2005 Heredity

Catfish belongs to Siluriformes (orders) with the predominant barbels, which resemble a cat's whiskers in most catfish species. Catfish also belong to a super order Ostariophysi, including the Cypriniformes, Characiformes, Gonorynchiformes, and Gymnotiformes. Catfish is one the most important species in the aquaculture industry in the United States, accounting for over 60% of all US aquaculture production. Catfish aquaculture is also growing very fast in Southeast Asia. Channel catfish (*Ictalurus punctatus* Rafinesque) is the major cultured catfish species, accounting for the majority of commercial aquaculture production. A closely related species, blue catfish (*I. furcatus* Valenciennes), is also considered important because of its ability to produce hybrids with the channel catfish [4], which has the desirable characteristics of improved capture by growth rate, feed conversion efficiency, processing yields, uniformity in body conformation, catchability, dress-out percentage, resistance to some bacterial diseases, and tolerance to low concentrations of dissolved oxygen [5-12]. In terms of disease resistance, channel catfish is superior in resistance to columnaris disease (caused by *Flavobacterium columnare*), while blue catfish is superior in resistance to enteric septicemia of catfish (ESC, caused by *Edwardsiella ictaluri*) (Dunham et al., 1993a). ESC and columnaris are the two most severe diseases in catfish accounting for over 78% of the disease problems (NAHMS, 1997). In terms of processing yield, blue catfish is superior to channel catfish, providing 5-8% more fillet yield than channel catfish. This inter-specific system, therefore, provides a model system for analysis of major QTLs involved in disease resistance and processing yield [11]. Hybrid catfish produced by inter-specific hybridization of channel catfish x blue catfish is one of the best catfish used in aquaculture [13].

Figure 2 U.S. catfish production



*<http://www.msstate.edu/dept/crec/aquaprod.html>

Through the evolution history, catfish diverged from zebrafish 110-160 million years ago (MYA) [14]. Unlike zebrafish and *Tetraodon*, Channel catfish is non-model species, but it serves as an important species for the study of comparative immunology, reproductive physiology, and toxicology. The channel catfish immune system is among the best characterized of any fish species, with decades of research leading to identification and characterization of catfish immune genes, such as CC chemokines [15, 16], establishment of clonal, functionally distinct, lymphocyte cell lines [17], characterization of much of the machinery of catfish innate [18, 19], adaptive immunity and production of panels of specific monoclonal antibodies for detection of catfish immunocytes [20-22]. Therefore, genomic studies of an aquaculture fish species might provide new insight addressing genetic mechanisms of the unique traits in aquatic environments as well as genome evolution.

Rapid progress in catfish genomics has been made in the last decade. The framework genetic linkage maps, with over 400 type I (type I marker is gene associated marker) microsatellite and SNP markers and 400 AFLP markers, have been constructed [23, 24]. Genome repeat structure has been characterized and several novel repetitive elements, e.g. *Xba* elements, TC-1 like elements tip1, tip2 and tipnon, short interspersed elements (SINE) *mermaid* and *merman*, have been identified from catfish genome through BAC end sequencing [25-28]. More than 55,000 expressed sequence tags (ESTs) have been generated from various tissues and organs [29-33], and an ongoing large-scale EST project by the Joint Genome Institute of the Department of Energy will significantly expand the EST resources in both channel catfish and blue catfish [34]. Two Microarrays have also been developed and utilized to study genome-wide expression in catfish [18, 19, 30, 31]. Two bacterial artificial chromosome (BAC) libraries (CHORI212 and CCBL1) using different restriction endo-nucleases (*EcoR* I and *Hind* III) have been previously constructed and characterized, CCBL1 (7.2X coverage) constructed using DNA from a homozygous gynogenetic female, and CHORI-212 (10.6X coverage) constructed using DNA from a normal male catfish where the genomic DNA contains all autosomes and sex chromosomes, and the normal level of polymorphism [35, 36]. Two physical maps have been constructed based on these libraries [37, 38]. Over 60,000 BAC end sequences (BES) have been generated from the CHORI-212 library, which provides over 6500 type II (type II markers are derived from unknown sequence in the genome) microsatellite markers from BES possessing enough flanking sequences to design primers for polymorphism test. Over 2,000 polymorphic type II microsatellites have been identified and utilized for the density linkage map construction and integration of genetic linkage and physical map.

The catfish virtual physical map has been constructed based on the catfish physical map and approximately 60,000 BES by comparing with zebrafish genome sequences [39]. The integration of linkage map and physical map will provide a framework for whole genome sequencing and comparative genomic study between catfish and other teleost fish. The first catfish SNP chips have been designed, which proved the feasibility of the EST-derived SNP genotyping application by using Illumina genotyping technology in the catfish [40]. The design of high-density Illumina SNP chip, including over 10,000 gene-associated SNPs, is on the way and will be available in 2010 for the catfish research community. The whole genome sequencing has been proposed in 2009 by the catfish genome consortium.

Expressed sequence tag (EST) and cDNA library construction

An expressed sequence tag, or EST, is a short sub-sequence of a transcribed spliced nucleotide sequence (either protein-coding or not), which can be used to identify gene transcripts, and are instrumental in gene discovery and gene sequence determination [41]. Currently, EST sequencing is one of the most efficient ways for gene discovery. The identification of ESTs has proceeded very quickly, with approximately 60 million ESTs now available in public databases, including 8.1 million from human and 4.8 million from mouse, which accounting for 20% of the total dbEST database (e.g. GenBank 2009 Apr release, all species).

An EST is produced by one-shot sequencing of a cloned mRNA (i.e., sequencing several hundred base pairs from an end of a cDNA clone taken from a cDNA library). The resulting sequence is a relatively low quality fragment whose length is limited to

approximately 500 to 800 nucleotides by traditional Sanger sequencing technology. Because clones consist of DNA that is complementary to mRNA, the ESTs represent portions of expressed genes. They may be present in the database as either cDNA/mRNA sequence or as the reverse complement of the mRNA, the template strand. ESTs can be mapped to specific chromosome locations using physical mapping techniques, such as radiation hybrid mapping or fluorescent *in situ* hybridization (FISH) [42]. Alternatively, if the genome of the organism that the EST originated from has been sequenced, one can align the EST sequence to that genome, which can help solve alternative splicing issues widely discovered in vertebrates [43, 44]. The current understanding of the human set of genes includes the existence of thousands of genes based solely on EST evidence. In this respect, ESTs have become a tool to refine the predicted transcripts for those genes, which leads to prediction of their protein products, and eventually of their function. Moreover, the situation in which those ESTs are obtained (e.g. cancer) gives information on the conditions in which the corresponding gene is acting [45]. ESTs contain enough information to permit the design of precise probes for DNA microarrays that then can be used to determine the gene expression [46, 47].

The cDNA library construction starts with the cDNA synthesis of mRNAs isolated from the tissues. The traditional cDNA library construction requires several steps, starting with reverse transcription (1st strand synthesis) followed by second strand synthesis and ligation. Currently, the most popular method for construction of cDNA libraries is SMART™ (Switching Mechanism at 5' End of RNA Template), which incorporate the first strand and second strand synthesis together, without adaptor ligation. The presence of these known sequences is crucial for a number of downstream applications including

amplification, Rapid Amplification of cDNA Ends (RACE), and library construction.

While a wide variety of technologies can be employed to take advantage of these known sequences, the simplicity and efficiency of the single-step SMART process permits unparalleled sensitivity and ensures that full-length cDNA are generated and amplified.

Because cDNA synthesis starts from 3' end of mRNA, it is sensitive to interruptions caused by secondary structures in the template RNA. The reverse transcription terminates before transcribing the complete mRNA sequence, and the 5' ends of genes (especially with longer mRNA sequences) are usually underrepresented by conventional cDNAs synthesis methods. Since the terminal transferase activity (and subsequent SMART switching process) occurs preferentially at the 5' ends of eukaryotic mRNAs by adding the additional Cs at the end of first strand, which are complemented with 5' adaptor with 3'-GGG tails. If the reverse transcription was terminated before the completion of transcribing mRNA sequences, truncated cDNA products are not able to base pair with the 5' adaptor, and therefore, get lost in the next step of PCR amplification. The SMART kit used for cDNA synthesis is designed to preferentially enrich for full-length cDNA, which will greatly benefit the full-length cDNA identification and characterization.

The mRNA concentration of the genes are not uniformed, the most prevalent mRNA in a typical cells accounts for over 60% of the total message. The frequency of cDNA in the library will correspond to the mRNA frequency in the transcripts, so in order to improve the sequence coverage to identify all potential genes and maximize the sequencing efficiency to reduce the number of sequencing, the cDNA library need to be normalized, and subtraction of highly abundant expressed genes, such as actin, is usually used to improve the normalization efficiency [48]. The current and most popular

normalization method involve the re-association of the denatured DNA, degradation of double-strand fraction formed by abundant transcripts, and PCR amplification of the equalized SS-DNA fraction. The principle of this method is degradation of the double-strand fraction formed during re-association of cDNA using Duplex-Specific Nuclease (DSN) enzyme [49]. DSN displays a strong preference for cleaving double-strand DNA in both DNA-DNA and DNA-RNA hybrids, compared to SS-DNA and RNA, regardless of the sequence length. During the normalization, the subtraction was also utilized by adding the PCR products selected from the highly abundantly genes, which could improve the normalization efficiency. A well-normalized and subtracted cDNA library will greatly improve the sequencing efficiency and gene discovery rate.

ESTs and microarray development

EST serves as the basis of microarray development for gene expression profiling studies. There are two major microarray technology based on the construction and sample labeling. Spotted microarrays are constructed by spotting oligos or cDNAs on the slides directly using printing robot. *In situ* arrays are constructed by directly synthesizing oligos on the slides using photolithography. The cDNAs utilized for spotted microarray are directly coming from the cDNA libraries. Through the EST sequencing, the genes can be identified from each cDNA clones. The cDNAs can be extracted from the containing cDNA clones where genes are identified. Through the ESTs cluster analysis, the contiguous sequences can be used for the development of oligo arrays, including both spotted and *in situ* synthesized arrays.

Full-length cDNA identification and characterization

The sequencing from cDNA library constructed by the SMART system greatly improves the possibility to recover full-length cDNA. The large scale EST sequencing provides a platform for full-length cDNA isolation and characterization [50, 51]. The length of EST sequences is usually limited, by our definition; full-length cDNA should contain the start codon (ATG), open reading frame, stop codon, 3' untranslated region (UTR) and presence of poly (A) tail in the cDNA clones. Full-length cDNA are derived from high quality sequencing of full-length insert cDNA clones containing complete ORF and 3' UTR. Full-length cDNA is vital for the accurate assembly of EST sequences and predication of protein sequence. Full-length cDNA will greatly benefit the future expression profiling by deep sequencing. The characterization of full-length cDNA also provides a platform for the study of differential polyadenylation and splice variation, including exon skipping, intro retention, 5' and 3' alternative splicing. The access to the UTR information will help understanding the non-coding mRNA that is important in gene expression regulation.

Type I marker identification

EST sequencing is also one of the most efficient ways to identify type I markers polymorphic markers, including microsatellites and single nucleotide polymorphisms (SNPs). Markers can be categorized to type I and type II markers. Type I markers are gene-associated markers. Type II markers are developed from anonymous genomic regions. Type I marker is more useful, not only for the construction of gene-based linkage map, but also for the comparative studies between catfish and map rich species, such as zebrafish.

Microsatellites, or simple sequences repeats (SSR), are polymorphic loci present in nuclear DNA and organellar DNA, which are tandem repeat DNA sequences and usually consist of repeated units of 1-4 base pairs in length. Microsatellites are highly-abundant in various eukaryotic genomes including aquaculture species. Through the genomic sequencing survey, 2.6% of microsatellites were identified from channel catfish [28]. Generally di-nucleotide repeats are the most abundant type of microsatellites, followed by tri- and tetra-nucleotide repeats. Over 60% of microsatellites are di-nucleotide repeats in the channel catfish [28]. Microsatellites are generally evenly distributed in the genome on all chromosome regions, including gene coding regions (exon), intron, and non-gene regions [52, 53]. The presence of microsatellites (non-codon repeats) in the gene coding regions can cause the frameshift mutation. Microsatellites are predominantly presented in the non-coding regions [54]. Microsatellite markers development requires sufficient sequence information. Generation of 30,000 EST and 20,000 BES sequences allow the identification of thousands of type I and type II microsatellite markers [28, 33]. JGI EST sequencing project will also provide tens thousands of type I microsatellite markers.

They are typically neutral, co-dominant and are used as molecular markers which have wide-ranging applications in the field of genetics, including genetics linkage mapping and population studies [55]. Microsatellites can also be used to study gene dosage (looking for duplications or deletions of a particular genetic region). Length variation in microsatellites is generated by two mechanisms: 1) the DNA polymerase slippage and 2) unequal crossover between the homolog chromosomes. When the DNA replicates, the polymerase loses track of its place, and either subtracts or adds repeat units. The result is that the new strand has a different number of repeats than the parent

strand. This is thought to explain small changes in numbers of repeats (adding or subtracting one or just a few repeats). During meiosis, the crossing-over happened between the homologous chromosomes. However, the cross-overs between the chromosomes are not equal every time, so unequal crossing-over could cause the microsatellites repeat units changes. This is thought to explain more drastic changes in numbers of repeats.

Single nucleotide polymorphisms (SNPs) are DNA sequence variation occurring at the single nucleotide locus in the genome, which can differ between two allele sequences inherited from parents, or among different individuals. In most case, SNP have only two alleles [56]. There are two types of mutation causing SNP, transition and transversion. Transition includes A-G and C-T SNP, and transversion includes A-C, A-T, C-G and T-G SNP. SNPs can be characterized as synonymous SNP and non-synonymous SNP. Synonymous SNPs can be located within the coding regions of genes (exon), non-coding regions of genes (intron), or inter-genic regions between genes in the chromosome. However, the SNPs within the coding regions of genes may not change the amino acid in the peptide sequences because of the degeneracy of the genetic code. If the presentation of SNPs in the coding regions of genes causes the change of peptide sequence, it will be defined as non-synonymous SNPs.

SNPs are one of the most abundant types of genetic variation. SNPs are estimated to occur once every 1.3 kb in humans when any two chromosomes are compared [57-59] while their frequencies have been estimated to be higher in other organisms, 3.42 SNPs per 100 bp in the medaka [3, 60]. This would make it possible to construct genetic maps with extremely high marker densities allowing identification of haplotype segments using

SNPs, especially for the species with a draft genome sequence [61]. In addition, SNPs offer several other advantages over other molecular markers. First, SNPs are the most fundamental causing of mutation in the genetic variation, especially in the protein coding genes, and their mapping would provide potential for the identification of the “causing” SNPs as well as the “tightly associated” SNPs with specific and complex traits, which will greatly benefit for the evaluation of personal risk, or whole genome selection in the animal and crop breeding [62-65]. Second, many technologies have been developed to genotype SNPs cost-effectively in an automated fashion, such as Illumina GoldenGate Assay [66-68]. Third, SNPs are sequence-tagged markers with co-dominant inheritance, suitable for comparative genome analysis [60, 69]; and finally, SNPs are highly stable genetic markers compared to tandem repeat markers where the high mutation rates can confound genetic analysis in populations [70, 71].

Identification and application of EST-derived SNP marker

Unlike other molecular markers, such as RAPD, AFLP, and microsatellites, identification of SNP marker requires massive sequencing effort, which could not be afforded by most research scientists in the aquaculture field because of limitation of funding and labors resources. Therefore SNP marker development and application were quite rare in aquaculture species in the last decade.

The first SNP mining from human EST resources and genotyping application was reported by in 1999. From a set of ESTs derived from 19 different cDNA libraries, 300,000 distinct sequences were assumed and 850 mismatches were identified from contiguous EST data sets (candidate SNP sites), without *de novo* sequencing [72]. Since

then, ESTs have been used as a major resource for the SNP identification, and a large number of SNP markers have identified. Through years, the genome resources, such as Expressed sequence tags (EST) and Genome sequence survey (GSS) resources have been harvested in the aquaculture species, such as Atlantic salmon, Rainbow Trout, tilapia, catfish, which make it reality for the SNP marker development.

Along with the development of high-throughput automated fashion SNP genotyping technology, the genotyping price was significantly reduced for each sample [73]. Several SNP studies have been conducted in aquaculture species which approved the feasibility of application of EST-derived SNP identification and genotyping [74-76]. A large number genomics resources have been successfully generated, including ESTs and GSS in several the major aquaculture species in the United States, such as salmon, catfish, rainbow trout. These genomic resources provide a platform for large scale marker development. ESTs served as a great resource for SNP development is the early SNP genotyping stages. The cDNA libraries were usually constructed with a variety of individuals. ESTs are the single pass sequence generated from cDNA sequencing. The ESTs obtained from different individuals, assembly of overlapping ESTs for the same region can lead to higher opportunity of identification of SNPs. ESTs-derived SNPs are usually associated with genes, which can help identify directly “associated” SNPs for complex traits. Several softwares and methods have been developed to identify putative SNPs from ESTs dataset. Polybayes is most popular software for SNP identification [77]. However, EST sequence quality scores and trace files are required for SNP identification using Polybayes. AutoSNP and QualitySNP were developed for the SNP identification without sequence quality scores or trace files, which are suitable for the SNP identification in most

aquaculture species [78, 79]. A significant risk in such an analysis is that many sequence variations are the result of poor quality sequence data typically found in single-pass EST data sets, especially using ESTs without sequence quality scores. The putative SNP identification using ESTs could lead to the identification of pseudo-SNPs, leading to subsequently great efforts and expense. In order to reduce the rate of pseudo-SNPs resulted from the sequencing errors; the development of a strategy for rapid and reliable identification of EST-derived SNPs is urgent and necessary.

Comparative studies with other model fishes based on EST

Comparative genomics is the study of relationships between the genomes of different species or strains. Comparative genomics attempts to take advantage of the information provided by the well-known studied species to understand the function and evolutionary processes that act on research species. Gene finding is an important application of comparative genomics, as is discovery of new, non-coding functional elements of the genome. For the EST sequences analysis, comparative genomics is a very powerful tool for the gene discovery and identification.

Except the gene discovery and identification, it would allow comparing the gene structures on the chromosomes, which could help understand the evolution history among the teleosts, such as zebrafish, *Tetraodon* and medaka. It also provides a feasibility to conduct the systematic and large-scale phylogenetic analysis of duplicated genes, thereby reducing the complexity for gene mapping with duplicated genes as well as setting a foundation for evolutionary and comparative genome analysis of duplicated genes, particularly for teleost-specific gene duplications.

Catfish genomics research “bottleneck”

Large scale ESTs can help identify genes and understand gene structures with future whole genome sequences, and it also can provide a large number of type I markers including microsatellites and SNPs, which are important for the study of complex traits and genome selection. However, the number of catfish ESTs is quite limited with less than 55,000 in the GenBank. Catfish genomic research is at a stage where the availability of a large number of ESTs is essential. Much of the catfish genome research in the last 5-10 years has built up into a research “bottleneck” because of the lack of a large number of ESTs. The limitation of the ESTs restricts the large scale of gene identification and marker development, which is important for the functional genomics study. In order to solve this “bottleneck”, we conducted this JGI catfish EST sequencing project.

