Understanding Teleost Genome Structure and Organization: Alternative Splicing, Gene Duplication, and Whole Genome Assembly

by

Jianguo Lu

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 4, 2012

Keywords: gene duplication, alternative splicing, next-generation sequencing, whole-genome sequence assembly, teleost, catfish

Approved by

Zhanjiang Liu, Chair, Alumni Professor of Fisheries and Allied Aquacultures
Nannan Liu, Professor of Plant Pathology and Entomology
Eric Peatman, Assistant Professor of Fisheries and Allied Aquacultures
Xiao Qin, Associate Professor of Computer Science and Software Engineering

Abstract

We conducted both same-species and cross-species analysis utilizing the Genome Mapping and Alignment Program (GMAP) and an AS pipeline (ASpipe) to study AS in four genome-enabled species (*Danio rerio*, *Oryzias latipes*, *Gasterosteus aculeatus*, *Takifugu rubripes*) and one species lacking a complete genome sequence, Ictalurus punctatus. AS frequency was lowest in the highly duplicated genome of zebrafish (17% of mapped genes). The compact genome of the pufferfish showed the highest occurrence of AS (43% of mapped genes). An inverse correlation between AS frequency and genome size was consistent across all analyzed species. Approximately 50% of AS genes identified by same-species comparisons were shared among two or more species.

We have analyzed gene duplication patterns and duplication types among the available teleost genomes and found that a large number of genes were tandemly and intrachromosomally duplicated, suggesting their origin of independent and continuous duplication. This is particularly true for the zebrafish genome. Further analysis of the duplicated gene sets indicated that a significant portion of duplicated genes in the zebrafish genome were of recent, lineage-specific duplication events. Most strikingly, a subset of duplicated genes is enriched among the recently duplicated genes involved in immune or sensory response pathways.

Because of the rapid improvements in cost and quality of sequencing data, de novo sequencing and assembly is possible not only in large sequencing centers, but also in small labs. This project addressed the Message Passing Interface (MPI) version assembler software, MPI-Velvet. It can process high coverage data sets and quickly reconstruct the underlying sequences.

The catfish genome database, cBARBEL(abbreviated from catfish Breeder And Researcher Bioinformatics Entry Location) is an online open-access database for genome biology of ictalurid catfish (*Ictalurus spp.*). It serves as a comprehensive, integrative platform for all aspects of catfish genetics, genomics and related data resources. cBARBEL provides BLAST-based, fuzzy and specific search functions, visualization of catfish linkage, physical and integrated maps, a catfish EST contig viewer with SNP information overlay, and GBrowse-based organization of catfish genomic data based on sequence similarity with zebrafish chromosomes.

Acknowledgments

It is a great pleasure to thank those who made this thesis possible. First and foremost, I am heartily thankful to my advisor, Dr. Zhanjiang Liu, whose encouragement, guidance, and support from the preliminary to the concluding level enabled me to complete this Ph.D. dissertation. I have been working for Dr. Liu for five years. Under his supervision, I learned how to do research and how to write technical papers from scratch. Without him, it would be impossible for me to finish this thesis.

I would like to express my gratitude to my committee: Dr. Nannan Liu, Dr. Eric Peatman, and Dr. Xiao Qin 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 have been working in a great research group. I would like to thank the group members especially Dr. Huseyin Kucuktas, Mila, who have helped me a lot in my research and study. Working with them is beneficial and pleasant.

My deepest gratitude goes to my beloved wife Min Zheng and my lovely son Ryan Lu and also my parents Shitou Lu and Shuhua Zhu for their years selfless support.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Fishes are a tremendously diverse group of vertebrate aquatic animals, breathing trough gills throughout their whole life and having fins and scales. With more than 23,500 species, ray-finned fishes (actinopterygians) are more than 95% of all living fish species. And also, more than 99.8% of ray-finned fishes were grouped into teleosts. Fish biodiversity is also very important to humans in terms of the economical, ecological, and cultural points. Several fish species and sequencing projects can provide the essential information of fish genomes organization and evolution. And those information can help us to underlying the evolutionary mechanisms in the fish lineage. Teleost fish, about half of the extant vertebrate species, exhibit tremendous biodiversity affecting their ecology, morphology, behavior, and many other biological aspects. This huge variability makes teleost fish extremely important for many biological related questions, especially of those related to evolution, gene regulation, and gene structure. There are several teleost fish species are particularly studied with huge genetic and genomic information. And some of them have been, are or will certainly be subjects of whole-genome sequencing projects (Figure 1.1 and Table 1.1), including zebrafish, Tetraodon [56], fugu [6], medaka [61], stickleback, and catfish. Access to entire genome sequences is revolutionizing our understanding of how genetic information is stored and organized in DNA, and how it has evolved over time.

Channel catfish (*I. punctatus*) is the most important freshwater aquacultured species in the United States. Within recent years, around 500-600 million pounds of catfish were harvested in the U.S., which represents about four billion dollars (Figure 1.2). Catfish production accounts for over 60% of all U.S. aquaculture production. It is anticipated that it

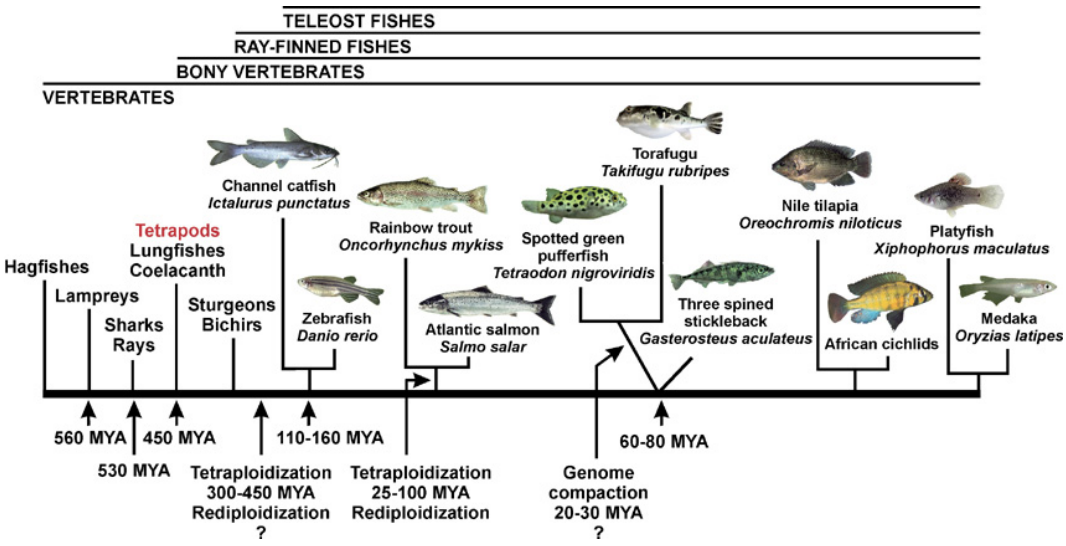


Figure 1.1: The fish lineage and origin of fish pictures

will become one of the most important aquaculture fish species in Asia as well. Channel catfish is the most economically important catfish species in the U.S. Blue catfish (I. furcatus) is also very important species because the hybrid between channel catfish and blue catfish can generate superior performance catfish in terms of several commercial traits [47]. It is possible to generate artificial breeds using introgression approach because these interspecific hybrids are fertile [76] [77]. Huge genetic/genomic catfish information have been generated in our lab, including large number of polymorphic markers [120] [146], high density linkage maps [136] [77] [67], BAC contig-based physical maps [147], over 63,000 BAC end sequences [146] [73], and a large number of ESTs [59] [21] [60] [66] [70] [138]. In addition to these traditional genome resources, the genome-scale gene-associated single nucleotide polymorphism (SNPs) in catfish is also available [75]. Also the high-density catfish SNP array is on the way, which will provide the material basis for genome association studies and whole genome-based selection in catfish.