References

1. Nelson JS: **Fishes of the World** 3rd edn. John Wiley and Sons: 1994, New York.
2. Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, Bouneau L, Fischer C, Ozouf-Costaz C, Bernot A: **Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype.** *Nature*. 2004, 431:946-957.
3. Kasahara M, Naruse K, Sasaki S, Nakatani Y, Qu W, Ahsan B, Yamada T, Nagayasu Y, Doi K, Kasai Y *et al*: **The medaka draft genome and insights into vertebrate genome evolution.** *Nature* 2007, **447**:714-719.
4. Chatakondi NG, Yant DR, Dunham RA: **Commercial production and performance evaluation of channel catfish, *Ictalurus punctatus* female x blue catfish, *Ictalurus furcatus* male F-1 hybrids.** *Aquaculture* 2005, **247**:8.
5. Giudice JJ, **Growth of a blue x channel catfish hybrid as compared to its parent species.** *Progressive Fish-Culturist* 1966, **26**:142–145.
6. Yant DR, Smitherman RO, Green OL: **Production of hybrid (blue x channel) catfish and channel catfish in ponds.** *Proceedings Annual Conference Southeastern Association of Game and Fish Commissioners* 1976, **29**:82–86.
7. Tave D, McGinty AS, Chappell JA, Smitherman RO: **Relatively harvestability by angling of blue catfish, and their reciprocal hybrids.** *North American Journal of Fisheries Management* 1981, **1**:73–76.
8. Dunham RA, Benchankan M, Smitherman RO, Chappell JA: **Correlations among morphometric traits of fingerling catfishes and the relationship to**

- dressing percentage at harvest.** *Journal of the World Mariculture Society* 1983, **14**:668–675.
9. Dunham RA, Smitherman RO, Webber C: **Relative tolerance of channel×blue hybrid and channel catfish to low oxygen concentrations.** *Progressive Fish-Culturist* 1983, **45**:55–57.
 10. Dunham RA, Joyce JA, Bondari K, Malvestuto SP: **Evaluation of body conformation, composition and density as traits for indirect selection for dress-out percentage of channel catfish.** *Progressive Fish-Culturist* 1985, **47**:169–175.
 11. Dunham RA, Hyde C, Masser M, Plumb JA, Smitherman RO, Perez R, Ramboux AC: **Comparison of culture traits of channel catfish, *Ictalurus punctatus*, and blue catfish, *I. furcatus*.** *Journal of Applied Aquaculture* 1993, **3**:257–267.
 12. Dunham RA, Lambert DM, Argue BJ, Ligeon C, Yant DR, Liu Z: **Comparison of manual stripping and pen spawning for production of channel catfish×blue catfish hybrids and aquarium spawning of channel catfish.** *North American Journal of Aquaculture* 2000, **62**:260–265.
 13. Argue BJ, Liu Z, Dunham RA: **Dress-out and fillet yields of channel catfish, *Ictalurus punctatus*, blue catfish, *Ictalurus furcatus*, and their F1, F2 and backcross hybrids.** *Aquaculture* 2003, **228**:81-90
 14. Volff JN: **Genome evolution and biodiversity in teleost fish.** *Heredity* 2005, **94**:280-294.
 15. Bao B, Peatman E, Peng X, Baoprasertkul P, Wang G, Liu Z: **Characterization of 23 CC chemokine genes and analysis of their expression in channel catfish**

- (*Ictalurus punctatus*). *Developmental and comparative immunology* 2006, **30**:783-796.
16. Peatman E, Liu Z: **Evolution of CC chemokines in teleost fish: a case study in gene duplication and implications for immune diversity.** *Immunogenetics* 2007, **59**:613-623.
 17. Miller N, Wilson M, Bengten E, Stuge T, Warr G, Clem W: **Functional and molecular characterization of teleost leukocytes.** *Immunological reviews* 1998, **166**:187-197.
 18. Peatman E, Baoprasertkul P, Terhune J, Xu P, Nandi S, Kucuktas H, Li P, Wang S, Somridhivej B, Dunham R *et al*: **Expression analysis of the acute phase response in channel catfish (*Ictalurus punctatus*) after infection with a Gram-negative bacterium.** *Developmental and comparative immunology* 2007, **31**:1183-1196.
 19. Peatman E, Terhune J, Baoprasertkul P, Xu P, Nandi S, Wang S, Somridhivej B, Kucuktas H, Li P, Dunham R *et al*: **Microarray analysis of gene expression in the blue catfish liver reveals early activation of the MHC class I pathway after infection with *Edwardsiella ictaluri*.** *Molecular immunology* 2008, **45**:553-566.
 20. Bengten E, Clem LW, Miller NW, Warr GW, Wilson M: **Channel catfish immunoglobulins: repertoire and expression.** *Developmental and comparative immunology* 2006, **30**:77-92.

21. Evenhuis J, Bengten E, Snell C, Quiniou SM, Miller NW, Wilson M: **Characterization of additional novel immune type receptors in channel catfish, *Ictalurus punctatus*. *Immunogenetics* 2007, **59**:661-671.**
22. Sahoo M, Edholm ES, Stafford JL, Bengten E, Miller NW, Wilson M: **B cell receptor accessory molecules in the channel catfish, *Ictalurus punctatus*. *Developmental and comparative immunology* 2008, **32**:1385-1397.**
23. Liu Z, Karsi A, Li P, Cao D, Dunham R: **An AFLP-based genetic linkage map of channel catfish (*Ictalurus punctatus*) constructed by using an interspecific hybrid resource family. *Genetics* 2003, **165**:687-694.**
24. Waldbieser GC, Bosworth BG, Nonneman DJ, Wolters WR: **A microsatellite-based genetic linkage map for channel catfish, *Ictalurus punctatus*. *Genetics* 2001, **158**:727-734.**
25. Kim S, Karsi A, Dunham RA, Liu Z: **The skeletal muscle alpha-actin gene of channel catfish (*Ictalurus punctatus*) and its association with piscine specific SINE elements. *Gene* 2000, **252**:173-181.**
26. Liu Z, Karsi A, Dunham RA: **Development of Polymorphic EST Markers Suitable for Genetic Linkage Mapping of Catfish. *Marine biotechnology* (New York, NY 1999, **1**:437-0447.**
27. Nandi S, Peatman E, Xu P, Wang S, Li P, Liu Z: **Repeat structure of the catfish genome: a genomic and transcriptomic assessment of Tc1-like transposon elements in channel catfish (*Ictalurus punctatus*). *Genetica* 2007, **131**:81-90.**
28. Xu P, Wang S, Liu L, Peatman E, Somridhivej B, Thimmapuram J, Gong G, Liu Z: **Channel catfish BAC-end sequences for marker development and**

- assessment of syntenic conservation with other fish species. *Animal genetics* 2006, **37**:321-326.
29. Cao D, Kocabas A, Ju Z, Karsi A, Li P, Patterson A, Liu Z: **Transcriptome of channel catfish (*Ictalurus punctatus*): initial analysis of genes and expression profiles of the head kidney.** *Animal genetics* 2001, **32**:169-188.
 30. Ju Z, Karsi A, Kocabas A, Patterson A, Li P, Cao D, Dunham R, Liu Z: **Transcriptome analysis of channel catfish (*Ictalurus punctatus*): genes and expression profile from the brain.** *Gene* 2000, **261**:373-382.
 31. Karsi A, Cao D, Li P, Patterson A, Kocabas A, Feng J, Ju Z, Mickett KD, Liu Z: **Transcriptome analysis of channel catfish (*Ictalurus punctatus*): initial analysis of gene expression and microsatellite-containing cDNAs in the skin.** *Gene* 2002, **285**:157-168.
 32. Kocabas AM, Kucuktas H, Dunham RA, Liu Z: **Molecular characterization and differential expression of the myostatin gene in channel catfish (*Ictalurus punctatus*).** *Biochimica et biophysica acta* 2002, **1575**:99-107.
 33. Li P, Peatman E, Wang S, Feng J, He C, Baoprasertkul P, Xu P, Kucuktas H, Nandi S, Somridhivej B *et al*: **Towards the ictalurid catfish transcriptome: generation and analysis of 31,215 catfish ESTs.** *BMC genomics* 2007, **8**:177.
 34. He C, Chen L, Simmons M, Li P, Kim S, Liu ZJ: **Putative SNP discovery in interspecific hybrids of catfish by comparative EST analysis.** *Animal genetics* 2003, **34**:445-448.

35. Quiniou SM, Katagiri T, Miller NW, Wilson M, Wolters WR, Waldbieser GC:
Construction and characterization of a BAC library from a gynogenetic channel catfish *Ictalurus punctatus*. *Genet Sel Evol* 2003, **35:673-683.**
36. Wang S, Xu P, Thorsen J, Zhu B, de Jong PJ, Waldbieser G, Kucuktas H, Liu Z:
Characterization of a BAC library from channel catfish *Ictalurus punctatus*: indications of high levels of chromosomal reshuffling among teleost genomes. *Marine biotechnology (New York, NY)* 2007, **9:701-711.**
37. Quiniou SM, Waldbieser GC, Duke MV: **A first generation BAC-based physical map of the channel catfish genome. *BMC genomics* 2007, **8**:40.**
38. Xu P, Wang S, Liu L, Thorsen J, Kucuktas H, Liu Z: **A BAC-based physical map of the channel catfish genome. *Genomics* 2007, **90**:380-388.**
39. Liu H, Jiang Y, Wang S, Ninwichian P, Somridhivej B, Xu P, Kucuktas H, Liu Z:
Comparative Genome Analysis Using BAC End Sequences of Catfish. *BMC Genomics* 2009, in review.
40. Wang S, Sha Z, Sonstegard TS, Liu H, Xu P, Somridhivej B, Peatman E, Kucuktas H, Liu Z: **Quality assessment parameters for EST-derived SNPs from catfish. *BMC genomics* 2008, **9**:450.**
41. Adams MD, Soares MB, Kerlavage AR, Fields C, Venter JC: **Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nature genetics* 1993, **4**:373-380.**
42. Horelli-Kuitunen N, Aaltonen J, Yaspo ML, Eeva M, Wessman M, Peltonen L, Palotie A: **Mapping ESTs by fiber-FISH. *Genome research* 1999, **9**:62-71.**

43. Mott R: **EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA.** *Comput Appl Biosci* 1997, **13**:477-478.
44. Wolfsberg TG, Landsman D: **A comparison of expressed sequence tags (ESTs) to human genomic sequences.** *Nucleic acids research* 1997, **25**:1626-1632.
45. Gress TM, Muller-Pillasch F, Geng M, Zimmerhackl F, Zehetner G, Friess H, Buchler M, Adler G, Lehrach H: **A pancreatic cancer-specific expression profile.** *Oncogene* 1996, **13**:1819-1830.
46. Khan J, Saal LH, Bittner ML, Chen Y, Trent JM, Meltzer PS: **Expression profiling in cancer using cDNA microarrays.** *Electrophoresis* 1999, **20**:223-229.
47. Carulli JP, Artinger M, Swain PM, Root CD, Chee L, Tulig C, Guerin J, Osborne M, Stein G, Lian J *et al*: **High throughput analysis of differential gene expression.** *J Cell Biochem Suppl* 1998, **30-31**:286-296.
48. Bonaldo MF, Lennon G, Soares MB: **Normalization and subtraction: two approaches to facilitate gene discovery.** *Genome research* 1996, **6**:791-806.
49. Shagin DA, Rebrikov DV, Kozhemyako VB, Altshuler IM, Shcheglov AS, Zhulidov PA, Bogdanova EA, Staroverov DB, Rasskazov VA, Lukyanov S: **A novel method for SNP detection using a new duplex-specific nuclease from crab hepatopancreas.** *Genome Res* 2002, **12**:1935-1942.
50. Min XJ, Butler G, Storms R, Tsang A: **TargetIdentifier: a webserver for identifying full-length cDNAs from EST sequences.** *Nucleic acids research* 2005, **33**:W669-672.

51. Baross A, Butterfield YS, Coughlin SM, Zeng T, Griffith M, Griffith OL, Petrescu AS, Smailus DE, Khattri J, McDonald HL *et al*: **Systematic recovery and analysis of full-ORF human cDNA clones.** *Genome research* 2004, **14**:2083-2092.
52. Liu Z, Li P, Kocabas A, Karsi A, Ju Z: **Microsatellite-containing genes from the channel catfish brain: evidence of trinucleotide repeat expansion in the coding region of nucleotide excision repair gene RAD23B.** *Biochem Biophys Res Commun.* 2001, **289**:317-24.
53. Toth G, Gaspari Z, and Jurka J: **Microsatellites in different eukaryotic genomes: survey and analysis.** *Genomes Research*, **10**:967-981
54. Metzgar D, Bytof J, Wills C: **Selection against frameshift mutations limits microsatellite expansion in coding DNA.** *Genome Research.* 2000, **10**:72-80
55. Todd JA: **La carte des microsatellites est arrivee! [The map of microsatellites has arrived!].** *Human molecular genetics* 1992, **1**:663-666.
56. Krawczak M: **Informativity assessment for biallelic single nucleotide polymorphisms.** *Electrophoresis* 1999, **20**:1676-1681.
57. Cooper DN, Smith BA, Cooke HJ, Niemann S, Schmidtke J: **An estimate of unique DNA sequence heterozygosity in the human genome.** *Hum Genet* 1985, **69**:201-205.
58. Li WH, Sadler LA: **Low nucleotide diversity in man.** *Genetics* 1991, **129**:513-523.

59. Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, Moulin DS, Clegg JB: **Archaic African and Asian lineages in the genetic ancestry of modern humans.** *Am J Hum Genet* 1997, **60**:772-789.
60. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ, Zody MC *et al*: **Genome sequence, comparative analysis and haplotype structure of the domestic dog.** *Nature* 2005, **438**:803-819.
61. Salisbury BA, Pungliya M, Choi JY, Jiang R, Sun XJ, Stephens JC: **SNP and haplotype variation in the human genome.** *Mutat Res* 2003, **526**:53 - 61.
62. Butcher LM, Davis OS, Craig IW, Plomin R: **Genome-wide quantitative trait locus association scan of general cognitive ability using pooled DNA and 500K single nucleotide polymorphism microarrays.** *Genes Brain Behav* 2007.
63. Kiyohara C, Yoshimasu K: **Genetic polymorphisms in the nucleotide excision repair pathway and lung cancer risk: a meta-analysis.** *Int J Med Sci* 2007, **4**:59 - 71.
64. Lazarus R, Vercelli D, Palmer LJ, Klimecki WJ, Silverman EK, Richter B, Riva A, Ramoni M, Martinez FD, Weiss ST *et al*: **Single nucleotide polymorphisms in innate immunity genes: abundant variation and potential role in complex human disease.** *Immunol Rev* 2002, **190**:9 - 25.
65. Rafalski A: **Applications of single nucleotide polymorphisms in crop genetics.** *Curr Opin Plant Biol* 2002, **5**:94-100.
66. Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS: **SNP discovery via 454 transcriptome sequencing.** *Plant J* 2007, **51**:910-918.

67. Fan JB, Oliphant A, Shen R, Kermani BG, Garcia F, Gunderson KL, Hansen M, Steemers F, Butler SL, Deloukas P *et al*: **Highly parallel SNP genotyping.** *Cold Spring Harb Symp Quant Biol* 2003, **68**:69 - 78.
68. Shen R, Fan JB, Campbell D, Chang W, Chen J, Doucet D, Yeakley J, Bibikova M, Wickham Garcia E, McBride C *et al*: **High-throughput SNP genotyping on universal bead arrays.** *Mutat Res* 2005, **573**:70 - 82.
69. Moreno-Vazquez S, Ochoa OE, Faber N, Chao S, Jacobs JM, Maisonneuve B, Kesseli RV, Michelmore RW: **SNP-based codominant markers for a recessive gene conferring resistance to corky root rot (*Rhizomonas suberifaciens*) in lettuce (*Lactuca sativa*).** *Genome* 2003, **46**:1059-1069.
70. Hastbacka J, de la Chapelle A, Kaitila I, Sistonen P, Weaver A, Lander E: **Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland.** *Nature genetics* 1992, **2**:204-211.
71. Marshall B, Leelayuwat C, Degli-Esposti MA, Pinelli M, Abraham LJ, Dawkins RL: **New major histocompatibility complex genes.** *Hum Immunol* 1993, **38**:24-29.
72. Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, Donaldson MA, Nickerson DA, Boyce-Jacino M: **Mining SNPs from EST databases.** *Genome Res* 1999, **9**(2):167-174.
73. Oliphant A, Barker DL, Stuelpnagel JR, Chee MS: **Beadarray technology: Enabling an accurate, cost-effective approach to high-throughput genotyping.** 2002, *Biotechniques*, Suppl, 56-58, 60-51.

74. Hayes B, Laerdahl JK, Lien S, Moen T, Berg P, Hindar K, Davidson WS, Koop BF, Adzhubei A, Hoyheim B: **An extensive resource of single nucleotide polymorphism markers associated with Atlantic salmon (*Salmo salar*) expressed sequences.** *Aquaculture* 2007, **265**(1-4):82-90.
75. Moen T, Hayes B, Baranski M, Berg P, Kjøglum S, Koop B, Davidson W, Omholt S, Lien S: **A linkage map of the Atlantic salmon (*Salmo salar*) based on EST-derived SNP markers.** *BMC Genomics* 2008, **9**(1):223.
76. Wang S, Sha Z, Sonstegard TS, Liu H, Xu P, Somridhivej B, Peatman E, Kucuktas H, Liu Z: **Quality assessment parameters for EST-derived SNPs from catfish.** *BMC genomics* 2008, **9**:450.
77. Marth GT, Korf I, Yandell MD, Yeh RT, Gu Z, Zakeri H, Stitzel NO, Hillier L, Kwok PY, Gish WR: **A general approach to single-nucleotide polymorphism discovery.** *Nat Genet.* 1999 **23**:452-456
78. Barker G, Batley J, H OS, Edwards KJ, Edwards D: **Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP.** *Bioinformatics* 2003, **19**:421-422.
79. Tang J, Vosman B, Voorrips RE, van der Linden CG, Leunissen JA: **QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species.** *BMC Bioinformatics.* 2006, **7**:438.

II. RESEARCH OBJECTIVES

Over 430,000 EST sequences were generated at Joint Genome Institute (JGI). The total available catfish EST sequences were approximately 500,000 now, which are ready for transcriptome analysis and large-scale marker development.

My dissertation was focused on the following objectives:

1. To establish quality assessment parameters for EST-derived SNP markers
 - 1) Selection of 384 SNPs from catfish GenBank EST dataset assembly.
 - 2) Validation of 384 SNPs using 192 samples.
 - 3) Evaluate the quality parameters for EST-derived SNPs

2. To analyze catfish transcriptome based on 500,000 EST sequences
 - 1) Assembly of blue catfish, channel catfish, and all catfish EST sequences
 - 2) Gene discovery, identification, and annotation after the EST assembly.
 - 3) Full length cDNA prediction and characterization
 - 4) Large scale type I gene associated marker development

III. QUALITY ASSESSMENT PARAMETERS FOR EST-DERIVED SNPS FROM CATFISH

Abstract

Background: SNPs are abundant, co-dominantly inherited, and sequence-tagged markers.

They are highly adaptable to large-scale automated genotyping, and therefore, are most suitable for association studies and applicable to comparative genome analysis.

However, discovery of SNPs requires genome-sequencing efforts either through whole genome sequencing or deep sequencing of reduced representation libraries. Such genome resources are not yet available for many species, including catfish. A large resource of ESTs is now available in catfish, allowing identification of large number of SNPs.

However, the reliability of EST-derived SNPs are relatively low because of sequencing errors. Thus, this project was designed to answer some of the questions relevant to quality assessment of EST-derived SNPs.

Results: Two factors were found to be most significant for validation of EST-derived SNPs: the contig size (i.e., number of sequences in the contig) and the minor allele sequence frequency. The larger the contigs were, the greater the validation rate (although the validation rate was reasonably high when the contigs contain four or more EST sequences) along with the minor allele sequence being represented at least twice in the contigs. Sequence quality surrounding the SNP under examination is also crucially important. PCR extension appeared to be limited to a very short distance, prohibiting successful genotyping when an intron was present.

Conclusions: Stringent quality assessment measures should be used when working with EST-derived SNPs. In particular, contigs containing four or more ESTs should be used

and the minor allele sequence should be represented at least twice. Genotyping primers should be designed from a single exon as to completely avoiding introns. Application of such quality assessment measures, along with large resources of ESTs, should provide effective means for SNP identification in species where genome sequence resources are lacking.

Background

Most performance traits of agricultural relevance are complex in that they are governed by multiple genes. Due to the large number of genes underlying a single trait and their complex interactions, direct genetic analysis of such traits has been difficult. In the past decade, genetic mapping has demonstrated great promise for the analysis of complex traits. In particular, wide applications of microsatellite markers in animal genome studies have allowed major progress in understanding of genes underlying performance traits [1, 2]. However, as larger genome datasets have become available, it is clear that microsatellites are not sufficiently dense to provide the genome coverage necessary for the dissection of many of the highly complex traits such as disease resistance, feed conversion efficiency, growth, and carcass traits [1]. In addition, large-scale automated genotyping of microsatellites has not been possible. Recently, much excitement was generated with the ability to analyze complex traits with new types of polymorphic markers, with efforts shifting to approaches such as single nucleotide polymorphisms (SNPs). SNPs are one of the most abundant types of genetic variation. In addition, SNPs provide several other advantages over other molecular markers, 1) the “tightly-associated” SNPs with specific and complex traits [3-6]; 2) automated SNPs genotyping with cost-effective [7-9].

In most cases, genome-wide SNP discovery has relied on the availability of a draft genome sequence, where SNPs can be detected during sequence assembly from the two chromosomes present in a diploid organism. This approach was initially feasible only for humans and model species. However, as the cost of genome sequencing decreases, now draft genome sequences have become available for several agriculturally important

species including cow, chickens, horses, and soon swine and tilapia. However, for most aquaculture species, it may take some time for the generation of entire genome draft sequences. Facing this, alternative approaches must be sought. It was able to identify a large number of SNPs from EST resources in Atlantic salmon, and it was recently demonstrated mapping of EST-derived SNPs to genetic linkage map [10, 11]. Their pioneering work with an aquaculture species set a great model for use of ESTs for the identification of SNPs, especially in non-model species [10-12]. In addition, BAC end sequences (BES) can also serve as sources for the identification of SNPs, and the combination of EST and BES could improve the SNP discovery accuracy comparing using only EST sequences [13].

Over 400,000 ESTs have been generated by the Joint Genome Institute of the Department of Energy 2008. Such EST sequences will provide an enormous resource for SNP identification. However, as most researchers have experienced, identification of SNPs using ESTs is not without problems. The most frequent is the high rate of sequencing errors, which can lead to the identification of pseudo-SNPs with subsequently great efforts and expense. Thus, the objective of this project was to develop a strategy for rapid and reliable identification and evaluation of EST-derived SNPs qualities to reduce the rate of pseudo-SNPs resulted from sequence errors typically found in single-pass EST datasets. This is especially important for those ESTs deposited in NCBI in other species, where sequence trace files may or may not be available. This pilot study was designed at 2007 before the releasing of JGI EST sequences data, so all the available catfish ESTs in the GenBank by April, 2007 was used in this study.

Results

Sequence Assembly

A total of 54,960 catfish ESTs available from GenBank including 44,437 ESTs from channel catfish and 10,523 ESTs from blue catfish were subjected to cluster analysis to identify putative SNPs. The contig assembly resulted in 5,670 contigs with an average size of 5.5 sequences per contig and an average length of 1,001 bp per contig. The assembly included 3,003 contigs with 2 ESTs, 980 contigs with 3 ESTs, and 1,687 contigs with 4 or more ESTs (Table 1).

Table 1. Summary of the EST Assembly

Number of sequences for assembly	54,960
blue catfish	10,523
channel catfish	44,437
Number of contigs	5,670
Number of singletons	23,598
Number of putative transcripts	29,268
Average contig size	5.5
Average contig length (bp)	1,001
No of contig with:	
2 ESTs	3,003
3 ESTs	980
4 ESTs	468
5 ESTs	263
6-10 ESTs	469
11-20 ESTs	246
21-30 ESTs	95
31-50 ESTs	72
>50 ESTs	74

Putative SNP discovery

Among 5,670 contigs, SNPs were detected in 4,387 contigs. The vast majority (73%) of the SNPs were identified from contigs with 2-3 sequences, the remaining SNPs were identified from contigs with 4 or more sequences (Table 2).

Table 2. Initial identification of SNPs as detected by AutoSNP software

No. of Sequences in each contig	No. of contigs with SNPs	No. of total SNPs	Total Consensus Length (bp)	SNP frequency (per 100 bp)
2	2,488	15,220	2,253,452	0.68
3	928	9,314	914,950	1.02
4	458	6,423	506,023	1.27
5	98	361	104,164	0.35
6-10	168	538	179,846	0.30
11-20	69	246	72,058	0.34
21-30	49	220	56,804	0.39
31-50	58	317	69,615	0.46
>50	71	955	93,065	1.03
Total	4,387	33,594	4,249,977	0.79*

*Average SNP frequency per 100 bp.

A total 33,594 SNPs were identified from the 4,387 contigs, with an average of 0.79 SNPs per 100 base pair. The putative SNP frequencies varied greatly among contigs of different sizes, ranging from 0.3 to 1.27 SNPs per 100 base pairs. It was apparent that the putative SNP frequency was greater within contigs containing fewer ESTs, an indication of significant sequence errors in contigs of 2 sequences (0.68 SNP per 100 bp), 3 sequences (1.02 SNPs per 100 bp), and 4 sequences (1.27 SNP per 100 bp). Clearly, this is also related to the parameters used in the AutoSNP software where any sequence variation is defined as a SNP in contigs of 2 sequences (1:1), 3 sequences (1:2), and 4 sequences (1:3 and 2:2), whereas the minor sequence allele must be at least twice with contigs of 5-6 sequences, at least 3 times with 7-8 sequences, etc. This observation, while within expectation, strongly demands validation of SNPs identified from EST sequences, especially from contigs with low numbers of sequences.

Validation of SNPs

To validate the putative SNPs identified from the ESTs, genotyping using the Illumina Bead Arrays was conducted with 192 fish, including 21 fish each from three strains of domestic catfish, 21 fish each from three wild populations collected from different watersheds, and 66 fish from the inter-specific mapping panel. Of the 266 successful genotyped SNPs, 156 (58.6%) were polymorphic among these 192 individuals. Of the 156 SNPs that were polymorphic, 49-97 were polymorphic in three domestic and wild catfish strain (Figure 3).

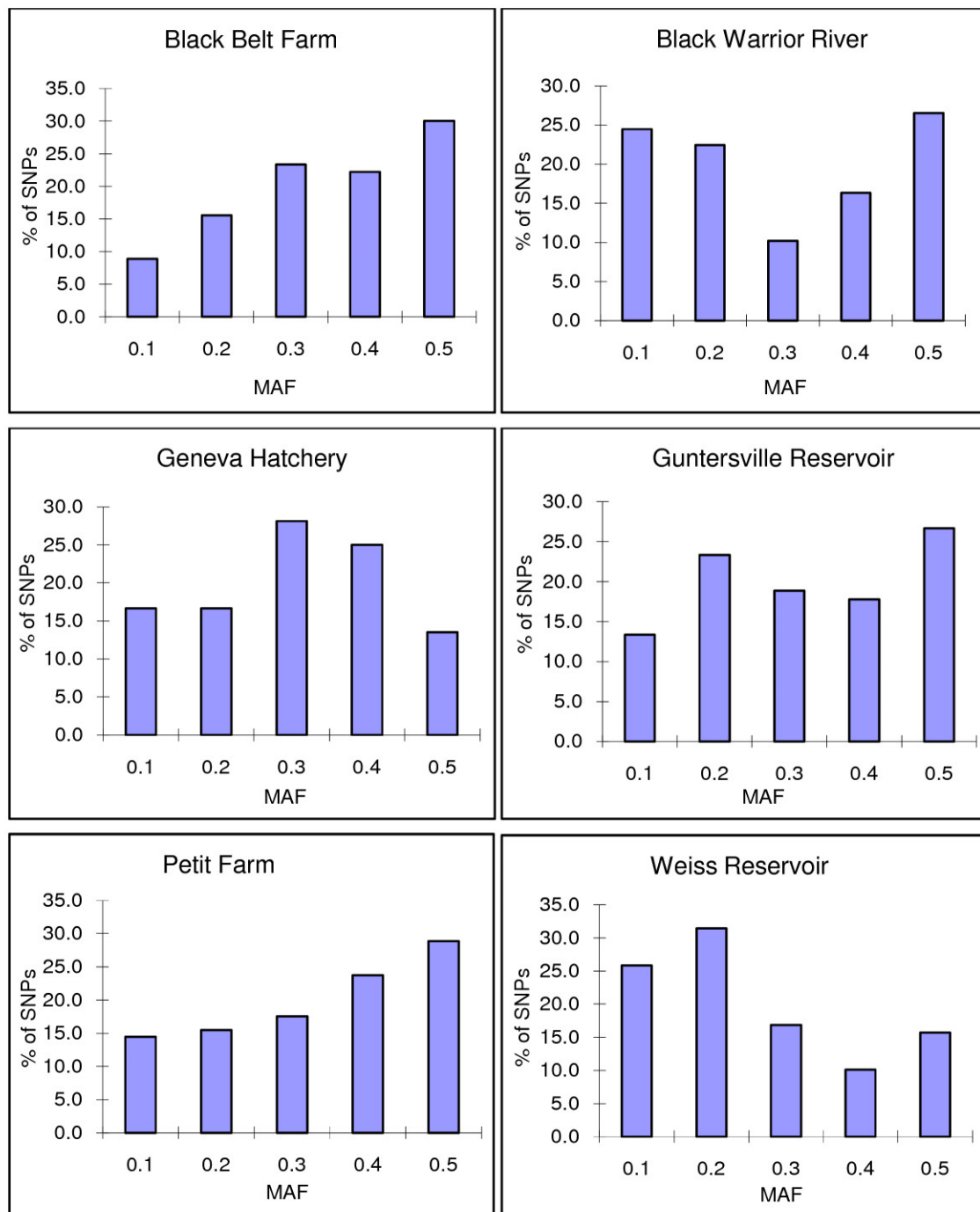
The Illumina's Quality Scores of SNPs did not affect SNP validation rates

Of the total of 384 SNPs tested, SNPs were selected with Quality Scores ranging from 0.5 to 1.0. As shown in Table 3, successful genotypes were obtained from 266 SNPs (of which 156 were polymorphic), while genotyping failed for 118 SNPs. Obviously, this failure rate is high, since we designed in the experiment several parameters (contigs with four or less sequences and minor sequence allele presence of once) to test SNP quality that obviously lowered the overall success rate. The very obvious question was if the Illumina's Quality Scores (as a reflection of the flanking sequence complexity and sequence context) affected the success rate. As indicated in Table 3, the Quality Scores were clearly not associated with the failures of SNP genotyping.

Table 3. Overall summary of the EST-derived SNP genotyping using the Illumina Bead Array technology

Categories	Number of SNPs	Average Quality Score
Successful genotype calling	266	0.87
Polymorphic SNPs	156	0.87
Non-polymorphic SNPs	110	0.87
Failed SNPs	118	0.90
Total number of loci tested	384	0.88

Figure 3. Distribution of minor allele frequency in domestic and wild channel catfish strains.



The name of the populations is labeled on the top of each panel. MAF: minor allele frequency.

Contig size and minor sequence allele frequency were the major determinants on SNP validation rates

The percentage of putative SNPs that was validated to be real (SNP validation rate) was found to be directly correlated with contig sizes (number of sequences in the contig) and the minor sequence allele frequencies (Table 4). In general, the smaller the contig size, the lower the SNP validation rate was. However, a consistently high SNP validation rate was obtained with contigs of at least 4 sequences, with minor sequence being present at least twice. The differences were observed within contigs with 4 sequences. While SNP polymorphic rate of 70.5% was achieved with contigs of sequences with two sequences of equal frequency (2:2), contigs of 4 sequences with 3:1 sequence frequency had only a 15.4% SNP validation rate, suggesting that the minor sequence allele frequency is crucially important. Overall, the average SNP validation rate was only 33.3% for contigs of 4 or fewer sequences with minor sequence allele present only once. However, the overall SNP validation rate for contigs of 4 or more sequences with minor sequence allele present at least twice was 70.9%, and up to 89.2% with contigs of 12 or more sequences (Table 4). Contig length was found not to be related with SNP validation rate. The average contig length of polymorphic SNPs was 1095 bp, 1071 bp for monomorphic SNPs was, 1080 bp for failed SNPs was.

Table 4. SNP polymorphic rates vs contig size and minor sequence allele frequency

# of sequences in the contig	# Successful Loci	Sequence ratio*	Minimal Minor Sequence Frequency	Polymorphic rate (%)
2	24	1:1	50%	33.3
3	37	1:2	33.3%	45.9
4	26	1:3	25%	15.4
Subtotal	87			33.3*
4	44	2:2	50%	70.5
5-6	60	2:3 & 2:4 & 3:3	33.3%	60.0
7-8	17	3:4 & 3:5 & 4:4	37.5%	64.7
9-12	21	4:5 & 4:6 & 4:7 & 4:8 & 5:5 & 5:6 & 5:7 & 6:6	33.3%	76.2
>12	37	5:7 & 6:6 & 5:8 & 6:7.....& 12:57	17.4%	89.2
Subtotal	179			70.9*
Total	266			58.6*

*Average polymorphic rate in respective categories.

Quality of sequences flanking SNPs is important

Flanking sequence quality greatly affected the SNP success rate. Among the contigs with SNPs, we identified 28 contigs with hot spots of SNP occurrence, where a region of sequence was highly variable with many “SNPs” detected. Examination of sequence quality suggested low quality scores in the sequencing reactions. We intentionally included these SNPs in this project to give an assessment of the effect sequence quality on the SNP validation. Of the 28 SNPs tested, 14 (50%) failed in genotyping, suggesting that high sequence quality is required in the SNP region as they are involved in the genotyping primer binding regions (data not shown).

The presence of intron(s) was the major cause for SNP genotyping failures

The presence of introns greatly reduced the SNP genotyping success rate. Among the contigs containing SNPs, only four known genes had genomic DNA information that allowed us to test if the involvement of introns has any effect on SNP genotyping and validation rates. All four SNPs failed to provide genotypes. Clearly, the Bead Array technology depends on very short extension and subsequent ligation for success.

Of the 118 failed SNPs, 14 were likely caused by low sequence quality flanking the SNP sites; and 4 were caused by the involvement of introns, as designed in the experiment. The causes for failure of the remaining 99 SNPs were then explored by *in silico* comparative analysis. Based on the fact that intron involvement led to the SNP genotyping failures, we conducted comparative sequence analysis of the catfish ESTs with corresponding zebrafish genes as references. The rationale is that if the gene organization is similar in catfish and zebrafish (diverged from 110-160 million years ago), then sequence similarity comparison would allow the location of SNP sites to be aligned to the zebrafish genome. If the SNP sites are close to the exon-intron junction, then that could have caused the genotyping failures, assuming conservation of gene structure and organization between catfish and zebrafish. As shown in Table 5, 92 of the 99 catfish SNP loci had significant BLAST hits with the zebrafish genome, but of these, only 50 allowed sequence alignment in the region containing the involved SNPs. Sequence alignment and gene structure in zebrafish indicated that 32 (64%) of the 50 SNPs were located at the exon-intron border, suggesting that the presence of the presumed introns was the major cause for the failures of the SNP genotyping.

Table 5. Effect of low sequence quality (as defined by the presence of hot spots of SNP occurrence) and the presence of predicted intron on success rate of SNP

genotyping

	Tested	Succeeded	Percentage
Number of loci with SNP located in regions containing low quality sequences	14	7	50%
Number of loci with known introns	5	5	100%
Number of failed loci without gene information	99		
With Significant Blast hits	92		92.9%
SNP positions can be located by similarity comparisons with zebrafish genome	50		54.3%
Number of Loci with SNP predicted to be positioned at exon-intron border	32		64%
Total number of loci potentially with SNP positioned at exon-intron border	37		67.3%

Discussion

ESTs proved to be efficient resources for putative SNP identification [10, 12, 14-16]. This study provides an assessment of nucleotide diversity in available catfish EST resources for putative SNP identification. Since our goal was to make quality assessment for the EST-derived SNPs, we designed this project to provide some answers as to how the sequence context (Illumina's Quality Score), contig size, minor sequence allele frequency, sequence quality flanking SNPs, and the distance between the SNP genotyping primers affect SNP validation rates.

When compared to SNPs identified from genomic sequences, EST-derived SNPs have several advantages. Since ESTs are transcribed sequences, EST-derived SNPs are

associated with actual genes, allowing use of gene-associated SNPs for mapping and subsequent use in comparative genome studies [17]. This is particularly important for species without a genome sequence, such as aquaculture species. In addition to be used as markers for mapping, SNPs are also considered a rich source of candidate polymorphisms underlying important traits leading to the identification of causative genes or quantitative trait nucleotide (QTN) [18]. However, several important factors need to be considered when using EST-derived SNPs. The major issue for development of SNPs from EST resources is not whether SNPs can readily be identified, but to what degree these SNPs would be reliable since parameters for quality assessment of EST-derived SNPs simply do not exist. This reliability issue was mostly due to sequence errors; assembled contigs with sequence variation could simply be sequence errors. Additionally, since SNPs derived from ESTs can only be identified from EST contigs where the same gene transcripts were sequenced at least twice and sequencing frequency of ESTs is not random, large-scale sequencing is required to identify SNP's from rarely expressed genes. Moreover, SNP rates could be lower in coding regions because of evolutionary constraints and/or selection pressure.


In this study, over 33,000 putative SNPs were identified from 55,000 catfish ESTs and 384 of these SNPs were tested using 192 catfish samples. We have found that the contig size (number of sequences in the contig) and minor sequence allele frequency were the two major factors affecting the validation rates of EST-derived SNPs. Small contigs had much lower SNP validation rates. Obviously, in small contigs with 2 or 3 sequences, the alternative base is represented only once, and this could be due to sequencing errors. Similarly, in contigs with 4 sequences and when the minor sequence

allele is represented only once, it is highly likely that the minor allele is due to sequencing errors. Contigs of 4 or more sequences with the minor sequence allele frequency being present at least twice in the contig provided high levels of SNP validation rates (average 70.9 % and up to 89.2%). This makes good sense because it is highly unlikely that sequencing errors of two independently sequenced ESTs to occur at the same base location. When at least two ESTs exhibit an alternative base at the putative SNP sites, it is highly likely that such sequence variations are real. All these findings were not unexpected, but for the first time, we provide experimental data to demonstrate the importance of contig size and minor sequence allele frequency. It is noteworthy that even though the larger contigs provided even greater SNP validation rates, contigs of four sequences with even sequence allele distribution (2:2) provided similarly high validation rates. Thus, a minimum of two sequences in the contigs representing the minor allele was required to provide a high SNP validation rate [10, 12].