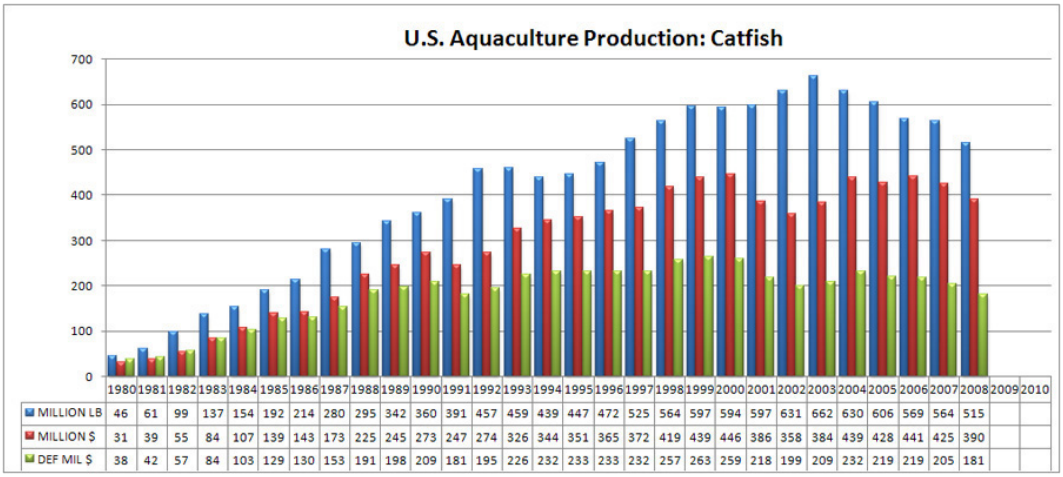
1.2 Teleost fish species gene duplication

Gene duplication is believed to play a major role in evolution. Ohno argued that gene duplication is the most important evolutionary force behind evolutionary novelty for species

Table 1.1: Teleost fish species with genomic resources

Scientific name	Common name	Genome size	References
Danio rerio	Zebrafish	1.7 Gbp	zfin.org
Oryzias latipes	Medaka	1.7 Gbp	Kasahara et al, 2007
Takifugu rubripes	Fugu	0.4 Gbp	Aparicio et al, 2002
Gasterosteus aculeatus	Stickleback	0.7 Gbp	Jones et al, 2012
Tetraodon nigroviridis	Pufferfish	0.4 Gbp	Jaillon et al, 2004
Ictalurus punctatus	Channel catfish	1.0 Gbp	Lu et al, 2011

(Ohno 1970). Three main mechanisms are believed to generate gene duplications; unequal crossing over, retrotransposition, and chromosomal (genome) duplication [58] [54]. Of these, localized (or tandem) duplication resulting from unequal crossing over and genome duplication are believed to be the two dominant mechanisms contributing to vertebrate genome evolution [34] [103]. Over four decades ago, Ohno (1970) suggested that two rounds of large-scale gene duplication had occurred early in vertebrate evolution. Sequencing analysis of Hox gene clusters from a spectrum of vertebrate species provided critical evidence in support of Ohno's hypothesis [50] [110] [88] [30] and indicated, in turn, an additional round of fish-specific genome duplication (FSGD) prior to the divergence of most teleost species [5] [131] [134] [28] [29]. Additional evidence supporting FSGD has been garnered from studies of pufferfish, Takifugu rubripes and Tetraodon nigroviridis. In these studies, hundreds of genes and gene clusters are present in duplicate in teleost fish but possessing only single copy in other vertebrates, illustrating fish-specific duplication of syntenic regions between humans and fish [56] [90] [112]. Ongoing examination of gene families across vertebrate evolution continues to provide general support for the three rounds of genome duplication (3R) hypothesis in teleost fish [142] [15] [153] [62] [4] [95]. In addition, chromosomal rearrangements scramble pieces of duplicated chromosomes around the genome hindering the identification of duplication signatures in a genome. Identifying these remnants of duplication evidence has been made possible with the development and maturation of computational tools.



	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
MILLION LB	46	61	99	137	154	192	214	280	295	342	360	391	457	459	439	447	472	525	564	597	594	597	631	662	630	606	569	564	515		
MILLION $	31	39	55	84	107	139	143	173	225	245	273	247	274	326	344	351	365	372	419	439	446	386	358	384	439	428	441	425	390		
DEF MIL $	38	42	57	84	103	129	130	153	191	198	209	181	195	226	232	233	233	232	257	263	259	218	199	209	232	219	219	205	181		

Figure 1.2: United State catfish production from 1980-2008

1.3 Teleost fish species alternative splicing

Alternative splicing (AS) is a cellular mechanism in eukaryotes that produces multiple mature mRNA isoforms from a pre-mRNA molecule. It is not only one of the major mechanisms for generating diversity of gene products, but also an important mechanism for modulating gene expression and function [42] [83] [10]. AS is known to impact protein usage by altering signals for hosphorylation, glycosylation, and trafficking [49]. Recent research in mammalian species is drawing attention to the fact that AS and its regulatory pathways can have important physiological consequences for an organism, including in the areas of cancer progression and immune repertoire generation [116]. The extent of AS has been shown to vary greatly among organisms. In some cases, however, the differences can be attributed to the extent that the genome of the organism has been studied. Approximately 40% of human genes have two or more AS products [93] [94] [16] [100]. Brett et al. (2002) compared seven different eukaryote species with sufficient coverage of ESTs and mRNA data, and reported rates of approximately 45%, 30%, 15%, 15%, 10%, 10%, and 6% AS in human, mouse, cow, fly, rat, worm, and plants, respectively. Despite the enormous potential to generate tremendous splice isoforms, only a small fraction of the variants have been observed in nature. This implies a complicated regulation mechanism that guides the expression of certain

splice instead of a random combinatorial output. Several attempts have been made to reveal the mechanism of regulation of alternative splicing leading to the identification of conserved sequences, which are associated with exon skipping. Both computational and experimental studies are underway to understand the mechanism and regulation of alternative splicing.

1.4 Next-generation sequencing technology

An understanding of the organization, expression and function, and evolutionary history of the aquaculture genomes requires knowing their primary structure, the linear order of the nucleotide base pairs of the genomes. DNA sequencing is undoubtedly the most important technology in biology. Over the past three decades, DNA sequencing technology has gone through remarkable advances: from radioactivity-labeled to fluorescence-labeled in chemistry, from slab gel to capillary in geometry, and from electrophoresis to ordered array in signal detection. The automated Sanger method is considered as a first-generation technology, and newer methods are referred to as next-generation sequencing (NGS). NGS technologies include various strategies that rely on a combination of template preparation, sequencing and imaging, and genome alignment and assembly methods. There are several commercially available technologies including Roche/454, Illumina/Solexa, Life/SOLiD, Helicos BioSciences, and Pacific Biosciences (see Table 1.2). The Roche/454 genome sequencer FLX system is based on sequencing by synthesis (pyrosequencing) technology [86]. In this system, the DNA sample is first sheared into fragments. Then two adaptors provide priming sites for amplification and sequencing, as well as a special key sequence. For DNA amplification, the Genome Sequencer FLX System uses emulsion-based clonal amplification, called emPCR [33]. The single-stranded DNA library is immobilized by hybridization onto primer-coated capture beads. After amplification, the microreactors are broken, releasing the DNA-positive beads for the enrichment. For sequencing, the DNA beads are layered onto a PicoTiterPlate device, depositing the beads into the wells, followed by enzyme beads and packing beads. Across multiple cycles, the pattern of detected incorporation events reveals

Table 1.2: Comparison of next-generation seuqencing platforms

Platform	Sequencing chemisitry	Template amplification	Read length (bp)	Run time (days)	Gb per run
Roche 454	Pyrosequencing	Emulsion PCR	500	0.4	0.45
Illumina	Reversible termination	Bridge PCR	200	4/9	25/50
SOLiD	Sequencing by liagation	Emulsion PCR	50	7/14	25/50
Helicos Bioscience	Reversible termination	Single molecule	32	8	37
Pacific Bioscience	Real-time	Single molecule	1,000	N/A	N/A

the sequence of templates represented by individual beads. Raw reads processed by the 454 platform are screened by various quality filters to remove poor-quality sequences, mixed sequences, and sequence without the initiating key sequence. Three different bioinformatic tools are available including GS de novo Assmebler, GS reference mapper, and GS Amplicon variant analyzer. The researchers can get meaningful results using these graphical analysis tools quickly. However, several limitations exist when using 454 technologies. A major limitation relates to resolution of homopolymer-containing DNA segements, such as CCCC and TTTT [115]. Because there is no terminating moiety preventing multiple consecutive incorporations at a given cycle, pyrosequencing relies on the magnitude of light emitted to determine the number of repetitive bases. But the key advantage of 454 platform is its long read length [89]. The 454 system can generate more than 1,000,000 individual reads, with high quality read length of 400 bases per run (see Table 1.2).

The Illumina/Solexa Genome Analyzer is based on sequencing by synthesis chemistry. Firstly, the input DNA is fragmented by shearing to short fragments ($\leq$ 800bp). Then the DNA fragments are ligated at both ends to adapters that have a single-base overhang. After denaturation, DNA fragments are immobilized at one end on a solid support-flow cell. The adaptors on the surface also act as primers for the following PCR amplification. Then the DNA fragments are amplified by bridge PCR [1] [37]. After incorporation into the DNA strand, the terminator nucleotide, as well as its position on the support surface, is detected and identified via its fluorescent dye by the CCD camera. A base-calling algorithm

assigns sequences and associated quality values to each read, and a quality checking pipeline evaluates the Illumina data from each run, removing poor-quality sequences. The Paired-End Module enables paired-end sequencing up to 2×100bp. Also, the Genome Analyzer II runs much quicker and the output per run can reach 45-50 Gbp (see Table 1.2). Comparing with Sanger sequencing, the Illumina system is more cost-effective and much quicker. However, the error rates are higher and read length is shorter [89]. Usually, the error rate can be overcome by coverage [68]. The Applied Biosystems SOLiD System is based on sequencing by ligation technology [52]. The fragment or mate-paired library can be constructed in this technology, which depends on the researchers purposes. DNA fragments are ligated to adapters and bound to beads. DNA fragments on the beads are amplified by emPCR. This method is called two-base encoding [104]. Two-base encoding is a unique and powerful approach designed to clearly discriminate measurement errors. The SOLiD 3 Plus System can generate more than 60Gb of mappable sequence or greater than 1 billion reads per run. Also it is very cost-effective comparing with other sequencers. However, the current read length, however, significantly limits its applications [89]. Helicos Biosciences is the first company to offer a next-next generation DNA sequencing system for single molecules. PacBio utilizes total internal reflection microarray (TIRM) technology to detect signals [46]. They observe and record the DNA synthesis process by DNA polymerase in real time. The fluorescence resonance energy transfer (FRET) is utilized between donor and receptor [12]. Pacific Biosciences (PacBio) is another new generation sequencing company. The single-molecule real-time (SMRT) technology is used in this platform. The zero-mode waveguide (ZMW) technology, single-molecule sequencing by synthesis with nanostructure, is used for real-time observation of DNA polymerization [35].

1.5 Whole genome sequence assembly

An assembly is a hierarchical data structure that maps the sequence data to a putative reconstruction of the target. Throng assembly, the short read sequence data can be grouped

into contigs and scaffolds. The next generation sequence assemblers can be divided into three groups based on the algorithm utilization, including the Overlap Layout Consensus (OLC) methods, the de Bruijn Graph (DBG) methods, and the greedy graph algorithms [91]. An overlap graph represents the sequencing reads and their overlaps [96]. The nodes in graph represent the reads and the edges represent overlaps. The de Bruijn graph was developed not from actual short DNA sequencing reads. The nodes in de Bruijn graph represent all possible fixed-length strings and the edges represent suffix-to-prefix prefect overlaps [151]. A repeat graph is an application of the K-mer graph [107]. K-mer graphs are more sensitive to repeats and read errors comparing with OLC assemblers. The greedy algorithms apply one basic operation: add one more read or contig onto given read or contig based on highest-scoring overlap to make the next join. The OLC approach was typical for Sanger-data assemblers. Several current assemblers were implemented with OLC algorithm, including Celera Assembler, Arachne, Newbler, and PCAP [97] [8] [55] [86]. OLC includes three phases: 1). Overlap discovery involves all-against-all, pair-wise read comparison; 2). Construction and manipulation of overlap graph to a read layout; 3). Multiple sequence alignment determines the precise layout and then the consensus sequence. The de Bruijn Graph approach is the most widely used in the short reads from the Solexa and SOLiD platforms. It relies on K-mer graphs, whose attributes make it attractive for huge amount of short reads. The K-mer graph would be a de Bruijn graph and it would contain an Eulerian path, a path that traverses each edge exactly once [108]. Obviously, K-mer graphs generation is much more complex from real sequencing data than ideal scenario. There are some problems when genomic repeats are introduced. For example, the repeats induce cycles in the K-mer graph would generate multiple possible reconstruction of the target.

1.6 Message Passing Interface (MPI)

The Message Passing Interface standard (MPI) is a message passing library standard used for the development of message-passing parallel programs [43]. The goal of MPI is to

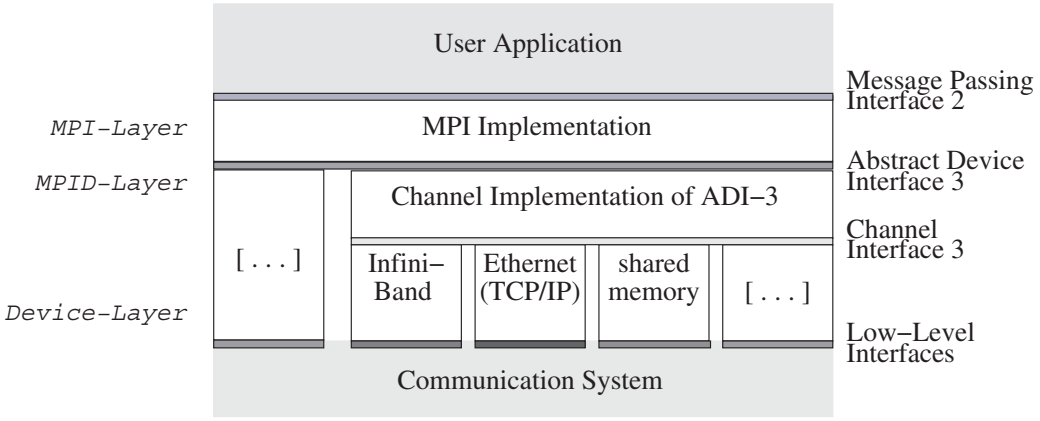


Figure 1.3: Hierarchical Structure of MPICH2

facilitate an efficient, portable, and flexible standard for parallel programs using message passing. MPICH2 - developed by the Argonne National Laboratory - is one of the most popular and widely deployed MPI implementations in cluster computing environments. Most MPI implementations consist of a specific set of routines (i.e., an API) directly callable from FORTRAN, C and C++ and from any language capable of interfacing with such libraries (such as C, Java or Python). The advantages of MPI over older message passing libraries are portability (because MPI has been implemented in almost every distributed memory architecture) and speed (because each implementation is in principle optimized for the hardware upon which it runs). MPICH2 provides an implementation of the MPI standard while supporting a large variety of computation and communication platforms like commodity clusters, high-performance computing systems, and high-speed networks [41].