The presence of minor allele sequence in relation to the contig size appears important. For instance, if the minor allele sequence was present only once, then the smaller the contig size, the more likely the SNP could be real. This is because the contig size of ESTs is simply a reflection of expression abundance. If a rarely expressed gene was sequenced twice, with the alternative allele being present once each, one can still expect that the allele frequency could be equal, or close to equal, when the transcript is sequenced 10 times. However, if the transcript was already sequenced 10 times, with the minor allele sequence being present only once, it is more likely that the minor allele could have been derived from sequencing errors (Figure 4). This relation is obvious when sequence heterozygosity is considered, as shown in Figure 2. A contig of two sequences with one

each of the alternative alleles would have a sequence heterozygosity of 0.5, while a contig with 10 sequences of 9 major allele:1 minor allele would have a sequence heterozygosity of only 0.18.

Figure 4. SNP quality assessment based on EST contig size and sequence frequency of the alleles.

# seq	Minor sequence frequency	Major sequence frequency	Sequence Heterozygosity	SNP quality trend
10 seq	1	9	0.18	
9 seq	1	8	0.20	
8 seq	1	7	0.22	
7 seq	1	6	0.24	
6 seq	1	5	0.28	
5 seq	1	4	0.32	
4 seq	1	3	0.38	
3 seq	1	2	0.44	
2 seq	1	1	0.50	
10 seq	2	8	0.32	
9 seq	2	7	0.35	
8 seq	2	6	0.38	
7 seq	2	5	0.41	
6 seq	2	4	0.44	
5 seq	2	3	0.48	
4 seq	2	2	0.50	
10 seq	3	7	0.42	
9 seq	3	6	0.44	
8 seq	3	5	0.47	
7 seq	3	4	0.49	
6 seq	3	3	0.50	
10 seq	4	6	0.48	
9 seq	4	5	0.49	
8 seq	4	4	0.50	
10 seq	5	5	0.50	

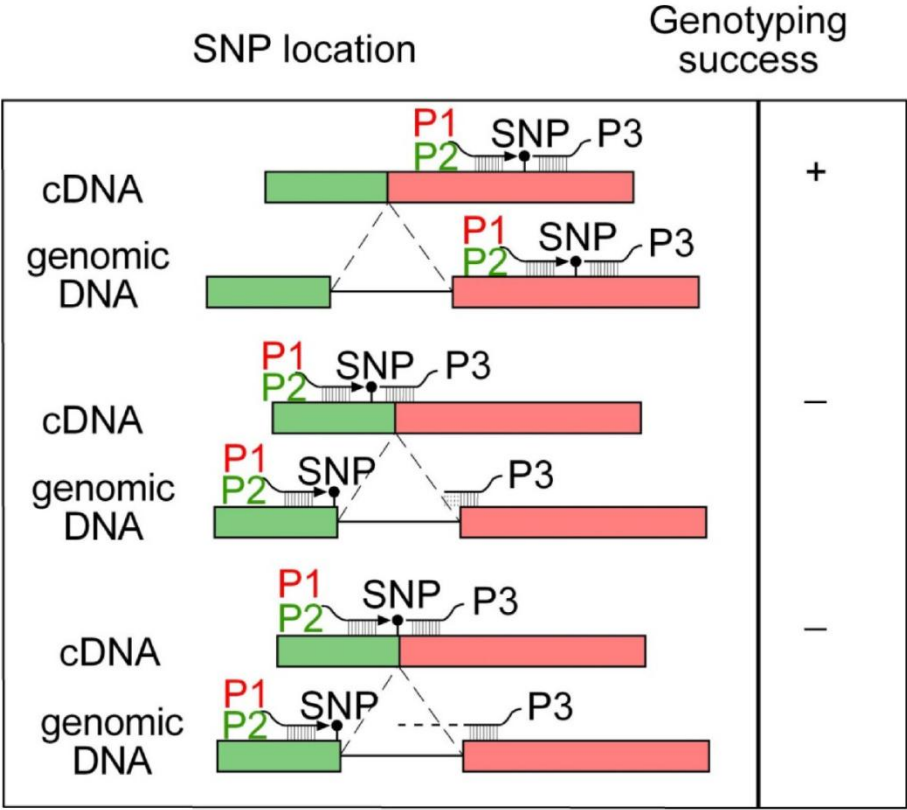
*Arrows indicate the trend of SNP quality, with the black arrows indicating trend of heterozygosity within a subset of contigs with the same number of the minor allele sequence, and the red arrow indicating overall SNP quality trend.

Another advantage of the SNP identification from EST sequences is its ability to identify uncommon sequence variants [16]. The monomorphic SNP rate was highly related to the number of samples tested, since the uncommon sequence variants possess very low minor allele frequency, which required a large number of samples. According to

our results, the monomorphic SNPs accounted for 28% of tested SNPs. However, these monomorphic SNPs could be false SNPs caused by sequencing errors. In addition, much smaller fish samples (10 fish) were used to construct the EST libraries than the number of fish samples used here to validate the SNPs, further supporting the possibility of sequencing errors related to monomorphic SNPs.

Sequence quality flanking the SNP sites was found to be important for successful SNP genotyping using Illumina's Bead Array technology, but not the flanking sequence context as referred to as the Quality Score by Illumina when above 0.5. It is probably true that SNP genotyping primers would have worked properly for the most part even if the sequence context was somewhat simple or A/T-rich, or G/C-rich. However, sequence errors in the SNP region could directly affect the base pairing of the SNP genotyping primers. Low quality sequences could easily generate false SNPs, especially at the beginning or end of the sequence. Therefore, sequence quality surrounding the SNP site should be used as one parameter to identify reliable SNPs. However, many EST sequences retrieved from NCBI do not have quality scores or trace files. In such cases, greater caution should be exercised. In particular, hot spot of SNP occurrence should be avoided if possible.

Figure 5. Schematic illustration of the effect of introns involved in SNP genotyping.



*In the first case, all the genotyping primers are located in the same exon nearby, leading to successful genotyping (+); in the second case (middle), one of the genotyping primers (P3 as shown) was located at the exon-intron border, causing non-base pairing that lead to failure of genotyping (-); and in the third case, even though all primers were located in exon regions. However, an intron was involved that demands PCR extension to across the intron. Apparently, the Bead array technology provide very limited extension capability, leading to genotyping failure (-) as well.

Selection of SNPs to allow both allele-specific and locus-specific primers to be located in a single exon is the key to achieving high success rate of SNP genotyping. We found that all tested SNP sites involving introns failed in genotyping. There seemed to be different reasons for such genotyping failures. The most notable cause is that the

genotyping primers are located at exon-intron boundary, leading to non-base pairing of the primers with DNA amplified from genomic DNA (Figure 5). In addition, it appeared that the extension of the genotyping primer P1 and/or P2 to reach P3 (see Materials and Methods above) is quite limited. In cases when even both genotyping primers had a perfect match with the template DNA, genotyping also failed simply because an intron was predicted to be present between the genotyping primers (Figure 5). This is somewhat unexpected as one would expect that the DNA polymerase should be able to extend easily a few hundred bases. In addition to the few tested loci, comparative gene organization analysis suggested that the vast majority of failed SNPs involved introns immediately flanking the SNP sites, further supporting the inability of genotyping when SNP is located at the exon-intron boundary or when introns are included in the extension reaction. Therefore, bioinformatics analysis using *in silico* comparative sequence and gene structural analysis is important when dealing with EST-derived SNPs.

Stringent quality assessment measures should be used when working with EST-derived SNPs, since ESTs are single pass read of cDNA sequences, and the quality is relative low. In particular, contigs containing four or more ESTs should be used and the minor allele sequence should be represented at least twice. Genotyping primers should be designed from a single exon as to completely avoiding introns because of the limitation of genotyping primer extension used by Illumina genotyping assay. Application of such quality assessment measures, along with large resources of ESTs, should provide effective means for SNP identification in species where genome sequence resources are lacking.

Methods

EST clustering and contig assembly

All catfish EST sequences were downloaded from NCBI dbEST database, including those of blue catfish and channel catfish. CAP3 was used to assemble the contigs with the parameters set at “minmatch 50, overlap similarity 0.95”, to have a minimal overlap of 50 bases and a minimal similarity of 95% [19]. For each contig generated from the CAP3 assembly, BLASTX was conducted against the non-redundant *nr* database to assist identification of any related ESTs in different contigs. A significant hit was defined as having an E-value below e^{-10} and 100 minimum of alignment length for all sequences. Following initial gene identification, related ESTs were further evaluated by manual inspection of the alignments.

SNP identification using EST resources

The autoSNP program was used to detect putative SNPs from the EST sequences [20]. The program utilized the CAP3 output files as input to detect SNPs based on the base redundancy in the sequence alignments. The autoSNP program generated two text files, a contig file including contig ID, consensus length, number of sequences in the contig, and the number of SNPs, a SNP file including Contig ID, SNP position, minor allele frequency, SNP allele, mutation type, and base alignment in the SNP position. The program also generated an *html* file for each contig, including the alignment information and SNP information. With the autoSNP program, the parameters for minimum minor allele frequency for SNP detection varied with the contig size (the number of sequences in the contig): 1) a sequence variation is declared as a SNP whenever a mismatch is

identified within contigs with four or fewer sequences; 2) a sequence variation is declared as a SNP when the minor allele sequence existed at least twice within contigs with 5-6 sequences; 3) a sequence variation is declared as a SNP when the minor allele sequence existed at least three times within contigs with 7-8 sequences; 4) similarly a sequence variation is declared as a SNP when the minor allele sequence existed at least four times within contigs with 9-12 sequences, and 5) when the minor allele sequence existed at least five times within contigs with 13-16 sequences and so on.

Selection of SNPs for this project

To evaluate the effect of contig size and minor allele sequence frequency on SNP reliability, the SNPs with different contig sizes and minor allele frequencies were selected for SNP validation. After initial submission of a set of SNPs to Illumina, GoldenGate assay functionality and designability scores were given by Illumina. SNPs with a range of functionality and designability scores were chosen for evaluation in this project. A total of 384 SNPs were selected for this project. Hot spots of SNP occurrence that may have been caused by low sequence quality were selected to test how sequence quality affects SNP genotyping and validation rates. In addition to sequence quality, the effect of intron presence on genotyping and validation rates was tested by including SNPs with four known genomic sequences.

Fish samples used for validation of SNPs

A panel of 192 samples were used for genotyping and validation of SNPs including 66 fish from our interspecific mapping resource family F1-2 x Channel catfish-6 (64

backcross progenies plus their two parents), and 21 fish each from three wild channel catfish populations including Black Belt Farm, Geneva Hatchery, and Petit Farm, and three domestic channel catfish populations including Black Warrior River, Guntersville Reservoir, and Weiss Reservoir [21].

SNP genotyping assay

Genomic DNA (250 ng per sample) was used as template for SNP genotyping using the Illumina's bead array technology according to the manufacturer's protocol for GoldenGate assay [22]. Briefly, two allele-specific primers labeled with Cy3 (P1) or Cy5 (P2) and a third locus-specific primer (P3) with an address sequence were first hybridized to the template and allele-specific primers were extended to cross the SNP site to reach the locus-specific primer. After this allele-specific extension, ligation was conducted between allele-specific primer(s) and the locus-specific primer, creating a PCR template. PCR reaction was conducted using both allele-specific primers and the locus-specific primer. The PCR reaction products were hybridized onto a chip (Illumina Inc., San Diego) containing bead types coated with oligo-nucleotides complementary to the locus-specific primer address on the PCR product. Each bead type is represented with an average redundancy of 30X on the array to optimize the accuracy of the final genotype signal. Following hybridization, the bead array signal was determined using a bead array reader, which could convert images to intensity data. The intensity data for each SNP for each sample was normalized and assigned a cluster position (and resulting genotype), and a quality score for each genotype was generated. Final genotyping results were automatically generated for downstream analysis using the BeadStudio software.

Data analysis

The BeadStudio software was used to analyze the SNPs data. The dye intensities are examined by the software to determine the genotype of each sample for that locus. A locus returning predominantly signal from Cy3 is AA, Cy5 is BB and an equal signal of Cy3 and Cy5 represents a heterozygous individual. Data is returned with the allele call for each locus as well as a GenTrain score, a measure that represents the reliability of that genotyping call. GenTrain scores was used to measure the reliability of SNP detection based on the distribution of genotypic classes, and the calling frequency was used to measure the successful SNP calling rate from all samples [23]. For this study, GenTrain score of 0.4, call rate of 90%, and minor allele frequency of 0.05 was used. After removing failed SNPs, the remaining SNPs were identified as successful SNPs in genotyping. Successful genotypes were used further for the analysis of minor allele frequencies, and for the calculation of SNP validation rate. Heterozygosity is defined with the formula $H=1-(p_a^2+p_b^2)$ where P_a is the allele frequency of the major allele and P_b is the allele frequency of the minor allele [24]. Chi-square test was conducted to test the relationship between minor allele sequence frequency and SNP successful rate.

Acknowledgement

This project was supported by a grant from USDA NRI Animal Genome Basic Genome Reagents and Tools Program (USDA/NRICGP award # 2006-35616-16685), and partially by an AAES Ag Initiatives grant.

References

1. Ron M, Weller JI: **From QTL to QTN identification in livestock--winning by points rather than knock-out: a review.** *Anim Genet* 2007, **38**(5):429-439.
2. Rothschild MF: **Porcine genomics delivers new tools and results: this little piggy did more than just go to market.** *Genet Res* 2004, **83**(1):1-6.
3. Butcher LM, Davis OS, Craig IW, Plomin R: **Genome-wide quantitative trait locus association scan of general cognitive ability using pooled DNA and 500K single nucleotide polymorphism microarrays.** *Genes Brain Behav* 2007.
4. Kiyohara C, Yoshimasu K: **Genetic polymorphisms in the nucleotide excision repair pathway and lung cancer risk: a meta-analysis.** *Int J Med Sci* 2007, **4**:59-71.
5. Lazarus R, Vercelli D, Palmer LJ, Klimecki WJ, Silverman EK, Richter B, Riva A, Ramoni M, Martinez FD, Weiss ST, Kwiatkowski DJ: **Single nucleotide polymorphisms in innate immunity genes: abundant variation and potential role in complex human disease.** *Immunol Rev* 2002, **190**:9-25.
6. Rafalski A: **Applications of single nucleotide polymorphisms in crop genetics.** *Curr Opin Plant Biol* 2002, **5**(2):94-100.
7. Barbazuk WB, Emrich SJ, Chen HD, Li L, Schnable PS: **SNP discovery via 454 transcriptome sequencing.** *Plant J* 2007, **51**(5):910-918.
8. Fan JB, Oliphant A, Shen R, Kermani BG, Garcia F, Gunderson KL, Hansen M, Steemers F, Butler SL, Deloukas P, Galver L, Hunt S, McBride C, Bibikova M, Rubano T, Chen J, Wickham E, Doucet D, Chang W, Campbell D, Zhang B, Kruglyak S, Bentley D, Haas J, Rigault P, Zhou L, Stuelpnagel J, Chee MS:

- Highly parallel SNP genotyping.** *Cold Spring Harb Symp Quant Biol* 2003, **68**:69-78.
9. Shen R, Fan JB, Campbell D, Chang W, Chen J, Doucet D, Yeakley J, Bibikova M, Wickham Garcia E, McBride C, Steemers F, Garcia F, Kermani BG, Gunderson K, Oliphant A: **High-throughput SNP genotyping on universal bead arrays.** *Mutat Res* 2005, **573**:70-82.
 10. Hayes B, Laerdahl JK, Lien S, Moen T, Berg P, Hindar K, Davidson WS, Koop BF, Adzhubei A, Hoyheim B: **An extensive resource of single nucleotide polymorphism markers associated with Atlantic salmon (*Salmo salar*) expressed sequences.** *Aquaculture* 2007, **265**(1-4):82-90.
 11. Moen T, Hayes B, Baranski M, Berg P, Kjolglum S, Koop B, Davidson W, Omholt S, Lien S: **A linkage map of the Atlantic salmon (*Salmo salar*) based on EST-derived SNP markers.** *BMC Genomics* 2008, **9**(1):223.
 12. Pavy N, Pelgas B, Beauseigle S, Blais S, Gagnon F, Gosselin I, Lamothe M, Isabel N, Bousquet J: **Enhancing genetic mapping of complex genomes through the design of highly-multiplexed SNP arrays: application to the large and unsequenced genomes of white spruce and black spruce.** *BMC Genomics* 2008, **9**(1):21.
 13. Guryev V, Koudijs MJ, Berezikov E, Johnson SL, Plasterk RH, van Eeden FJ, Cuppen E: **Genetic variation in the zebrafish.** *Genome Res* 2006, **16**(4):491-497.
 14. Hayes BJ, Nilsen K, Berg PR, Grindflek E, Lien S: **SNP detection exploiting multiple sources of redundancy in large EST collections improves validation**

- rates.** *Bioinformatics* 2007, **23**:1692-1693.
15. He C, Chen L, Simmons M, Li P, Kim S, Liu ZJ: **Putative SNP discovery in interspecific hybrids of catfish by comparative EST analysis.** *Anim Genet* 2003, **34**(6):445-448.
 16. Picoult-Newberg L, Ideker TE, Pohl MG, Taylor SL, Donaldson MA, Nickerson DA, Boyce-Jacino M: **Mining SNPs from EST databases.** *Genome Res* 1999, **9**(2):167-174.
 17. Sarropoulou E, Nousdili D, Magoulas A, G K: **Linking the genomes of nonmodel teleosts through comparative genomics.** *Mar Biotechnol (NY)* 2008, **10**(3):227-233.
 18. Jalving R, van't Slot R, BA vO: **Chicken single nucleotide polymorphism identification and selection for genetic mapping.** . *Poult Sci* 2004, **83**(12):1925-1931.
 19. Huang X, Madan A: **CAP3: A DNA sequence assembly program.** *Genome Res* 1999, **9**(9):868-877.
 20. Barker G, Batley J, H OS, Edwards KJ, Edwards D: **Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP.** *Bioinformatics* 2003, **19**(3):421-422.
 21. Simmons M, Mickett K, Kucuktas H, Li P, Dunham R, Liu ZJ: **Comparison of domestic and wild channel catfish (*Ictalurus punctatus*) populations provides no evidence for genetic impact.** *Aquaculture* 2006, **252**(2-4):133-146.
 22. Shen R, Fan JB, Campbell D, Chang W, Chen J, Doucet D, Yeakley J, Bibikova M, Wickham Garcia E, McBride C, Steemers F, Garcia F, Kermani BG,

- Gunderson K, Oliphant A: **High-throughput SNP genotyping on universal bead arrays**. *Mutat Res* 2005, **573**:70-82.
23. Fan JB, Oliphant A, Shen R, Kermani BG, Garcia F, Gunderson KL, Hansen M, Steemers F, Butler SL, Deloukas P, Galver L, Hunt S, McBride C, Bibikova M, Rubano T, Chen J, Wickham E, Doucet D, Chang W, Campbell D, Zhang B, Kruglyak S, Bentley D, Haas J, Rigault P, Zhou L, Stuelpnagel J, Chee MS: **Highly parallel SNP genotyping**. *Cold Spring Harb Symp Quant Biol* 2003, **68**:69-78.
24. Liu ZJ: Chapter 5 **Microsatellite markers and assessment of marker utility**. In: *Aquaculture Genome Technologies* (edited by Liu ZJ.), 2007, pp. 43-58, Blackwell Publishing, Ames, IA.