MPI uses Language Independent Specifications (LIS) for calls and language bindings. The first MPI standard specified ANSI C and Fortran-77 bindings together with the LIS. The draft was presented at Supercomputing 1994 (November 1994) [17] and finalized soon thereafter. About 128 functions constitute the MPI1.3 standard which was released as the final end of the MPI1 series in 2008. At present, the standard has several popular versions: version 1.3 (shortly called MPI1), which emphasizes message passing and has a static run-time environment, and MPI2.2 (MPI2), which includes new features such as parallel I/O,

dynamic process management and remote memory operations. MPI2's LIS specifies over 500 functions and provides language bindings for ANSI C, ANSI Fortran (Fortran90), and ANSI C++. Object interoperability was also added to allow for easier mixed-language message passing programming. A side effect of MPI2 standardization (completed in 1996) was clarification of the MPI1 standard, creating the MPI1.2. Figure 1.3 shows the hierarchical structure of the MPICH2 implementation, where there are four distinct layers of interfaces to make the MPICH2 design portable and flexible. The four layers, from top to bottom, are the message passing interface 2 (MPI-2), the abstract device interface (ADI3), the channel interface (CH3), and the low-level interface. ADI3 - the third generation of the abstract device interface - in the hierarchical structure (see Figure 1.3) allows MPICH2 to be easily ported from one platform to another. Since it is non-trivial to implement ADI3 as a full-featured abstract device interface with many functions, the CH3 layer simply implements a dozen functions in ADI3 [74]. As shown in Figure 1.3, the TCP socket Channel, the shared memory access (SHMEM) channel, and the remote direct memory access (RDMA) channel are all implemented in the layer of CH3 to facilitate the ease of porting MPICH2 on various platforms. Note that each one of the aforementioned channels implements the CH3 interface for corresponding communication architecture like TCP sockets, SHMEM, and RDMA. Unlike an ADI3 device, a channel is easy to implement since one only has to implement a dozen functions relevant for with the channel interface.

Chapter 2

Alternative splicing in teleost fish genomes: Same-species and cross-species analysis and comparisons

2.1 Abstract

Alternative splicing (AS) is a mechanism by which the coding diversity of the genome can be greatly increased. Rates of AS are known to vary according to the complexity of eukaryotic species, potentially explaining the tremendous phenotypic diversity among species with similar numbers of coding genes. Little is known, however, about the nature or rate of AS in teleost fish. We report here the characteristics of AS in teleost fish and classification and frequency of five canonical AS types. We conducted both same-species and cross-species analysis utilizing the Genome Mapping and Alignment Program (GMAP) and an AS pipeline (ASpipe) to study AS in four genome-enabled species (*Danio rerio*, *Oryzias latipes*, *Gasterosteus aculeatus*, *Takifugu rubripes*) and one species lacking a complete genome sequence, Ictalurus punctatus. AS frequency was lowest in the highly duplicated genome of zebrafish (17% of mapped genes). The compact genome of the pufferfish showed the highest occurrence of AS (43% of mapped genes). An inverse correlation between AS frequency and genome size was consistent across all analyzed species. Cross-species comparisons utilizing zebrafish as the reference genome allowed the identification of additional putative AS genes not revealed by zebrafish transcripts. Approximately 50% of AS genes identified by same-species comparisons were shared among two or more species. A searchable website, the Teleost Alternative Splicing Database, was created to allow easy identification and visualization of AS transcripts in the studied teleost genomes. Our results and associated database should further our understanding of alternative splicing as an important functional and evolutionary mechanism in the genomes of teleost fish.

2.2 Introduction

Alternative splicing (AS) is a cellular mechanism in eukaryotes that produces multiple mature mRNA isoforms from a pre-mRNA molecule. It is not only one of the major mechanisms for generating diversity of gene products, but also an important mechanism for modulating gene expression and function [42] [83] [10]. AS is known to impact protein usage by altering signals for phosphorylation, glycosylation, and trafficking [49]. Recent research in mammalian species is drawing attention to the fact that AS and its regulatory pathways can have important physiological consequences for an organism, including in the areas of cancer progression and immune repertoire generation [116]. The extent of AS has been shown to vary greatly among organisms. In some cases, however, the differences can be attributed to the extent that the genome of the organism has been studied. Approximately 40% of human genes have two or more AS products [93] [94] [16] [100]. Brett et al.(2002) compared seven different eukaryote species with sufficient coverage of ESTs and mRNA data, and reported rates of approximately 45%, 30%, 15%, 15%, 10%, 10%, and 6% AS in human, mouse, cow, fly, rat, worm, and plants, respectively. While the highest quality predictions of AS are generated from same-species transcript-genome alignments, these studies have traditionally been hindered by the lack of complete genome sequences, small transcript resources, or both. Cross-species approaches to detection of AS allow the utilization of a related genome sequence or transcript set to aid in analysis. Additionally, cross-species approaches offer an evolutionary assessment of conservation of AS events and mechanisms and may highlight important functional requirements for AS that have been maintained across species [14]. Alignment of mouse, rat, and human ESTs in cross-species fashion, allowed the identification of novel, previously unannotated exons and AS events that were subsequently validated by Reverse transcription PCR (RT-PCR) [25]. Similarly, cross-species approaches in legumes allowed the identification of novel and conserved AS events [137]. No systematic analysis of rates and types of AS has been conducted in teleost fish. Two groups have previously included a teleost species, zebrafish, in their analysis of

general vertebrate AS levels, both without further analysis of teleost results [117] [63]. Teleost species have only recently obtained sufficient levels of transcript sequences [138] and assembly of multiple whole genome sequences to allow meaningful analysis of AS. Given the diversity of fish species, the wide variety of fish genome sizes and complexities, and continued research into fish-specific genome duplication [119] [98], an investigation of teleost AS is particularly relevant to our understanding of teleost genome evolution. In this study, all available transcript sequences from 4 genome-enabled fish species fugu (*Takifugu rubripes*), medaka (*Oryzias latipes*), zebrafish (*Danio rerio*) and stickleback (*Gasterosteus aculeatus*) were aligned to their genome sequences for same-species analysis of AS using an analysis pipeline modified from Wang et al. 2008. We found that rates of AS vary widely across the studied species, even when normalizing for Expressed Sequence Tag (EST) gene coverage. Rates of AS appear to be inversely correlated with genome size. Cross-species alignments of medaka, fugu, stickleback, and channel catfish, a species currently lacking a genome sequence, onto the zebrafish genome allowed the identification of additional putative AS genes. Our results represent the first major genome-wide analysis of AS in teleost fish and a major step toward understanding genome structure and evolution in this large species group.

2.3 Materials and Methods

2.3.1 Datasets for AS analysis

The genome sequences from zebrafish, medaka, fugu and stickleback were downloaded from the Ensembl database. The zebrafish genome sequences version is Zv8. Other genome sequence data sets used in this study were current as of January 20, 2010. All EST/cDNA sequences from zebrafish, medaka, fugu, stickleback, and catfish were retrieved from GenBank and Ensembl genome databases.