**IV. ASSEMBLY OF 500,000 INTER-SPECIFIC CATFISH EXPRESSED
SEQUENCE TAGS AND LARGE SCALE GENE-ASSOCIATED
MARKER DEVELOPMENT FOR GENOME SELECTION STUDIES**

Abstract

Background

EST sequencing is one of the most efficient means for gene discovery and gene expression profiling. With a good resource of ESTs, a large number of molecular markers can be identified and issues related to alternative splicing and differential polyadenylation can be addressed at the genome-wide scale. Through the Community Sequencing Program, a catfish EST sequencing project was selected by the DOE's Joint Genome Institute (JGI).

Results

In this project, a total of 12 cDNA libraries were constructed, including eight from channel catfish (*Ictalurus punctatus*) and four from blue catfish (*I. furcatus*). A total of 600,000 sequencing attempts were made, generating a total of 438,321 quality ESTs. With previously existing ESTs in the GenBank, this project brings the total of ESTs to nearly 500,000 for catfish. The JGI EST sequencing had an overall sequencing success rate of 73%, with an average length of 576 bp. All the ESTs were assembled using CAP3, resulting in 111,578 unique sequences, including 45,306 contigs and 66,272 singletons. Of these unique sequences, over 35% had significant similarities to known genes by BLASTX searches, which allowed the identification of 14,776 unique genes in the catfish. A total of 1,350 and 849 full-length cDNAs have been identified from channel catfish and blue catfish, respectively. The ESTs are an enormous resource for SNP identification. Over 300,000 putative SNPs have been identified, of which over 48,000 are high quality

SNPs as defined by contig size of at least 4 sequences and the minor allele presence of at least twice in the contig. The EST resource should also be valuable for identification of microsatellites and comparative genome analysis.

Conclusions

This large scale EST sequencing project would allow the identification of a majority of catfish transcriptome. The parallel analysis of ESTs from the two closely related ictalurid catfishes should also provide powerful means for the evaluation of ancient and recent gene duplications, and for the development of high-density microarrays in catfish. The inter- and intra- specific SNPs identified from all catfish EST dataset assembly will greatly benefit the catfish introgression breeding selection and whole genome association studies. All ESTs have been deposited in GenBank.

[Supplement materials are available on line. The ESTs from blue catfish and channel have been deposited in GenBank under accession numbers. **FC996013-FC999999, FD000001-FD380635 and GH640296-GH693994**]

Background

Catfish is one of the major aquaculture species in the United States. However, the genome research falls behind other aquaculture species, such as salmon and rainbow trout. The genome resources are quite limited. Genome research requires the development of a number of resources that facilitate both structural and functional analysis of the genome. Many of the required resources have been developed in catfish, including a large number of polymorphic markers [1, 2], linkage maps [3-5], bacterial artificial chromosome (BAC) libraries [6, 7], physical maps [8, 9], and BAC end sequences (BES) [10]. However, expressed sequence tag (EST) resources were low from catfish [11-15], hindering both functional and comparative genome analysis. Large numbers of ESTs have been produced for most model species as well as a number of agriculturally important species [16-21] including bovine (1.5 million), swine (1.4 million), chicken (600,000), Atlantic salmon (471,000), and rainbow trout (281,000). The availability of such EST resources has allowed efficient gene discovery and gene identification in these species, and rapid progress has been made through comparative genome analysis in understanding the structural, organizational, and functional properties of the genomes of these species.

Whole genome sequences are not available for most aquaculture species, but will be available for tilapia soon. In absence of the whole genome sequence, we initiated a large-scale EST project to provide transcriptomic resources in channel catfish and blue catfish. These ESTs will serve as resources for gene discovery and gene identification, will supply the framework for high-density microarray platforms, will provide a foundation for the analysis of full-length cDNAs, and will assist in the identification of genetic markers such as microsatellites and single nucleotide polymorphisms (SNPs).

These resources will also be of great use for comparative genome analysis. In this study, we have taken an inter-specific EST approach to produce a parallel EST resource from two closely related *Ictalurid* species to resolve some of the most difficult issues in teleost genome research, such as paralog confusions involving duplicated genomes [22-24].

Here, we report the generation and analysis of nearly 500,000 ESTs from catfish, including 354,377 ESTs from channel catfish and 139,475 ESTs from blue catfish. Channel catfish and blue catfish EST assembly allowed identification of 45,306 contigs and 66,272 singletons, suggesting a majority of the catfish transcriptome was captured. The analysis of the inter-specific ESTs resulted in the identification of 20,757 gene-associated microsatellites and over 300,000 putative SNPs, of which over 48,000 were generated with presence of minor allele at least twice. The inter- and intra- specific SNPs identified from all catfish EST dataset assembly will greatly benefit the catfish introgression breeding selection and whole genome association studies.

Results

cDNA libraries and EST sequencing

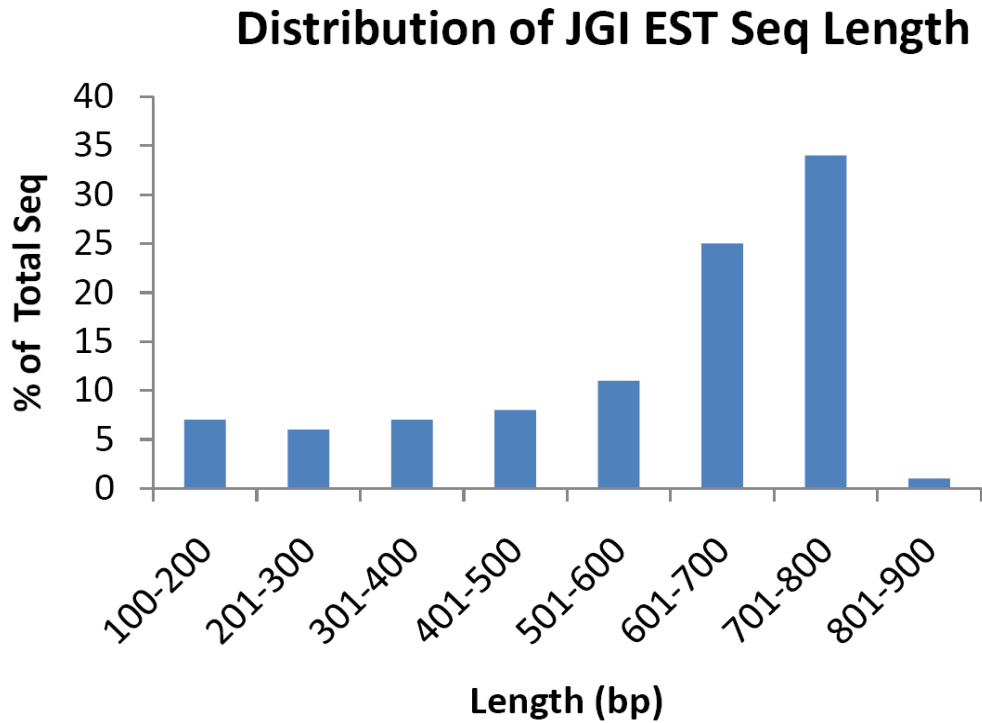
A total of 12 libraries were constructed from various tissues, organs, and cell lines, including four blue catfish libraries and eight channel catfish libraries (Table 6). More than 600,000 sequencing reactions were attempted to sequence a total of 307,296 cDNA clones from both ends.

Table 6. cDNA library information and sequencing summary. Library names were designated by the Joint Genome Institute.

Library	Species	Nature of library	Organ, tissue, or cell line	Total sequences
CBFH	Blue catfish	Normalized	stomach, muscle, olfactory tissue and trunk kidney	37,314
CBZC	Blue catfish	Normalized	stomach, muscle, olfactory tissue and trunk kidney	30,902
CBNH	Blue catfish	Normalized	head kidney, gill, intestine, spleen, skin and liver	9,323
CBZF	Blue catfish	Normalized	head kidney, gill, intestine, spleen, skin and liver	51,172
Subtotal				128,711
CBCZ	Channel catfish	Non-normalized	Mixed leukocytes	16,168
CBFA	Channel catfish	Normalized	catfish whole fry library	63,602
CBNG	Channel catfish	Normalized	kidney, gill, intestine, spleen, skin and liver	2,982
CBZB	Channel catfish	Normalized	kidney, gill, intestine, spleen, skin and liver	57,772
CBNI	Channel catfish	Normalized	stomach, muscle, olfactory tissue and trunk kidney	17,023
CBZA	Channel catfish	Normalized	stomach, muscle, olfactory tissue and trunk kidney	61,320
CBPN	Channel catfish	Subtracted	liver, pituitary, ovary and testis	62,058
CBPO	Channel catfish	Normalized	peripheral blood leukocytes	28,685
Subtotal				309,610
NCBI	Blue catfish			10,764
NCBI	Channel catfish			44,767
Total				493,852

A total of 438,321 ESTs were generated from this project, of which 128,711 sequences were from blue catfish and 309,610 were from channel catfish (Table 6). Of these EST sequences, 219,831 were sequenced from the 5' end of the transcripts, and 218,490 were sequenced from the 3' end of the transcripts. A total of 194,136 clones have paired reads from both 5' and 3' ends of the same transcripts. The lengths of the ESTs range from 100 bp to 877 bp, with an average length of 576 bp and a median length of 655 bp (Figure 6). There were 10,764 ESTs of blue catfish and 44,767 ESTs of channel catfish existing in the GenBank before the start of this project; this project, therefore, brings the total of catfish ESTs to almost a half million sequences (139,475 blue catfish ESTs and 354,377 channel catfish ESTs; Table 6).

Figure 6. Length distribution of JGI EST sequences



EST Assembly

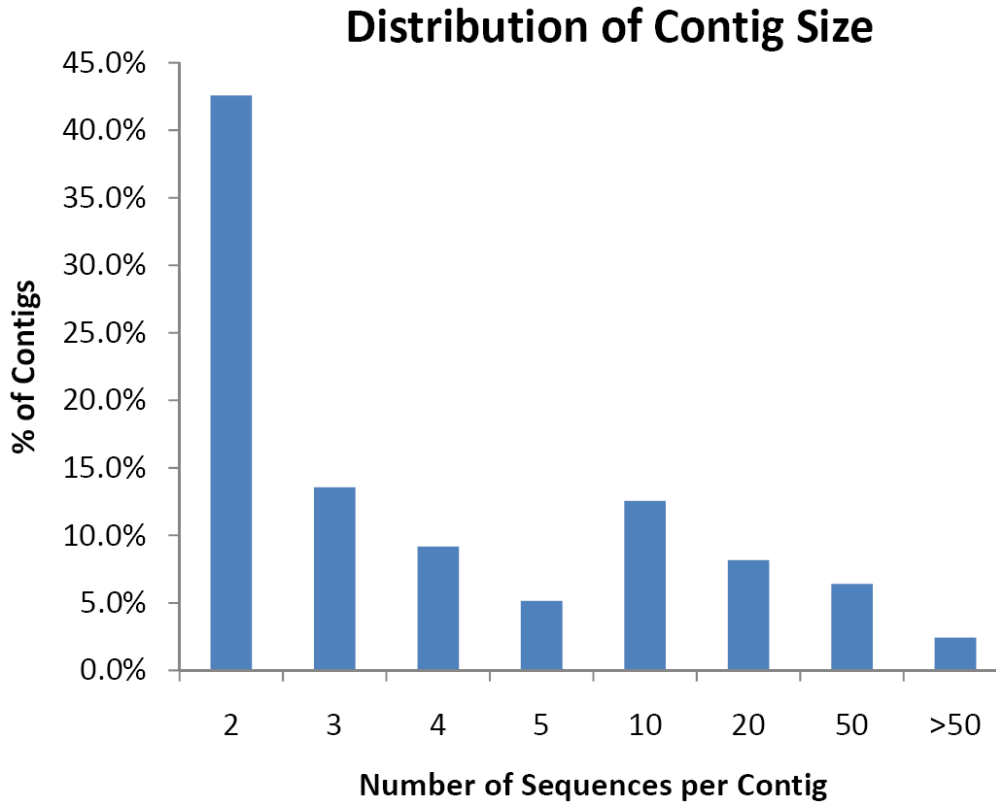
All catfish EST sequences, including those from this project and ones already present in GenBank, were used for the assembly. Three assemblies were conducted: 1) assembly of blue catfish ESTs; 2) assembly of channel catfish ESTs; and 3) assembly of all blue catfish and channel catfish ESTs for inter-specific analysis.

Table 7. EST Assembly statistics

	Blue catfish	Channel catfish	All catfish
Total number of sequences	139,475	354,377	493,852
Short and simple sequences removed	2,735	6,230	8,965
Sequences for assembly	136,740	348,147	484,887
Contigs	22,009	28,941	45,306
Singletons	32,806	41,776	66,272
Average number of sequences per contig	4.72	10.6	9.2
Total unique sequences	54,815	70,717	111,578

As summarized in Table 7, the assembly of the 139,475 blue catfish ESTs resulted in the identification of 54,815 unique EST sequences, including 22,009 contigs and 32,806 singletons; the assembly of the 354,377 channel catfish ESTs resulted in the identification of 70,717 unique EST sequences, including 28,941 contigs and 41,776 singletons. In order to identify inter-specific SNPs, we also conducted the assembly of all available 493,852 ESTs from blue catfish and channel catfish. This assembly allowed the formation of 45,306 contigs, from which potential inter-specific SNPs can be identified. The distribution of contig sizes from the assembly of all catfish ESTs is shown in Figure 7; 43% contigs with 2 sequences, 13% contigs with 3 sequences; and the remaining 44% contigs with 4 or more sequences. The average contig size was approximately nine ESTs per contig. With the ESTs being sequenced mostly from normalized libraries, the vast majority of contigs had 50 or fewer sequences. However, some extremely large contigs were found, including the largest contig with 7,208 ESTs. The putative identity of this contig is apolipoprotein, and it was repeatedly sequenced from all libraries, including high numbers being sequenced from non-normalized libraries already existing in GenBank before this project. As previously reported [34], contig size (number of sequences in the contig, not the consensus sequence length) is one of the two most important factors affecting EST-derived SNP qualities. Therefore, the information on contig sizes is practical and highly useful.

Figure 7. Distribution of contig sizes



To assess the level of common gene discovery from both blue catfish and channel catfish, unique sequences from the EST assemblies were used for BLAST searches. A total of 34,466 (~63%) blue catfish unique sequences, including 16,646 contigs and 19,136 singletons, had hits to at least one unique sequences from channel catfish (E-10), while 20,349 blue catfish unique sequences had no hits to the channel catfish ESTs, suggesting that they were sequenced only from blue catfish. Similarly, 45,171 (~64%) channel catfish unique sequences, including 20,951 contigs and 24,220 singletons, had hits to at least one blue catfish unique sequences (Table 8), while the remaining 25,549 channel catfish unique sequences had no hits to the blue catfish ESTs, suggesting that

they were sequenced only from channel catfish. The identities between homologous blue catfish sequences and channel catfish sequences range from 77% to 100%, with an average of 95%.

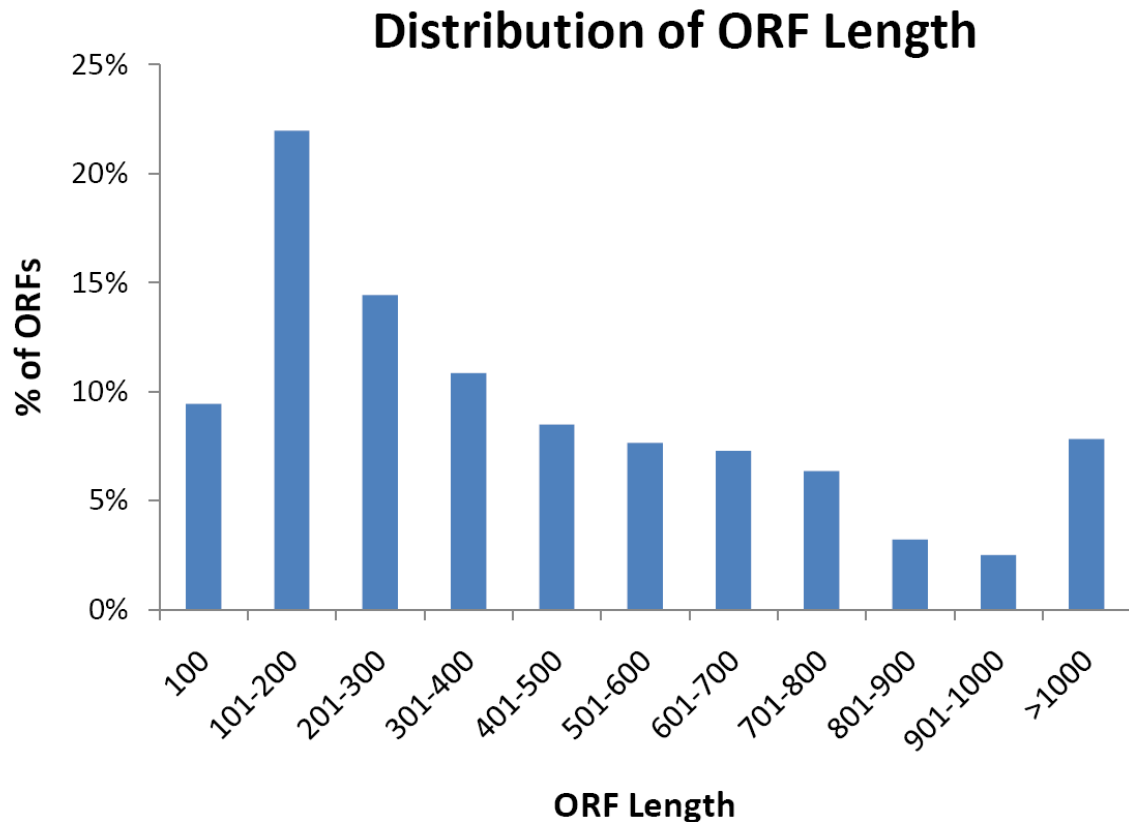
Table 8. Inter-specific similarity comparison of blue catfish and channel catfish unique sequences.