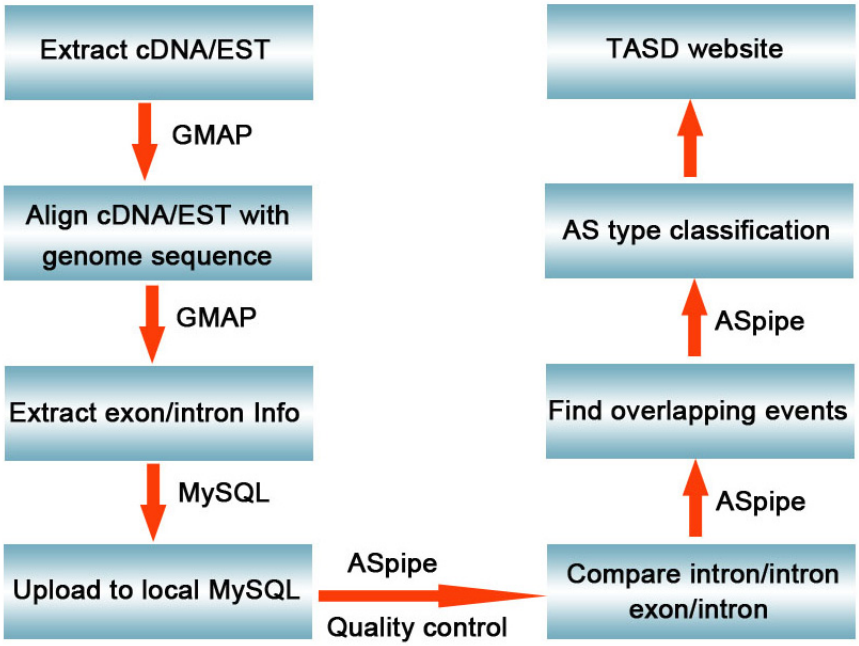


Figure 2.1: Work flowchart for detecting AS events in teleost species

2.3.2 Alignment of transcripts to genome sequences

The EST/cDNA sequences of the four genome-enabled species were mapped to their respective genome sequences using the GMAP computer program [144]. Default parameters were used for GMAP. Then using the GMAP alignment output as the input for ASpipe 1.0, coordinates and scores for high-quality predicted intron/exon/alignment were extracted from the GMAP program outputs and uploaded into a local MySQL5.0 database. In order to decrease AS artifacts, highly stringent parameters were used. For same-species EST/cDNA alignment, the parameters were >95% sequence identity and 80% alignment coverage. For cross-species EST/cDNA alignment, the parameters were decreased to 80% sequence identity with alignment coverage remaining at 80% (Figure 2.1).

2.3.3 Alternative splicing types and alternative splicing identification

The coordinate information of predicted introns and exons were compared in pair-wise fashion to identify AS candidates and particular types of splicing events following the

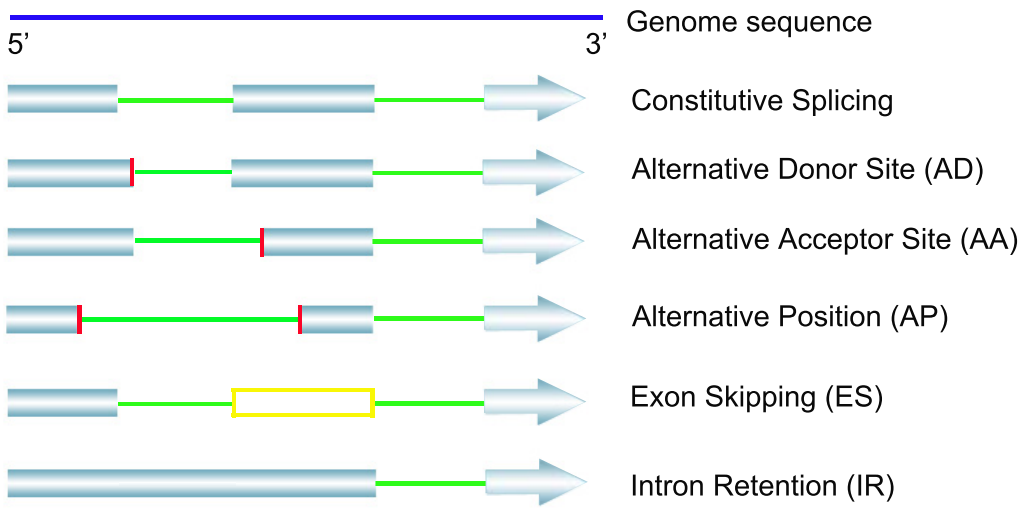


Figure 2.2: Visualization of five AS types in teleost species

methodology of Wang et al. (2008). In the case of intron vs. intron comparisons, if two exons had the same 5' end but different 3' ends, this case was classified as Alternative Donor Site (AltD). If two exons had the same 3' end but differed only in the 5' ends, this case was classified as Alternative Acceptor Site (AltA). If both 3' ends and 5' ends differed but had overlapping introns, this case was classified as Alternative Position Site (AltP). For intron vs. exon comparisons, if the intron was completely replaced by an exon, this case was classified as Intron Retention (IntronR). Alternatively, if the exon was completely replaced by an intron, this case was classified as Exon Skipping (ExonS) (Figure 2.2).

AS events were identified by two approaches: 1) same species AS events identification was based on EST/cDNA alignment with their own genome sequence; 2) cross-species AS events were identified between the zebrafish genome and the transcript sequences of 4 other teleost species. For the first case, all the EST/cDNA of zebrafish, medaka, fugu and stickleback were aligned with their own genome sequences, respectively. Only those with identical coordinates of an alternatively processed intron/exon were regarded as same species AS events. For the second case, using the EST/cDNA from channel catfish, fugu,

medaka and stickleback, alignments were conducted on the zebrafish genome sequence before input into ASpipe for detection of AS events. Based on the AS genes identified using the same-species approach for fugu, medaka, stickleback and zebrafish, BLAST searches were conducted against the Uniprot database. Genes with the same Uniprot top hits were recorded and used to assess potential levels of conservation of AS genes in teleost species.

2.4 Results

2.4.1 Same-species transcript/genome alignments

The EST/cDNA information of four teleost species (zebrafish, medaka, fugu and stickleback) were extracted from NCBI and Ensembl databases for AS analysis (Figure 2.1). A total of 1,780,568 EST/cDNA from zebrafish, 638,483 EST/cDNA from medaka, 304,239 EST/cDNA from stickleback and 73,945 EST/cDNA from fugu were retrieved (Table 2.1). The Genome Mapping and Alignment Program (GMAP) was then used to align the EST/cDNAs with their respective genome sequences. The percentage of mapped transcripts varied from 61.2% to 93.2%. A majority of transcript information from fugu came from predicted Ensembl gene transcripts, leading to a high rate of successful mapping to its genome sequence. The number of identified transcription units (genes) were similar among medaka, fugu, and stickleback (21,613 to 25,443), but largerly higher in zebrafish (41,365), whose genome is characterized by high rates of gene duplications [114] [105]. Analysis of AS was naturally restricted to those transcription units with multiple ESTs (58.8-71.7%). There were large disparities in average numbers of ESTs per gene among the four species, reflecting the depth of transcript resources that have been generated to-date. Zebrafish, with the largest EST resources, averaged 28.9 ESTs/gene while fugu, with minimal transcripts available, averaged only 3.3 ESTs/gene. Additional information obtained from the GMAP alignments reflected the genome characteristics of the four species. Intron sizes were dramatically larger in zebrafish (average of 6,767.4 bp) than in the compact genome of fugu (average of 687.7 bp), with medaka and stickleback falling in between these two extremes (2,317.5 bp and 1,216.7

bp, respectively). Similarly, average exon sizes ranged from 257.5 bp in zebrafish to 157.9 bp in fugu (Table 2.1).