	Blue catfish vs channel catfish	Channel catfish vs blue catfish
Contig:Contig	12,840 : 9,958	14,713 : 9,989
Contig:Singleton	3,806 : 3,303	6,238 : 5,070
Singleton:Contig	11,753 : 7,853	15,684 : 7,468
Singleton:Singleton	6,067 : 4,585	8,536 : 5,204
Total	34,466 : 21,362	45,171 : 21,690

Gene Identification and Annotation

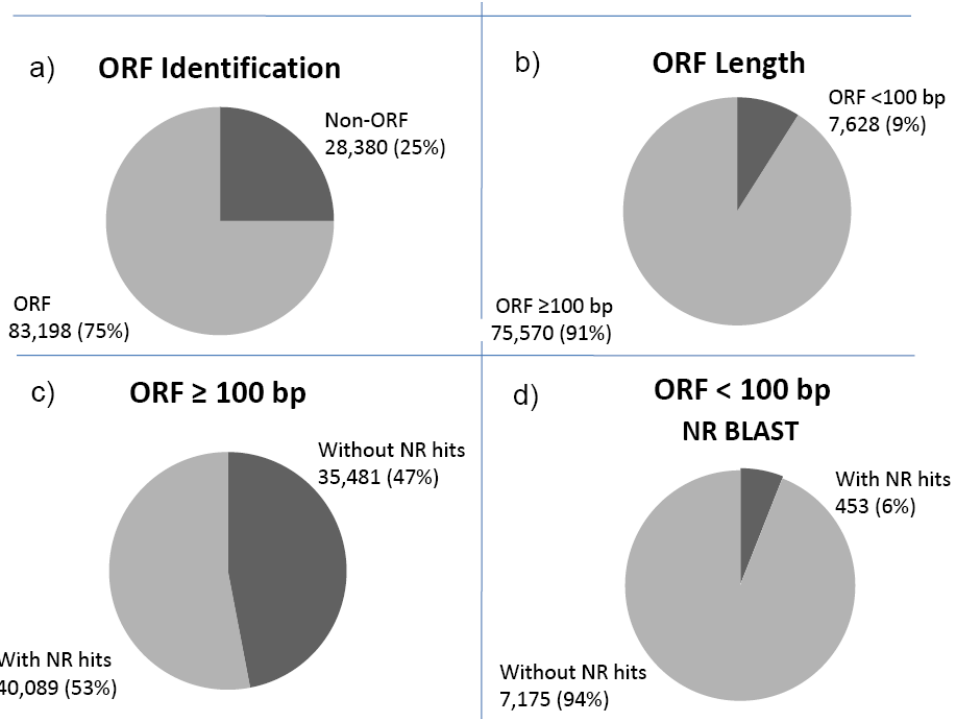
Putative gene identification was conducted using two approaches. The first was to identify open reading frames (ORFs), and the second was to conduct BLASTX searches for similarities with known genes in the public protein databases. Of the 111,578 total catfish unique sequences (total catfish EST assembly), ORFs were detected from 83,198 (75%) unique sequences, with an average ORF length of 450 bp (min=51 bp, max=14,674 bp; Figure 8), and the remaining 28,380 sequences (25%) contained no ORFs (Figure 9a). These ORF-less ESTs were likely ESTs sequenced within the untranslated regions (UTRs). The approach of analysis through the identification of ORFs has the strength of detecting protein-coding capacity without showing any similarities with known genes, but it is incapable of revealing the nature of the involved genes [ORF].

Figure 8. Open reading frame (ORF) length distribution from unique sequences of all catfish assembly.



In order to demonstrate that the vast majority of the identified ORFs were actually gene coding regions, BLASTX searches were carried out based on the size of ORFs. It appeared that the larger the ORFs, the greater the level of putative gene identification through BLASTX searches. Of the identified ORFs, 91% had a length of more than 100 bp. Within these ORFs, 53% had significant BLASTX hits ($1E-10$) (Figure 9b and 9c). However, only 9% of the ORFs with less than 100 bp had significant BLASTX hits ($1E-10$), suggesting that many of these ESTs may either represent novel genes or that the short ORF would not support the similarities using BLAST at the cutoff value of significance (Figure 9d).

Figure 9. Analysis of open reading frames (ORFs).



*a) Percentage of ORFs among unique sequences from all catfish EST assembly; b) Percentage of ORF greater than 100 bp among unique sequences from all catfish EST assembly; c) Percentage of ORFs equal to or greater than 100 bp with significant BLASTX hits; d) Percentage of ORFs smaller than 100 bp with significant BLASTX hits.

A total of 41,311 (37%) unique sequences had significant BLASTX hits within the *nr* database ($1E-10$), and 34,860 (31%) had significant BLASTX hits within Uniprot database ($1E-10$). Over 98% of unique sequences with significant hits were identified with ORFs, which indicated the reliability of ORF searching. After examination of putative protein identities from the BLASTX searches, homologous sequences were identified from the catfish ESTs. Of the 41,311 sequences with BLASTX hits, 22,642 (~55%) and 17,948 (~43%) unique proteins were identified through searches against the *nr* and the Uniprot protein databases, respectively.

To assist in gene annotation, gene ontology (GO) categories were assigned to 16,394 unique catfish sequences with significant BLASTX hits ($1E-10$). At the 2nd level GO terms, 6,266 were assigned to the Biological Processes category (Appendix figure 1), 4,524 to the Cellular Component category (Appendix figure 2), and 7,525 to the Molecular Function category (Appendix figure 3). Figure 6 shows the percentage distributions of GO terms (2nd level). From the GO category of Biological Process, Cellular Process (74 %) was the most dominant 2nd level term, followed by Metabolism (58%). In the Molecular Function category, Binding (61%) was the most dominant, followed by Catalytic Activity (51%).

Assessment of the sequenced catfish transcriptome

In order to assess the level to which the catfish transcriptome has been discovered, the unique sequences were also searched against the NCBI Refseq and Ensemble databases. A total number of significant hits identified within zebrafish, medaka, *Tetraodon*, human, mouse, and chicken reference protein database ($1E-10$) were listed in Table 4. Following removal of duplicates, the unique reference proteins were identified, which represented 12,470 (58%), 12,920 (66%), 10,322 (53%), 9,668 (44%), 11,518 (49%), 8,717 (52%) unique genes from zebrafish, medaka, *Tetraodon*, human, mouse, and chicken database respectively (Table 4). A total of 14,776 unique genes were identified from the catfish based on the BLASTX searches against Ensemble database (Table 9). The majority (>80%) of the unique protein and gene hits were from the contigs

Table 9. Summary of BLASTX searches analysis of catfish ESTs.

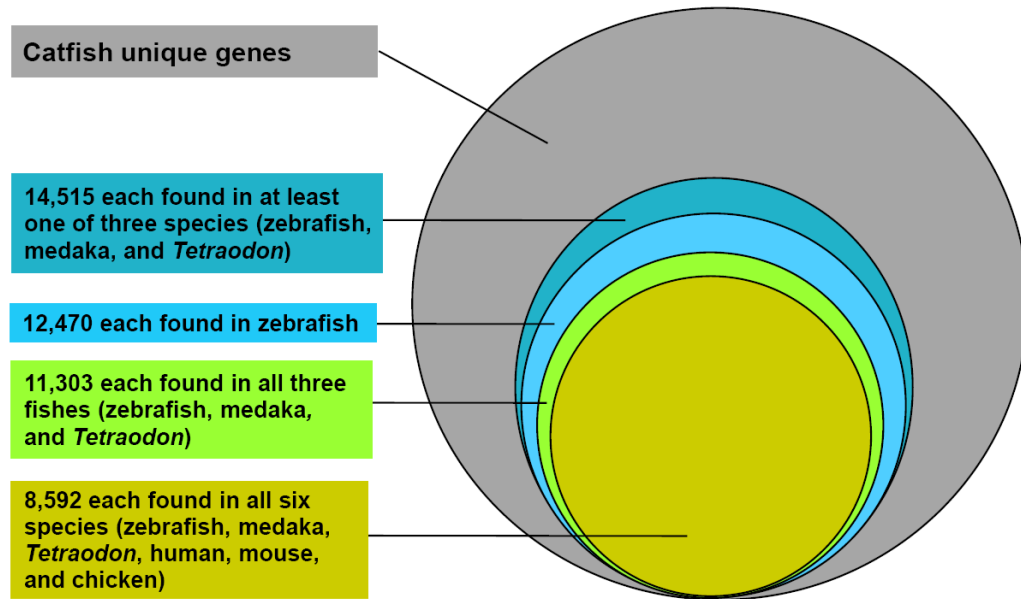
Database	Total ^a	Unique protein ^b	% of the unique protein	Unique gene
NR	41,311 (24,651/16,660)	22,642 (16,265/11,183)		
Uniprot	34,860 (21,412/13,448)	17,948 (13,583/8,782)		
Refseq/Ensemble				
Zebrafish	39,546 (23,487/16,059)	14,988 (12,346/8,534)	54% of 27,996	12,470
Medaka	36,641 (21,835/14,806)	13,588 (11,088/7,641)	56% of 24,461	12,920
Tetraodon	34,418 (20,953/13,465)	13,132 (10,743/7,398)	57% of 23,118	10,322
Human	33,847 (21,038/12,809)	12,621 (10,595/6,924)	33% of 38,342	9,668
Mouse	33,594 (20,942/12,652)	12,267 (10,323/6,808)	35% of 35,236	11,518
Chicken	31,646 (19,661/11,985)	11,059 (9,267/6,319)	50% of 22,194	8,717
Total (E-10)	42,669 (24,880/17,788)	16,439 (13,154/9,416)		14,776
Total (E-5)	47,576 (26,431/21,145)	17,060 (13,485/10,407)		16,173

^aThe first number in the bracelet is the number of contig sequences, and the second number is the number of singleton sequences.

^bThe first number in the bracelet is the number of proteins hit by contig sequences, and the second number is the number of proteins hit by singleton sequences

To assess the evolutionary conservation of the identified unique genes, the number of hits to unique genes in each species of zebrafish, medaka, *Tetraodon*, human, mouse, and chicken were compared. A total of 8,592 (58%) putative known unique genes were found in all six species; 11,303 (76%) were found in all three fish species and 14,515 (98%) were found in at least one of the three fish species (Figure 10).

Figure 10. Number of catfish homologous genes identified from other species using BLASTX searches.



Prediction of full-length cDNAs

The catfish EST sequences provide a platform for the identification and characterization of full-length cDNA clones without having to use expensive and labor-intensive primer walking sequencing. In the context of this presentation, full-length cDNA inserts were defined as a cDNA with the start codon ATG and presence of poly (A) tail in the cDNA clones. In order to determine if the identified ATG in cDNAs were potential “true” start codons rather than in frame internal ATG codons, the putative full-length cDNAs were searched against the Uniprot protein database. If the catfish sequence aligns well with the protein with the best hit and the catfish ATG codon is further upstream, at the same position, or within the first 10 amino acids as compared to the reference sequence, the catfish cDNA clone was considered to harbor a full-length

cDNA. A total of 7,382 blue catfish and 10,037 channel catfish unique cDNA clones with full-length inserts were identified from the assembly with a cutoff E-value of 1E-5, which represented 5,293 unique genes in blue catfish, 6,098 unique genes in channel catfish, and a total of 8,336 unique genes from catfish (Table 10). The full-length cDNA clones provide a convenient way for the complete cDNA sequences, simply by completion of sequencing of the cDNA clones.

Table 10. Full-length cDNA identification

	Blue catfish	Channel catfish
Unique cDNA with full-length	7,382	10,037
Unique gene with full-length	5,293	6,098
Unique full-length cDNA	849	1,350
Unique full-length gene	721	1,159

The full-length cDNA analysis also allowed us to obtain full-length cDNA sequences sequenced from the same clone. To clarify, the full-length cDNA sequences were generated from single-pass sequencing, rather than from assembly of sequences in the same contig. First, the cDNA clones had to qualify for containing the full-length cDNA as defined above; and second, the 5' and the 3' sequences generated from the same cDNA clones overlapped. A total of 849 blue catfish and 1,350 channel catfish unique full-length cDNAs were obtained, representing 721 unique blue catfish and 1,159 channel catfish genes, and a total of 1,260 catfish genes (Table 10), after removing related full-length cDNAs derived from alternative splicing and differential polyadenylation.

Microsatellite and SNP marker identification

A total of 20,757 microsatellites were initially identified from 15,082 unique sequences, including di-, tri-, tetra-, penta- and hexa-nucleotide (Table 11). After removing the microsatellites without enough flanking sequence for primer design, 13,375 unique sequences with microsatellites have sufficient flanking sequences (50 bp) on both sides of the microsatellites to design primers for genotyping.

Table 11. Summary of microsatellite marker identification from catfish ESTs.

Total number of unique sequences	111,578
Microsatellites identified	20,757
Di-nucleotide repeats	12,367
Tri-nucleotide repeats	5,506
Tetra-nucleotide repeats	2,664
Penta-nucleotide repeats	182
Hexa-nucleotide repeats	38
Number of unique sequences containing microsatellites	15,082
Number of unique sequences containing microsatellites with sufficient flanking sequences for PCR primer design	13,375

A total of 48,702 putative SNPs were identified from the blue catfish EST dataset assembly while 102,252 putative SNPs were identified from channel catfish EST dataset assembly (Table 7). These putative SNPs indicated an SNP rate of 3.2 SNPs per kilobase of transcribed sequences in blue catfish, and 4.1 SNPs per kilobase of transcribed sequences in channel catfish. These SNP rates were calculated from the total consensus sequence length and, therefore, the deeper the EST sequencing was, the greater the possibility for the identification of an SNP within the consensus sequences.

For practical applications, catfish breeding programs involve the use of channel catfish x blue catfish hybrids and introgression. Genetic linkage mapping has been conducted in both the intra-specific resource families involving only channel catfish [5]

and the inter-specific resource families made from backcrosses of the channel catfish x blue catfish hybrids [3,4]. Therefore, we also conducted EST assembly using both blue catfish and channel catfish ESTs, and we referred to this assembly as the “all catfish EST assembly”. Over 303,000 putative SNPs and 100,000 putative indels were identified from the all catfish EST assembly results (Table 12).

Table 12. Summary of SNP identification from the catfish ESTs

Putative SNPs identified from the catfish ESTs			
	Blue catfish	Channel catfish	All catfish
Transitions	29,305	61,184	172,746
Transversions	19,397	41,068	130,254
Total SNPs	48,702	102,252	303,000
Indels	14,803	41,660	100,636
SNP rate (kb)	3.2	4.1	7.7
Filtered putative SNPs identified from the catfish ESTs			
Transitions	2,886	11,012	32,235
Transversions	1,005	4,815	16,359
Total SNPs	3,891	15,827	48,594
Indels	1,070	6,707	19,398
Filtered/Non filtered rate	7.8%	15.7%	16.2%
SNP rate* (kb)	0.25	0.64	1.6

*SNP rate was calculated by dividing the total number of SNPs excluding indels with the total length (bp) of the consensus sequences of the contigs.

EST-derived SNPs are often prone to sequencing errors. Therefore, the putative SNPs were subjected to filtering using only those with contig sizes of at least four sequences and the minor allele presence of at least twice in the contigs, and indels were not used for further analysis [25]. After filtering, 3,891 and 15,827 SNPs were identified from blue catfish and channel catfish EST dataset assembly, respectively. A subset of 48,594 filtered

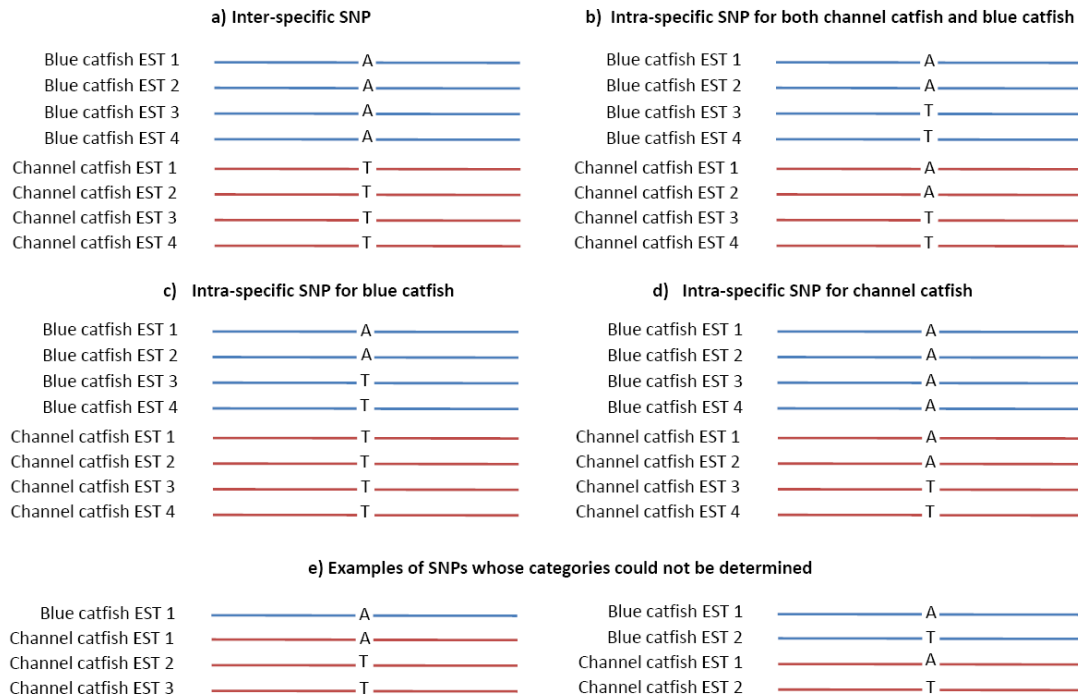
SNPs were obtained from all catfish EST assembly; these SNPs included 32,235 transitions and 16,359 transversions (Table 7). The filtered SNP frequency in the transcribed sequences was 0.25 SNP in blue catfish, 0.64 SNP in channel catfish, and 1.6 SNP in all catfish assembly per kilobase. Of the 48,594 SNPs, over 90% were identified from the contigs containing 5 or more sequences (Table 13).

Table 13. Quality assessment of the filtered putative SNPs identified from the catfish ESTs based on the number of sequences per contig and the sequence frequencies of the minor alleles

No. of sequences in the contig	No. of contigs with SNPs	No. of SNPs	SNP rate (per kb)
2 (1:1)	16,567	96,565	5.2
3 (2:1)	8,374	86,686	10.8
4 (3:1)	5,136	71,155	13.0
Subtotal	30,077	254,406	8.0
4 (2:2)	1,528	5,008	0.9
5-6 (2)	3,099	13,725	2.0
7-8 (3)	805	2,659	0.7
9-12 (4)	730	2,376	0.5
13-20 (5)	629	2,307	0.6
21-30 (5)	628	2,864	1.3
31-50 (6)	730	5,052	3.0
51-100 (6)	542	6,379	6.0
101-500 (6)	316	6,580	13.4
>500	31	1,644	15.0
Subtotal	9,038	48,594	1.6
Total	39,115	303,000	7.7

The assessment of the rates of inter-specific SNPs and intra-specific SNPs may have practical applications. We therefore assessed these SNP rates using the EST data. First, SNPs were identified from contigs containing at least four sequences with at least two sequences from either channel catfish or blue catfish in the all catfish EST assembly. Inter-specific SNPs were defined as those that have sequence variations between blue catfish and channel catfish, but no sequence variations within blue catfish or within channel catfish; similarly, SNPs were identified within blue catfish but not within channel catfish or vice versa; and SNPs were identified within both channel catfish and blue catfish at the same SNP positions (Figure 11).

Figure 11. Categorization of four different types of SNPs (a-d) that can be identified from the all catfish EST assembly, and examples of SNPs whose categories could not be determined due to the minor allele sequence from a given species is fewer than two (e).



Of the 48,594 filtered SNPs, 42,080 were identified from contigs comprising both channel catfish and blue catfish ESTs, and 6,514 were identified from contigs composed of ESTs from either channel catfish or blue catfish, including 5,396 from channel catfish contigs and 1,118 were identified from blue catfish contigs. Of the contigs containing ESTs from blue catfish and channel catfish, the estimation of percentage of inter- and intra-specific SNPs was conducted based on the identification of SNPs from 1000 randomly selected contigs (Table 14). Although a large number of filtered inter-specific SNPs were identified (18,000 out of 48,000 total filtered SNPs), they were identified from a relatively small number of contigs.

Table 14. Estimation of proportions of inter-specific and intra-specific SNPs from the set of filtered SNPs identified from the inter-specific all catfish EST assembly

SNP type*	From 1,000 random contigs	Estimated from all catfish assembly	Estimated % of total filtered SNP
Inter-specific SNP ¹	430	18,731	39
Intra-specific SNP, blue catfish ²	12	523	1
Intra-specific SNP, channel catfish ³	54	2,352	5
Intra-specific SNP, blue catfish & channel catfish ⁴	87	3,790	8
Undetermined ⁵	383	16,683	34
Subtotal	966	42,080	87
SNP from only blue catfish ESTs ⁶	N/A	1,118	2
SNP from only channel catfish ESTs ⁶	N/A	5,396	11
Subtotal	N/A	6,514	13
Total SNP	N/A	48,594	100

*SNPs were identified from contigs containing at least four sequences with at least two sequences from either channel catfish or blue catfish in the all catfish EST assembly: ¹ where there were no intra-specific blue catfish SNPs or intra-specific channel catfish SNPs, but the sequence differed between the two species at the inter-specific SNP position; ² where there were SNPs within blue catfish, but not within channel catfish; ³ where there were SNPs within channel catfish, but not within blue catfish; ⁴ where there were SNPs within both blue catfish and channel catfish; ⁵ undetermined because overall the SNPs qualified as SNPs with at least two minor allele sequences, but only one of the minor allele sequences was from one of the two species of blue catfish or channel catfish; ⁶ these SNPs were identified from ESTs that have been only sequenced from one of the two species, blue catfish or channel catfish to date.

Discussion

This project represented one of the major milestones in catfish genome research, and it brings the catfish EST resources to almost a half million sequences, including previously existing ESTs in GenBank [11-15]. The EST resource will prove to be useful for gene discovery, molecular marker development, and genetic linkage and comparative mapping. Such a resource should facilitate whole genome sequencing and annotation of the catfish genome. The parallel EST sequencing in two closely related species, *Ictalurus punctatus* and *I. furcatus*, may also provide material for the analysis of genome duplication and genome evolution.

The single most important function of EST sequencing is for gene discovery. However, the assessment of the numbers of genes discovered in an EST project depends on assembly using bioinformatic tools, which in turn depends on sequence identities, EST sequence lengths, and the relationship of the species under study relative to information available in existing databases. In this project, the assembly of the channel catfish ESTs allowed identification of 28,941 contigs and 41,776 singletons, resulting in 70,717 unique sequences in channel catfish; similarly, assembly of the blue catfish ESTs allowed identification of 22,009 contigs and 32,806 singletons, resulting in 54,815 unique sequences in blue catfish. Obviously, a larger fraction of the channel catfish transcriptome was captured because more clones were sequenced from channel catfish than from blue catfish. While it is certainly true that not every contig represented a unique gene, the majority of the contigs, however, should represent unique genes.

For gene discovery purposes, we also conducted EST assembly using ESTs from both channel catfish and blue catfish. Our previous reports indicated that the channel

catfish and blue catfish shared 98.7% identities across EST sequences [26]. Thus, bringing all ESTs from both species together should provide a more complete picture as to what fraction of the catfish transcriptome was captured to date. For instance, if four genes A, B, C, and D have been found from channel catfish, and four genes A, B, D, and E have been found from blue catfish, we can regard that five genes: A, B, C, D, and E, have been found from catfish. Such an approach was also taken because of practical considerations. Hybrid catfish produced by inter-specific hybridization of channel catfish x blue catfish is one of the best catfish used in aquaculture, and many believe that industry-wide application of this hybrid may have a revolutionary impact on the catfish industry [27]. One of the major catfish breeding programs is based on introgression of beneficial genes from blue catfish into channel catfish breeds. Most of the catfish linkage mapping has been conducted using the inter-specific hybrid resource panels that can exploit inter-specific polymorphisms [3, 4].

Assembly of all the catfish ESTs allowed identification of 45,306 contigs and 66,272 singletons, resulting in 111,578 unique sequences. Since blue catfish and channel catfish are from the same genus, most of the contigs from blue and channel EST assembly are expected to merge together in an all catfish EST assembly. However, the all catfish EST assembly generated 45,306 contigs, which are much larger than the contigs generated in either blue catfish (22,009) or channel catfish (28,941) EST assembly. There could be several reasons for this major increase in contig numbers with the all catfish EST assembly. First, some ESTs belonging to the contigs were only sequenced in blue catfish but not in channel catfish, and vice versa; second, singletons in either blue catfish or channel catfish are now brought together to form new contigs; third, splice variations

involving two species may have led to the formation of a larger number of contigs under our assembly parameters. Of these reasons, it appeared that the differences in coverage of the transcriptome in two species may account for the major fraction of this increase in contig numbers. When BLASTN searches were conducted between blue catfish and channel catfish unique sequences, only 12,840 blue catfish contigs (58.3%) had significant hits to channel catfish contigs, and 14,713 channel catfish contigs (50.8%) had significant hits to blue catfish contigs (Table 8).

Analysis of the all catfish unique sequences suggests that a major fraction of the catfish transcriptome has been captured. The 111,578 unique catfish sequences had hits to 22,642 unique proteins in *nr* database, and to 17,948 unique proteins in Uniprot database. When compared to well-characterized fish species such as zebrafish, medaka, and *Tetraodon*, the 111,578 unique catfish sequences had hits to 54-57% of their respective unique proteins (Table 9). Taken the comparison with zebrafish as an example, 39,546 catfish unique sequences (including 23,487 contigs and 16,059 singletons) had hits to 14,988 of the 27,996 total unique proteins of zebrafish (54%), i.e., on a one-on-one relationship, the 39,546 of the 111,578 unique catfish sequences covered 54% of the zebrafish transcriptome. In other words, equivalent to 54% of the zebrafish transcriptome has been captured by approximately 51.8% of the contigs and 24.2% of the singletons of the catfish EST assembly. While 46% of the zebrafish transcriptome was not covered, there are still large numbers of contigs (21,819) and singletons (50,213) of catfish having no hits to the zebrafish reference proteins. Part of the reason for these large numbers of EST contigs and singletons without significant hits to the zebrafish reference protein databases could be resulted from high sequence variation and short ORF representation in

these ESTs. For instance, when the cutoff E-value was increase from 1E-10 to 1E-5, the number of genes that can be identified increased from 14,716 to 16,173 (Table 9). This alone should be enough reason to carry out projects for the characterization of full-length cDNAs in the future, or whole genome sequencing in catfish. In this regard, analysis of full-length cDNA inserts allowed identification of 10,037 cDNA clones containing unique full-length cDNAs in channel catfish and 7,382 unique full-length cDNAs in blue catfish (Table 10). Direct sequencing of these clones in the near future should greatly enhance the genome resources for catfish research.

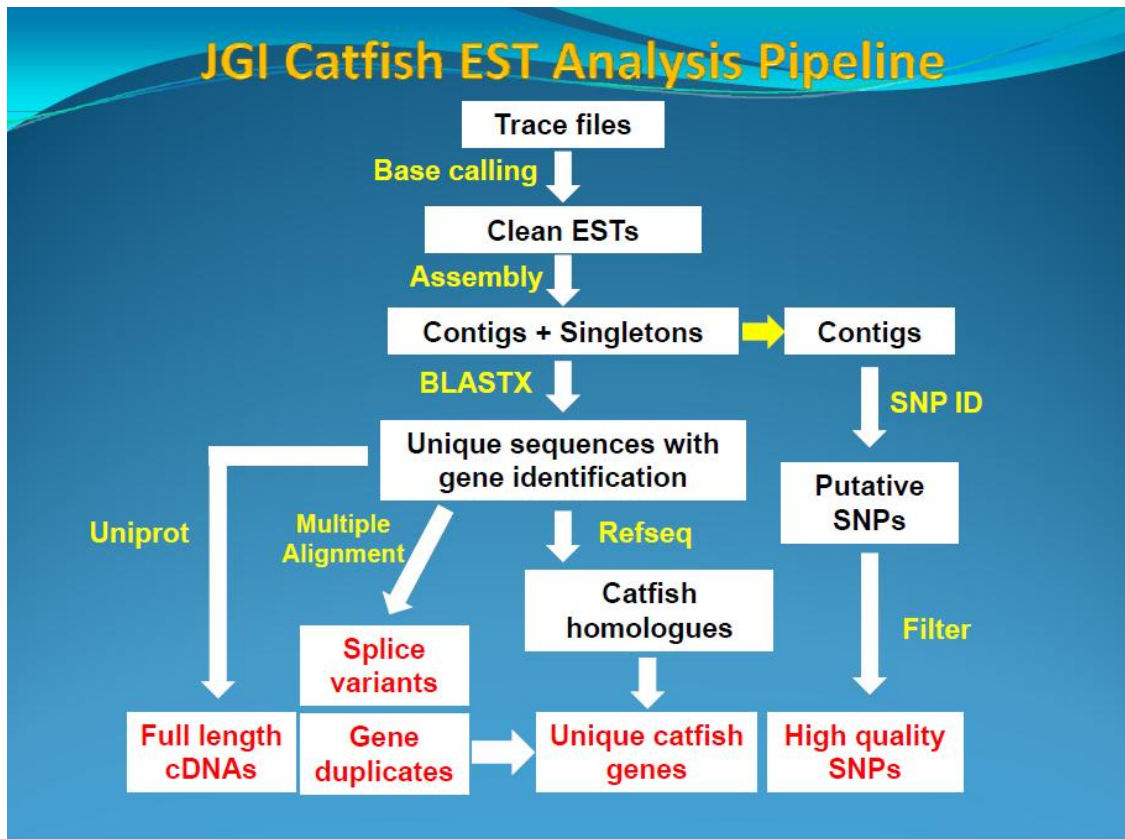
Large-scale EST sequences provide an enormous resource for molecular marker development. This project allowed identification of over 20,000 microsatellites within ESTs, of which 13,375 were located within unique ESTs and had sufficient flanking sequences for microsatellite primer design for genotyping (Table 11). Therefore, these microsatellites will be a major resource for genetic linkage and comparative mapping [12]. In addition, over 300,000 putative SNP sites were identified, of which over 48,000 were identified from contigs with at least four ESTs and the minor sequence was represented at least twice (Table 12). The 48,000 filtered SNPs should be highly useful for the development of a SNP panel for whole genome association studies [34].

The parameters of quality SNP assessment may not be applied to the very large contigs. The utilization of minor allele frequency of six for all the contig containing 30 sequences or more resulted in higher SNP frequency from these contigs (Table 13), such as 13.4 SNP per kb in the contigs with 100-500 sequences, and 15 SNP per kb in the contigs with 500 sequences or more. The information of contigs over 500 sequences can be found in Appendix table 1. High SNP frequency within these large contigs might be

caused by the accumulation of sequencing errors, so these SNPs from large contigs should not be selected for future SNP genotyping.

This large scale EST sequencing project would allow the identification of a majority of catfish transcriptome. The parallel analysis of ESTs from the two closely related ictalurid catfishes should also provide powerful means for the evaluation of ancient and recent gene duplications, and for the development of high-density microarrays in catfish. The inter- and intra- specific SNPs identified from all catfish EST dataset assembly will greatly benefit the catfish introgression breeding selection and whole genome association studies.

Figure 12 JGI Catfish EST Analysis Pipeline



Methods

cDNA Library Construction, EST Sequencing and Processing

The cDNA libraries were constructed from various tissues, organs, and cell lines, including stomach, muscle, olfactory tissue, trunk kidney, head kidney, gill, intestine, spleen, skin, liver, pituitary, ovary and testis (Table 6). Within these libraries, one had no modification, one was subtracted, and 10 were normalized. All cDNA libraries were constructed using the pSPORT-1 and pDNR superscript plasmid cloning system (Invitrogen, Carlsbad, CA). This cloning system provides a vector with capacity of uni-directional cloning of cDNAs that support choices of EST sequencing from either the 5'-, or 3'-end of the transcript. In this work, all ESTs were sequenced from both ends of the transcript (or clone), which provide sequences for further full-length cDNA assembly and characterization. Clone selection, arraying, and sequencing of the 12 libraries were performed at the Joint Genome Institute (JGI) from the Department of Energy (DOE). The cDNAs were sequenced from both the 5' and 3' ends using Big Dye Terminator (V3.1) sequencing chemistry (Applied Biosystem, Foster city, CA). Base calling and sequence trimming were also conducted at the JGI. Phred was utilized for sequence base calling with cutoff Q20, and cross-match was utilized for removing vectors [28,29].

EST Assembly

Assembly was conducted on the blue catfish EST dataset, channel catfish EST dataset, and all catfish EST dataset (Figure 12). The JGI EST sequences and GenBank EST sequences from channel catfish and blue catfish (directly downloaded from dbEST database) were used in clustering and assembly by PTA (Paracel Transcript Assembler,

based on CAP3 program) [30]. Contaminant sequences like *E. coli*, mitochondrial, cloning vector, and RNA were filtered during the cleanup stage by matching these sequences in the database. Repeat sequences and poly (A) tails are masked and annotated. Before the assembly, all the EST sequences were compared to the catfish full-length or partial cDNA sequences in the GenBank, which referred to as seed sequences. Sequences sharing 80% similarity to these seed sequences were grouped to clusters first, and the assembled to generate seed-cluster contigs with criteria of 95% identity with at least 50 bp overlap. The seed cluster assembly would reduce the number of sequences for final assembly, which could reduce the calculation and speedup the assembly process. All the remaining EST sequences are then clustered based on local similarity scores of pairwise comparison using 88% similarity with alignment at least 100 bp. Clusters containing only one sequence are grouped as singletons. The EST clusters were assembled into contiguous sequences (contigs) by multiple-sequence alignment, which generates a consensus sequence for each cluster; with criteria of 95% identity with at least 50 bp overlap. Multiple contigs may be generated from each cluster, since EST clusters may not share enough similarity over their entire length to be assembled as single contig. Multiple contigs may also be generated when ESTs in a cluster represent splice variant forms or paralogs of the gene. The ESTs remaining in a cluster after the formation of contigs were designated as cluster singletons. The unique sequences for each assembly included the seed-cluster contigs, cluster contigs, cluster singletons, and singletons.

ORF searching, gene identification and gene ontology annotation

All the unique sequences obtained after the assembly were analyzed by ESTScan [31]

to search for open reading frames (ORF) which could be used to distinguish coding and non-coding sequences [31, 32]. All the unique sequences were used to search against the *nr* database and Uniprot database using BLASTX to obtain the putative identity at a cutoff E-value of 1E-10. The NCBI Refseq protein and Ensemble database (zebrafish, medaka, *Tetraodon*, human, mouse, and chicken) were also used to identify the catfish unique genes and homologous genes in other species. The *nr* BLASTX results were input in Blast2GO [39] to obtain the Gene Ontology.

Full length cDNA identification

The program TargetIdentifier [23, 34] was used to identify the putative full-length cDNA using BLAST comparisons to full-length genes in Uniprot databases and Start signals. The cutoff E-value of 1E-5 was applied to identify all putative full-length cDNA. Once the start codon (ATG) was identified, the cDNA sequence was considered as a full-length cDNA insert. If a single pass sequences from start codon (ATG) to stop codon were completely sequenced from a single clone rather than from contig assembly, the sequences were considered as a full-length cDNAs.

Microsatellite and SNP marker identification

All unique sequences were used to search the microsatellite makers by using Msatfinder [35]. The repeat threshold for di-nucleotide repeats was eight, and five for tri-, tetra- penta-, and hexa-nucleotide repeats. The microsatellites with 50 bp sequences on both sides were considered microsatellites with sufficient flanking sequences for primer design [36].

The blue catfish, channel catfish, and all catfish EST assembly results were used for further SNP identification. The identification of putative SNPs from the EST sequences was conducted using autoSNP [37], which utilizes the assembly output files as input to detect SNPs based on the base redundancy in the sequence alignments. With the autoSNP program, the parameters for minimum, minor allele frequency for SNP detection varied with the contig size (the number of sequences in the contig) [37]. In order to estimate the inter-specific and intra-specific SNPs within the filtered SNPs, 1,000 contigs were randomly selected to identify the inter- and intra-specific SNPs by visual inspection.

Acknowledgement

This project was supported by the Community Sequencing Program of the Joint Genome Institute of the DOE, and partially by grants from USDA NRI Animal Genome Basic Genome Reagents and Tools Program (USDA/NRICGP award # 2006-35616-16685 and USDA/NRICGP award# 2009-35205-05101). All the sequencing was conducted at the JGI. Thanks are given to Alabama Supercomputer Center for providing the computer capacity for the bioinformatics analysis of the ESTs.

References

1. Serapion J, Kucuktas H, Feng J, Liu Z: **Bioinformatic mining of type I microsatellites from expressed sequence tags of channel catfish (*Ictalurus punctatus*)**. *Mar biotechnol* (New York, NY) 2004, **6**:364-377.
2. Somridhivej B, Wang S, Sha Z, Liu H, Quilang J, Xu P, Li P, Hu Z, Liu Z: **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* 2008, **275**:76-80.
3. Kucuktas H, Wang S, Li P, He C, Xu P, Sha Z, Liu H, Jiang Y, Baoprasertkul P, Somridhivej B, Wang Y, Abernathy J, Guo X, Liu L, Muir W, Liu Z: **Construction of Genetic Linkage Maps and Comparative Genome Analysis of Catfish Using Gene-associated Markers**. *Genetics* 2009. In press
4. Liu Z, Karsi A, Li P, Cao D, Dunham R: **An AFLP-based genetic linkage map of channel catfish (*Ictalurus punctatus*) constructed by using an interspecific hybrid resource family**. *Genetics* 2003, **165**:687-694.
5. Waldbieser GC, Bosworth BG, Nonneman DJ, Wolters WR: **A microsatellite-based genetic linkage map for channel catfish, *Ictalurus punctatus***. *Genetics* 2001, **158**:727-734.
6. Quiniou SM, Katagiri T, Miller NW, Wilson M, Wolters WR, Waldbieser GC: **Construction and characterization of a BAC library from a gynogenetic channel catfish *Ictalurus punctatus***. *Genet Sel Evol* 2003, **35**:673-683.
7. Wang S, Xu P, Thorsen J, Zhu B, de Jong PJ, Waldbieser G, Kucuktas H, Liu Z:

- Characterization of a BAC library from channel catfish *Ictalurus punctatus*: indications of high levels of chromosomal reshuffling among teleost genomes.** *Mar biotechnol (New York, NY)* 2007, **9**:701-711.
8. Quiniou SM, Waldbieser GC, Duke MV: **A first generation BAC-based physical map of the channel catfish genome.** *BMC genomics* 2007, **8**:40.
 9. Xu P, Wang S, Liu L, Thorsen J, Kucuktas H, Liu Z: **A BAC-based physical map of the channel catfish genome.** *Genomics* 2007, **90**:380-388.
 10. Xu P, Wang S, Liu L, Peatman E, Somridhivej B, Thimmapuram J, Gong G, Liu Z: **Channel catfish BAC-end sequences for marker development and assessment of syntenic conservation with other fish species.** *Anim Genet* 2006, **37**:321-326.
 11. Cao D, Kocabas A, Ju Z, Karsi A, Li P, Patterson A, Liu Z: **Transcriptome of channel catfish (*Ictalurus punctatus*): initial analysis of genes and expression profiles of the head kidney.** *Anim Genet* 2001, **32**:169-188.
 12. Ju Z, Karsi A, Kocabas A, Patterson A, Li P, Cao D, Dunham R, Liu Z: **Transcriptome analysis of channel catfish (*Ictalurus punctatus*): genes and expression profile from the brain.** *Gene* 2000, **261**:373-382.
 13. Karsi A, Cao D, Li P, Patterson A, Kocabas A, Feng J, Ju Z, Mickett KD, Liu Z: **Transcriptome analysis of channel catfish (*Ictalurus punctatus*): initial analysis of gene expression and microsatellite-containing cDNAs in the skin.** *Gene* 2002, **285**:157-168.
 14. Kocabas AM, Kucuktas H, Dunham RA, Liu Z: **Molecular characterization and differential expression of the myostatin gene in channel catfish (*Ictalurus punctatus*).** *Biochimica et biophysica acta* 2002, **1575**:99-107.