Table 2.1: Transcript genome alignments and intron and exon features in 4 fish species

	Zebrafish	Medaka	Fugu	Stickleback
EST/cDNA	1,780,568	638,483	73,945	304,239
Mapped to genome	1,120,795 (62.9%)	522,516 (81.8%)	68,928 (93.2%)	186,324 (61.2%)
Transcription Units (Genes)	41,365	25,443	21,613	23,188
Multi EST TU(Genes)	24,305 (58.8%)	18,250 (71.7%)	13,293 (61.5%)	15,889 (68.5%)
Average ESTs/gene	28.9	21.0	3.3	8.5
Number of Introns	180,717	198,676	257,697	214,516
Avg intron size (bp)	6,767.4	2,317.5	687.7	1,216.7
Long intron ($\geq 1,000$ nt)	51.0%	28.6%	15.1%	17.3%
Number of internal exons	147,318	174,879	240,429	196,533
Avg internal exon size (bp)	257.5	197.5	157.9	200.5

2.4.2 Rates of alternative splicing vary among teleost species

GMAP alignment outputs were fed into ASpipe to identify AS genes and categorize splicing mechanisms (Table 2.2). Interestingly, the largest number of AS events was identified from fugu (20,676) and the smallest number from zebrafish (12,222). Medaka and stickleback had similar numbers of AS events (13,246 and 12,241, respectively) to that of zebrafish. A similar pattern was observed when the number of unique AS genes was considered. A total of 9,336 unique genes were AS in fugu, compared to 7,036 in zebrafish. Not only was the number of unique AS genes greater in fugu than the other teleost species, but also the number of AS events per gene was larger. On average, 2.21 AS events per gene were identified from fugu, compared to approximately 1.6-1.7 events/gene in the other three species. Most striking was the comparison of percentages of AS genes among the four species. Greater than 43% of all fugu genes were detected to be AS, while only 17% of zebrafish genes were alternatively spliced. Based on our analysis, medaka and stickleback again fell in between these two extremes, with 31.2% and 32.4% of their genes being AS, respectively (Table 2.2).

Table 2.2: Comparison of AS types and frequencies in 4 fish species

	Zebrafish	Medaka	Fugu	Stickleback
AltD	1,525 (12.5%)	1,427 (10.8%)	1,967 (9.5%)	1,203 (9.7%)
AltA	1,966 (16.1%)	2,272 (17.2%)	3,011 (14.6%)	2,046 (16.5%)
AltP	1,767 (14.5%)	2,391 (18.1%)	4,977 (24.1%)	2,427 (19.5%)
ExonS	4,180 (34.2%)	4,062 (30.7%)	7,255 (35.1%)	3,533 (28.4%)
IntronR	2,784 (22.8%)	3,094 (23.4%)	3,466 (16.8%)	3,212 (25.9%)
Total AS events	12,222	13,246	20,676	12,421
Unique AS genes (% of total genes)	7,036 (17.0%)	7,929 (31.2%)	9,336 (43.2%)	7,513 (32.4%)
Events/gene	1.74	1.67	2.21	1.65

We next asked whether the differing coverage of ESTs/gene (Table 2.1) affected the detected frequency of AS genes. We, therefore, sought to normalize our analysis by EST levels. We plotted AS frequency against ESTs per gene for each of the four teleost species (Figure 2.3). Our results revealed that although AS frequency did generally increase with depth of EST coverage, the AS frequencies of the four species varied greatly regardless of EST level until coverage reached greater than 25 ESTs per gene. Taking, for example, genes with 4 ESTs/gene across all four species, greater than 80% of these genes were AS in fugu, compared to around 10% in zebrafish. These trends continued until the number of ESTs/gene rose largely. The number of fugu genes with EST coverage greater than 11 ESTs/gene and stickleback genes with greater than 35 ESTs/gene were too small for analysis, and, therefore, these categories were removed from consideration.

2.4.3 Similar distribution of alternative splicing types among teleost species

AS genes from the four teleost species were additionally characterized as to their type as defined in the Methods (Figure 2.2). Five types were quantified for each species, Alternative Donor (AltD), Alternative Acceptor (AltA), Alternative Position (AltP), Exon Skipping (ExonS), and Intron Retention (IntronR). Exon skipping was the most abundant AS type in all four species, varying from 28.4% to 35.1% of genes (Table 2.2). Intron retention was the next most prevalent type, comprising 16.8% -25.9% of the cases. The least abundant form

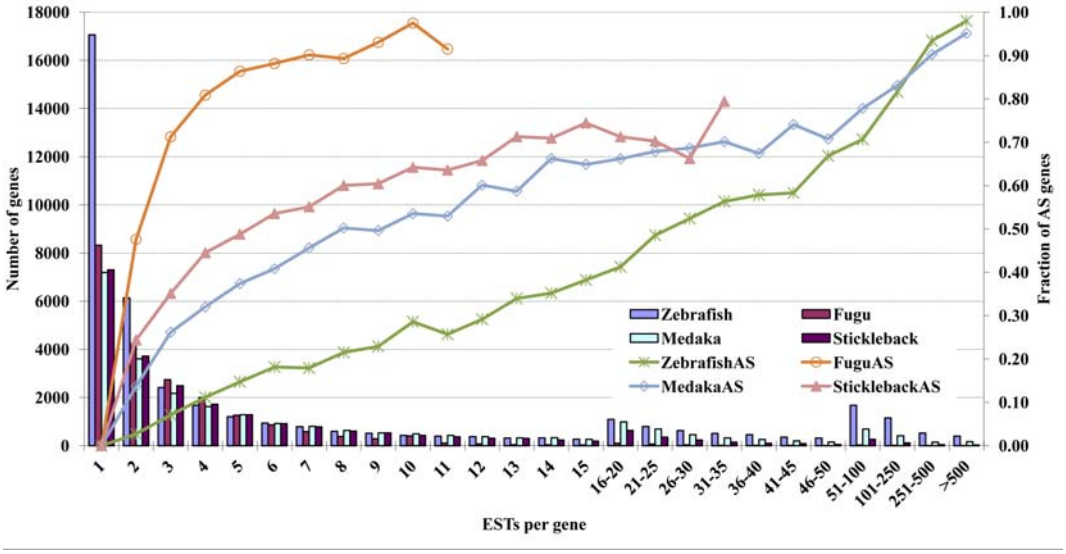


Figure 2.3: Correlation between AS frequency and EST coverage

of AS in the tested species was the alternative donor type, accounting for only 9.5%-12.5% of the AS genes in each species (Figure 2.2).

2.4.4 Gene ontology of teleost AS genes does not indicate category enrichment

To assess whether AS genes from teleost species are evenly distributed across molecular functions, cellular components, and biological process categories, gene ontology assessments were carried out using BLAST2GO based on the UniProt database (Figure 2.5). Results were compared to those obtained using non-AS gene sets from each species (data not shown). Distributions of AS genes among the GO categories did not differ greatly from those of non-AS genes, indicating that AS genes, on the whole, are likely not highly enriched in any of the broader categories of cell structure, molecular function, or biological process across teleost species.

2.4.5 Conservation of AS genes among four teleost species

We also compared the Uniprot top-hit gene identities of the four species to determine the degree of conservation of the AS gene set across teleost genomes. The proportion of

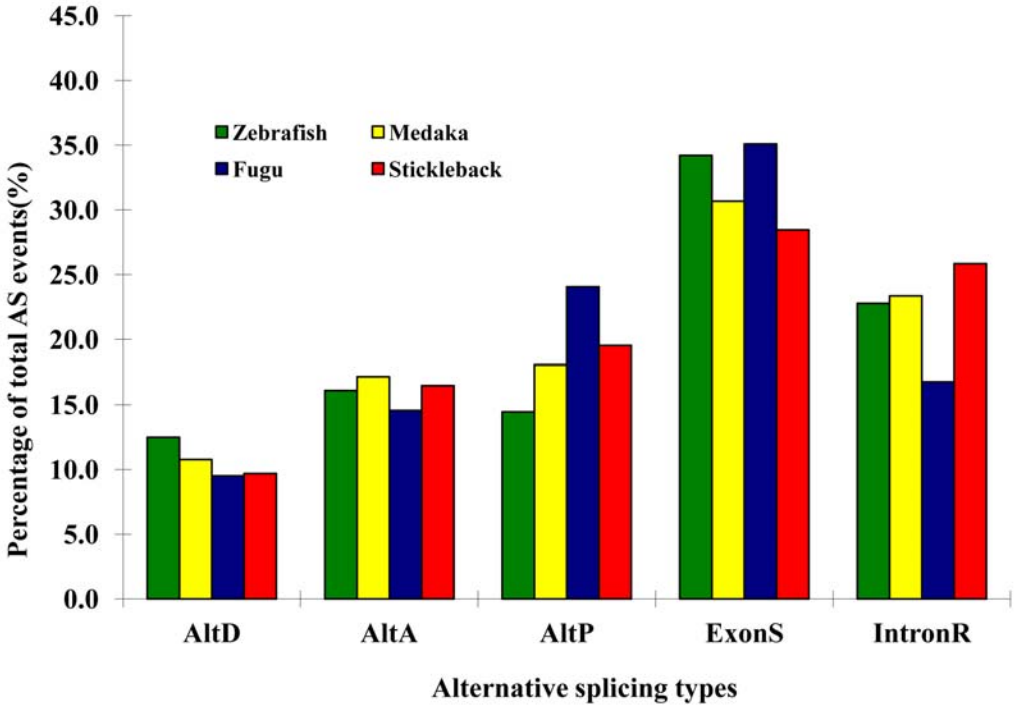


Figure 2.4: The distribution of different AS types in teleost species

putative AS genes with shared identities with one or more of the other three species varied from 47.2% in zebrafish to approximately 57% in both medaka and stickleback (Figure 2.6). Fugu, medaka, and stickleback, which share closer evolutionary relationships with each other than with the more distantly related zebrafish, shared larger numbers of AS gene identities with each other than with zebrafish. A total of 820 AS genes were shared among fugu, medaka, and stickleback. A smaller subset of 182 AS genes were detected in same-species alignments of all four teleost genomes. AS in these genes may play crucial roles that have aided in their maintenance throughout millions of years of evolutionary drift. The shared AS gene set identities and splicing mechanisms are included.

2.4.6 Cross-species alignments for AS detection

Cross-species alignments of transcripts to target genomes have been used to identify novel AS events as well as study conserved patterns of AS across multiple species. Additionally, cross-species alignments can allow the utilization and analysis of ESTs from species

lacking their own genome sequence. All available EST/cDNA sequences from channel catfish (no whole genome sequence), medaka, fugu, and stickleback were aligned against the zebrafish genome. Due to the stringent criteria necessary for accurate mapping of transcript sequences to the exon/intron boundaries of the zebrafish genome, only a small subset of highly conserved genes from each species were mapped (Table 2.3). The smallest number of genes (597) was mapped utilizing the transcript set from channel catfish which, unlike the other three species, lacks full-length Ensembl cDNA data. The largest number of genes was mapped from medaka (2,097). Similarly, ASpipe analysis revealed 39 AS genes based on the transcripts of channel catfish, and 132 AS genes based on medaka transcripts. A large percentage of the cross-species identified AS genes in zebrafish had not previously been identified by analysis with zebrafish ESTs alone. Percentages of novel AS genes identified with the cross-species transcripts varied considerably depending on species, from 31.8% of AS genes supported by medaka sequence evidence to 75.0% of AS genes from fugu (Table 2.3).

Table 2.3: Cross-species EST alignment results with zebrafish genome sequences

Species	EST/cDNA	Mapped to genome	Genes number	AS Genes	Novel
Channel catfish	354,377	5,313 (1.5%)	597	39	20 (51.3%)
Medaka	638,483	58,976 (9.2%)	2,097	132	42 (31.8%)
Fugu	73,945	3,423 (4.6%)	1,099	84	63 (75.0%)
Stickleback	304,239	17,916 (5.9%)	1,117	87	54 (62.1%)

The distribution of AS types identified from cross-species alignments was similar to that found with the same-species approach. On average, exon-skipping was the most abundant AS type identified by both cross-species and same-species approaches, and alternative donor and alternative acceptor AS were the rarest in both cases (Table 2.4).

2.4.7 Teleost Alternative Splicing Database

An online database for teleost alternative splicing was created using the ASviewer format [137]. The database is publically available and can be found at `http://asviewer.acesag.auburn.edu/ASviewer/index.htm`. Users of the database can search for the number, type,

Table 2.4: AS events and types predicted from cross-species EST alignment with zebrafish

Species	Events	AltD	AltA	AltP	ExonS	IntronR
Channel catfish	46	2 (4.3%)	4 (8.7%)	10 (21.7%)	24 (52.3%)	6 (13.0%)
Medaka	177	13 (7.3%)	12 (6.8%)	38 (21.5%)	68 (38.4%)	46 (26.0%)
Fugu	114	4 (3.5%)	12 (10.5%)	32 (28.1%)	34 (29.8%)	32 (28.1%)
Stickleback	115	14 (12.2%)	5 (4.3%)	26 (22.6%)	34 (29.6%)	36 (31.3%)
Avg % cross-species		7.6%	7.0%	23.9%	35.4%	26.2%
Avg% same-species		10.6%	16.1%	19.1%	32.1%	22.2%

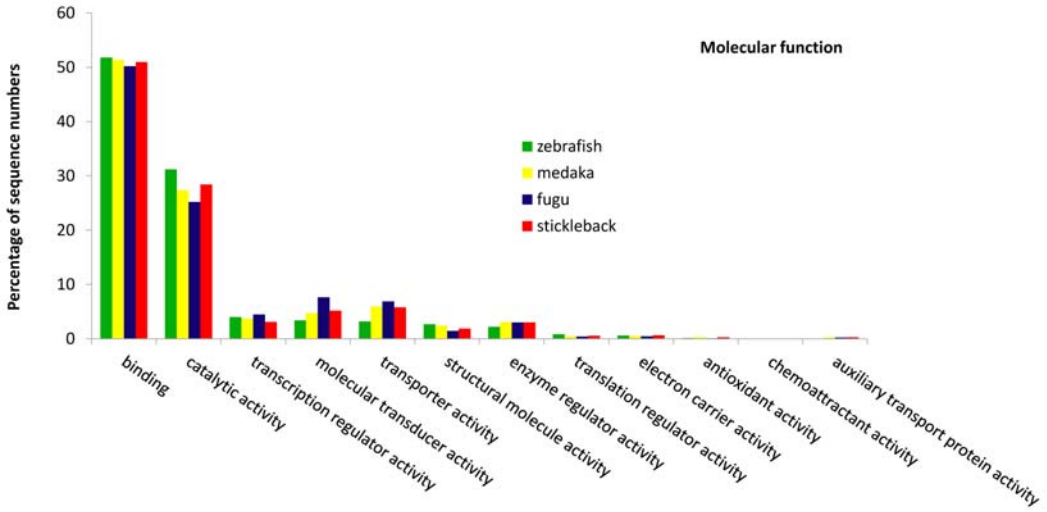
and location of AS genes in channel catfish, fugu, medaka, zebrafish, and stickleback as well as visualizing the AS event (Figure 2.7). AS information will be added for additional species as warranted.