15. Li P, Peatman E, Wang S, Feng J, He C, Baoprasertkul P, Xu P, Kucuktas H, Nandi S, Somridhivej B, Serapion J, Simmons M, Turan C, Liu L, Muir W, Dunham R, Brady Y, Grizzle J, Liu Z: **Towards the ictalurid catfish transcriptome: generation and analysis of 31,215 catfish ESTs.** *BMC genomics* 2007, **8**:177.
16. Clark MS, Edwards YJ, Peterson D, Clifton SW, Thompson AJ, Sasaki M, Suzuki Y, Kikuchi K, Watabe S, Kawakami K, Sugano S, Elgar G, Johnson SL: **Fugu ESTs: new resources for transcription analysis and genome annotation.** *Genome Res* 2003, **13**:2747-2753.
17. Lo J, Lee S, Xu M, Liu F, Ruan H, Eun A, He Y, Ma W, Wang W, Wen Z, Peng J: **15000 unique zebrafish EST clusters and their future use in microarray for profiling gene expression patterns during embryogenesis.** *Genome Res* 2003, **13**:455-466.
18. Poustka AJ, Groth D, Hennig S, Thamm S, Cameron A, Beck A, Reinhardt R, Herwig R, Panopoulou G, Lehrach H: **Generation, annotation, evolutionary analysis, and database integration of 20,000 unique sea urchin EST clusters.** *Genome Res* 2003, **13**:2736-2746.
19. Rise ML, von Schalburg KR, Brown GD, Mawer MA, Devlin RH, Kuipers N, Busby M, Beetz-Sargent M, Alberto R, Gibbs AR, Hunt P, Shukin R, Zeznik JA, Nelson C, Jones SR, Smailus DE, Jones SJ, Schein JE, Marra MA, Butterfield YS, Stott JM, Ng SH, Davidson WS, Koop BF: **Development and application of a salmonid EST database and cDNA microarray: data mining and interspecific hybridization characteristics.** *Genome Res* 2004, **14**:478-490.

20. Udall JA, Swanson JM, Haller K, Rapp RA, Sparks ME, Hatfield J, Yu Y, Wu Y, Dowd C, Arpat AB, Sickler BA, Wilkins TA, Guo JY, Chen XY, Scheffler J, Taliercio E, Turley R, McFadden H, Payton P, Klueva N, Allen R, Zhang D, Haigler C, Wilkerson C, Suo J, Schulze SR, Pierce ML, Essenberg M, Kim H, Llewellyn DJ, Dennis ES, Kudrna D, Wing R, Paterson AH, Soderlund C, Wendel JF: **A global assembly of cotton ESTs.** *Genome Res* 2006, **16**:441-450.
21. Gorodkin J, Cirera S, Hedegaard J, Gilchrist MJ, Panitz F, Jørgensen C, Scheibye-Knudsen K, Arvin T, Lumholdt S, Sawera M, Green T, Nielsen BJ, Havgaard JH, Rosenkilde C, Wang J, Li H, Li R, Liu B, Hu S, Dong W, Li W, Yu J, Wang J, Staefeldt HH, Wernersson R, Madsen LB, Thomsen B, Hornshøj H, Bujie Z, Wang X, Wang X, Bolund L, Brunak S, Yang H, Bendixen C, Fredholm M: **Porcine transcriptome analysis based on 97 non-normalized cDNA libraries and assembly of 1,021,891 expressed sequence tags.** *Genome Biol* 2007, **8**:R45.
22. Gut IG, Lathrop GM: **Duplicating SNPs.** *Nat Genet* 2004, **36**:789-790.
23. Koop BF, von Schalburg KR, Leong J, Walker N, Lieph R, Cooper GA, Robb A, Beetz-Sargent M, Holt RA, Moore R, Brahmabhatt S, Rosner J, Rexroad CE 3rd, McGowan CR, Davidson WS: **A salmonid EST genomic study: genes, duplications, phylogeny and microarrays.** *BMC genomics* 2008, **9**:545.
24. Taylor JS, Braasch I, Frickey T, Meyer A, Van de Peer Y: **Genome duplication, a trait shared by 22000 species of ray-finned fish.** *Genome Res* 2003, **13**:382-390.
25. Wang S, Sha Z, Sonstegard TS, Liu H, Xu P, Somridhivej B, Peatman E,

- Kucuktas H, Liu Z: **Quality assessment parameters for EST-derived SNPs from catfish.** *BMC genomics* 2008, **9**:450.
26. He C, Chen L, Simmons M, Li P, Kim S, Liu ZJ: **Putative SNP discovery in interspecific hybrids of catfish by comparative EST analysis.** *Anim Genet* 2003, **34**:445-448.
 27. Chatakondi NG, Yant DR, Dunham RA: **Commercial production and performance evaluation of channel catfish, *Ictalurus punctatus* female x blue catfish, *Ictalurus furcatus* male F-1 hybrids.** *Aquaculture* 2005, **247**:8.
 28. Ewing, B. and P. Green. 1998. **Base-calling of automated sequencer traces using phred. II. Error probabilities.** *Genome Res* **8**: 186-194.
 29. Ewing, B., L. Hillier, M.C. Wendl, and P. Green. 1998. **Base-calling of automated sequencer traces using phred. I. Accuracy assessment.** *Genome Res* **8**: 175-185.
 30. Huang X, Madan A: **CAP3: A DNA sequence assembly program.** *Genome Res* 1999, **9**:868-877.
 31. Iseli C, Jongeneel CV, Bucher P: **ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences.** *Proc Int Conf Intell Syst Mol Biol, ISMB* 1999:138-148.
 32. Lottaz C, Iseli C, Jongeneel CV, Bucher P: **Modeling sequencing errors by combining Hidden Markov models.** *Bioinformatics (Oxford, England)* 2003, **19 Suppl 2**:ii103-112.
 33. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional**

- genomics research**. *Bioinformatics (Oxford, England)* 2005, **21**:3674-3676.
34. Min XJ, Butler G, Storms R, Tsang A: **TargetIdentifier: a webserver for identifying full-length cDNAs from EST sequences**. *Nucleic Acids Res* 2005, **33**:W669-672.
35. Thurston MI, Field D: **Msatfinder: detection and characterisation of microsatellites**. 2005, Distributed by the authors at <http://www.genomics.ceh.ac.uk/msatfinder/>. CEH Oxford, Mansfield Road, Oxford OX13SR.
36. Rozen S, Skaletsky H: **Primer3 on the WWW for general users and for biologist programmers**. *Methods Mol Biol (Clifton, NJ)* 2000, **132**:365-386.
37. Barker G, Batley J, H OS, Edwards KJ, Edwards D: **Redundancy based detection of sequence polymorphisms in expressed sequence tag data using autoSNP**. *Bioinformatics (Oxford, England)* 2003, **19**:421-422.

V. CONCLUSIONS

In the EST-derived SNP quality assessment project

- 1) A total of 384 SNPs were selected based on the catfish EST (all catfish GenBank EST by April 2007) assembly results with an average validation rate of 70%.
Overall, the average SNP validation rate was only 33.3% for contigs of 4 or fewer sequences with minor sequence allele present only once. The overall SNP validation rate for contigs of 4 or more sequences with minor sequence allele present at least twice was 70.9%, and up to 89.2% with contigs of 12 or more sequences, which suggested the EST-derived SNPs with minor sequence allele present at least twice will greatly improve the SNP validation rate.
- 2) Comparative genomics studies of 50 failed SNPs revealed that 32 (64%) SNPs were located at the exon-intron border, suggesting that the presence of the presumed introns was the major cause for the failures of the EST-derived SNP genotyping.

In the catfish transcriptome analysis project:

- 1) A total of 438,321 ESTs were generated from JGI EST sequencing project, which allowed the capture of majority of transcriptome of catfish. A total of 14,776 unique genes were identified from the catfish based on the BLASTX searches against Ensemble database.
- 2) A total of 7,382 blue catfish and 10,037 channel catfish unique cDNA clones with full-length inserts were identified from the assembly with a cutoff E-value of $1E-5$, which represented 5,293 unique genes in blue catfish, 6,098 unique genes in channel catfish, and a total of 8,336 unique genes from catfish.
- 3) A total of 20,757 microsatellites were initially identified and 13,375 unique sequences with microsatellites have sufficient flanking sequences (50 bp) on both sides of the microsatellites to design primers for genotyping.
- 4) A total of 48,702 putative SNPs were identified from all catfish EST assembly including inter-specific and intra-specific SNPs

This large scale EST sequencing project would allow the identification of majority of catfish transcriptome. This also provides an platform for the characterization of full-length cDNA. The parallel analysis of ESTs from the two closely related ictalurid catfishes should also provide powerful means for the evaluation of ancient and recent gene duplications, and for the development of high-density microarrays in catfish. The high-density SNP genotyping will greatly benefit the complex trait study and introgression breeding selection and whole genome association studies in the catfish.

APPENDIX

Figure 1.

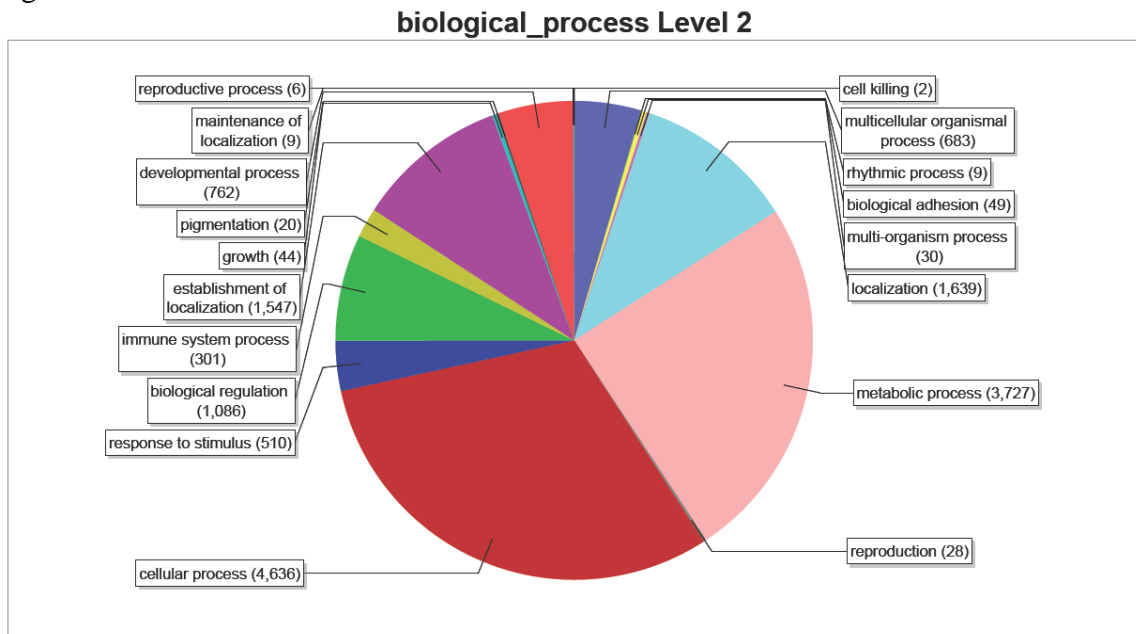


Figure 2.

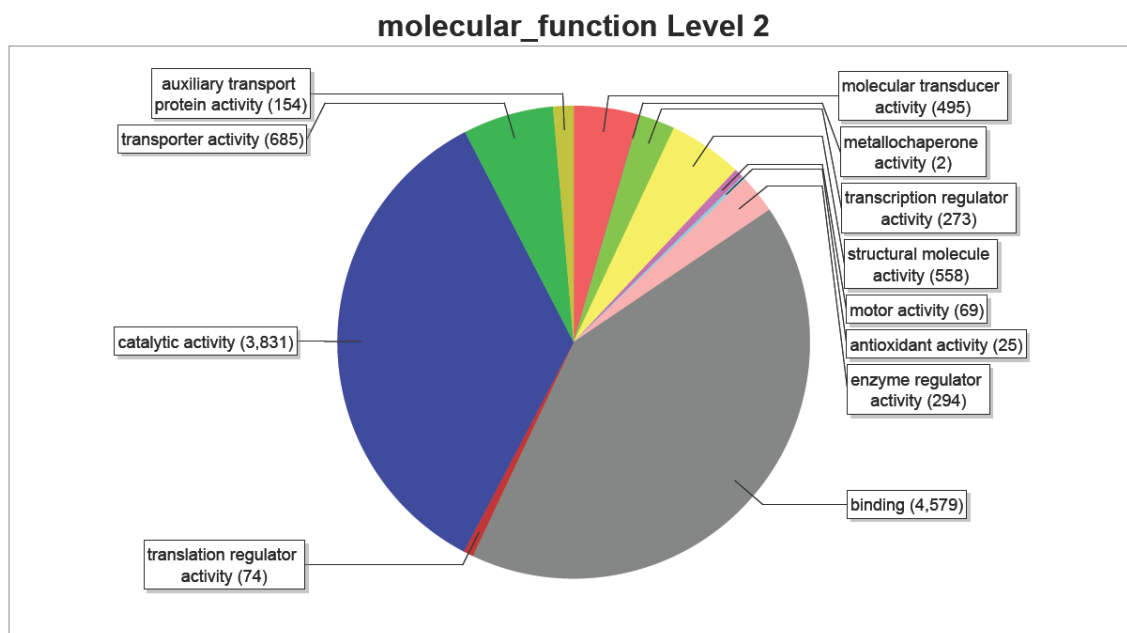


Figure 3

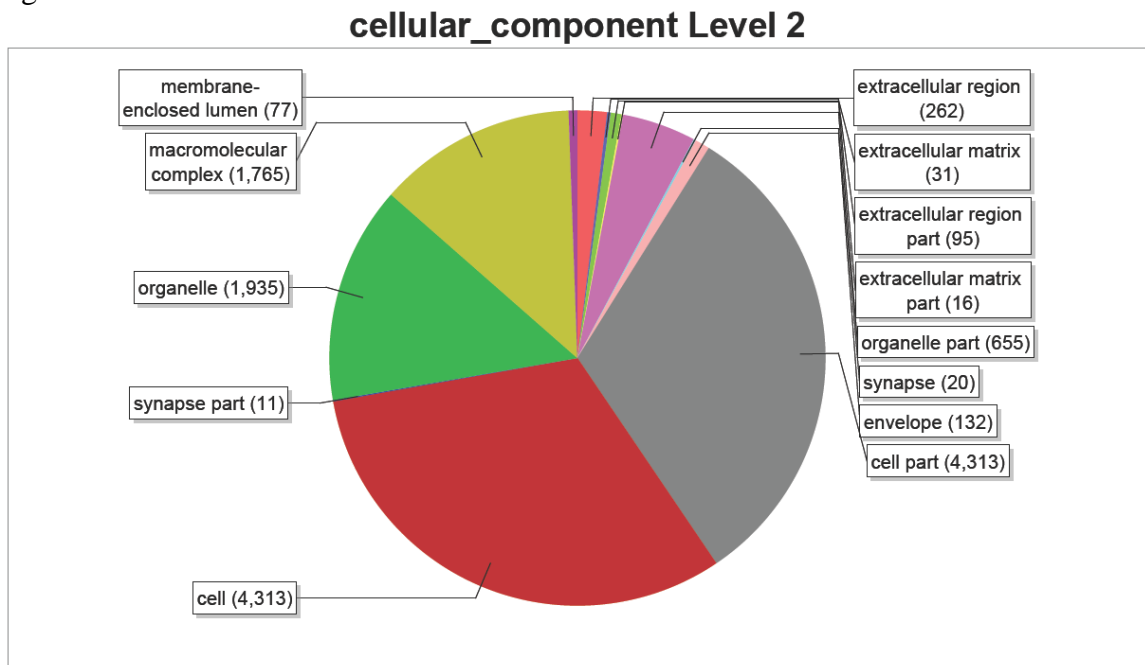


Table 1 Contig information with 500 sequences or more

Contig_ID	Length	No_SEQ	No_SNP	Identitiy
Contig08984	1,076	7,208	740	Apolipoprotein
Contig10817	925	5,315	559	Parvalbumin
Contig03478	1,767	4,820	640	Actin alpha
Contig03221	1,856	1,959	219	creatine kinase M3-CK
Contig16290	1,356	1,870	243	Apolipoprotein A-I
Contig09046	1,520	1,390	126	Apolipoprotein E-1
Contig11831	2,041	1,389	138	beta-actin
Contig19091	2,031	1,087	95	elongation factor 1-alpha
Contig18459	1,503	1,073	128	myosin regulatory light chain
Contig16302	1,102	944	103	liver-type fatty acid-binding protein
Contig06893	1,359	926	83	Prothymosin
Contig17816	1,119	906	126	40S ribosomal protein S2
Contig17170	1,418	898	89	60S acidic ribosomal protein P0
Contig14182	2,096	834	117	creatine kinase M2-CK
Contig09727	1,171	824	72	60S ribosomal protein L7a
Contig16217	1,084	794	107	Elastase 2 like
Contig03229	1,118	762	95	trypsinogen
Contig17925	762	747	100	alpha-globin
Contig08847	1,988	731	69	leucine rich repeat and Ig domain containing 1
Contig17244	1,104	681	36	beta-actin
Contig17797	1,275	638	84	beta thymosin
Contig18149	2,132	636	74	beta-actin
Contig10609	1,126	598	75	Nonspecific cytotoxic cell receptor protein 1
Contig15136	2,447	593	80	procollagen-proline
Contig08974	774	575	48	apolipoprotein C-I
Contig18427	1,413	562	75	actin, alpha, cardiac muscle
Contig16332	1,294	560	68	guanine nucleotide binding protein
Contig01797	4,095	555	159	skeletal muscle myosin heavy chain
Contig18935	823	539	45	NK-lysin type 3
Contig02866	1,452	512	63	skeletal muscle tropomyosin1-1
Contig09022	974	506	64	ribosomal protein L7