2.5 Discussion

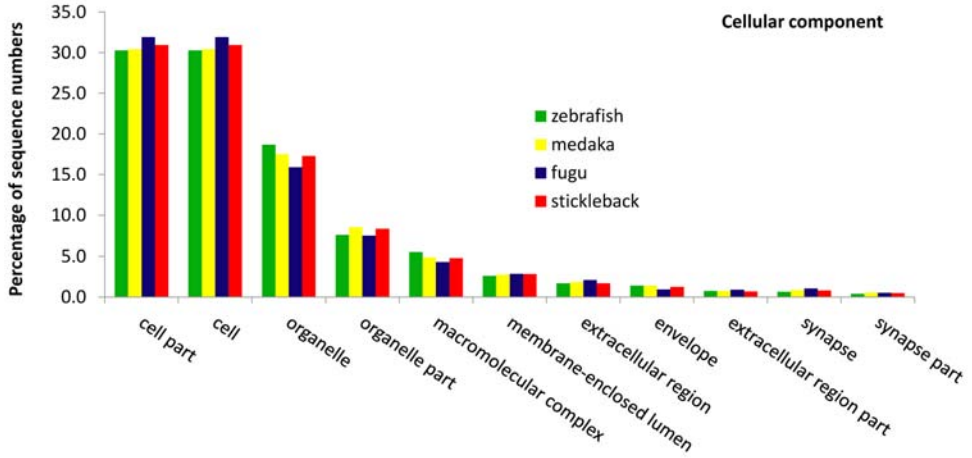
Alternative splicing is recognized as one of the mechanisms by which the coding capacity and diversity of the genome can be amplified. Rates of AS have been shown to vary according to the complexity of eukaryotic species, potentially explaining the tremendous phenotypic diversity among species with similar numbers of coding genes. Little is known, however, regarding mechanisms and rates of AS in teleost fish, a major vertebrate group encompassing over 27,000 species and more than half of vertebrate diversity. In this study, we have conducted the first major genome-wide analysis of AS in teleost fish. Based on same-species and cross-species alignment approaches, we report here that rates of AS appear to vary widely across teleost species, even when accounting for differing EST gene coverage rates. Our analysis indicates that rates of AS are inversely correlated with genome size, potentially shedding light on coding mechanisms underlying differing genome evolutionary strategies within teleost fish. EST coverage rates varied largely among the analyzed teleost species, potentially impacting our ability to detect AS events. Wang et al. (2008) reported that apparent differences in AS frequency between legume species could be attributed to differences in numbers of ESTs available for analysis. Legume species with larger EST resources had higher detected AS frequencies. After normalizing for EST number per gene,

the observed frequency of AS was similar across all tested legume species. We observed a markedly different phenomenon in teleost species. EST coverage ranged from 3.3 ESTs per gene in fugu to 28.9 ESTs per gene in zebrafish. However, despite low transcript coverage, the highest frequency of AS was detected in fugu. Also, normalizing data for ESTs per gene did not diminish differences in AS frequency among the four species. Pronounced differences in AS frequency were observed at low to medium levels of EST gene coverage, with convergence of AS frequencies occurring only above 25 ESTs per gene (Figure 2.3). These differences in AS frequency, therefore, appear to reflect genuine differences in genomic architecture and protein coding strategy among teleost species. Indeed, detected AS frequencies for the four teleost species are inversely correlated with their genome sizes. Fugu (43.2% AS frequency) has a compact genome size of 0.4 Gb. Stickleback (32.4% AS) and medaka (31.2% AS) have intermediate genome sizes of 0.6 Gb and 0.7 Gb, respectively. Zebrafish, with only 17% of genes being detected as AS, has the largest genome by far-1.7 Gb [6] [135] [61] [53]. Fish species with smaller genomes may rely more heavily on AS to generate necessary protein diversity. Conversely, species such as zebrafish with larger, more duplicated genomes may not require the generation of as many AS transcripts to augment protein diversity [133]. Researchers have previously suggested an inverse relationship between rates of gene duplication and AS in animals [145] and, more recently, in plants [150] based on single gene or gene family investigations. However, the present genome-wide analysis of AS in several teleost species offers one of the first wide-ranging examinations of this relationship. Further work is needed to quantify more accurately levels of gene duplication in individual teleost species and to begin to examine the evolutionary interactions between duplication and AS across the teleost radiation. Our study also examined, in part, the utility of cross-species alignments for identification of novel and conserved AS events. Our results indicated that additional putative AS events could be detected applying the transcript sets of related species to a target genome. However, detection power appeared to be limited by transcript length (full-length cDNAs vs. short transcripts) and evolutionary distance between the aligned species.

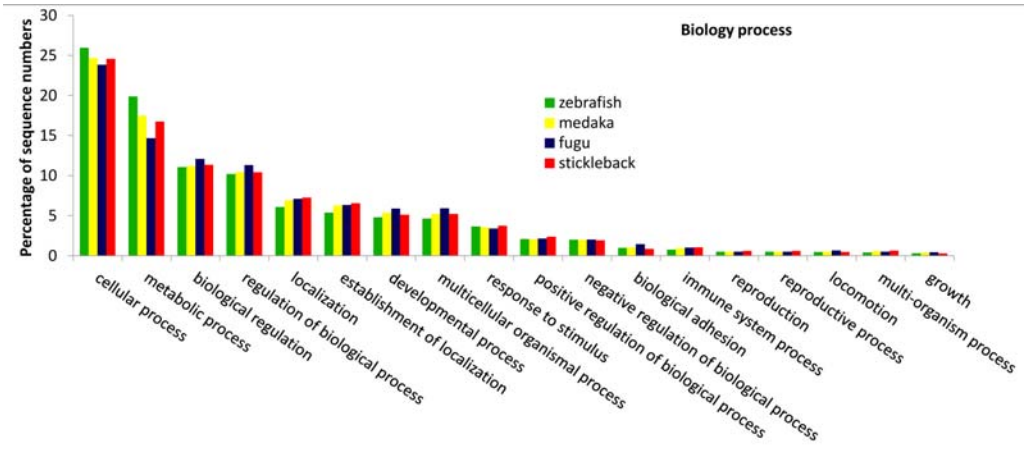
Additional cross-species examination of AS patterns may be informative for transcript-rich, genome-poor species groups such as tilapias, carps and salmonids. Analysis of shared gene identities across the AS gene sets of the four examined teleost species may be informative in identifying biological processes for which AS-generated diversity is essential. In particular, we identified 182 genes that were alternatively spliced in the genomes of each of the four species. This gene set appeared to be moderately enriched for genes regulating developmental processes, anatomical structure formation, and immune system processes (Figure 2.5). Examples of conserved AS gene identities included Fc gamma binding protein, disheveled-associated activator of morphogenesis 1, ligand of numb protein X, ephrin A3, macrophage mannose receptor 1, rhamnose-binding lectin, and complement C3-1. The functional consequences of AS of these genes in teleost fish are unknown in all cases. Additional research is clearly needed to examine the mechanisms supporting maintenance of specific AS genes during teleost species evolution.



(a) molecular function



(b) cellular component



(c) biological process

Figure 2.5: Gene ontology classification results of the AS genes identified from same-species

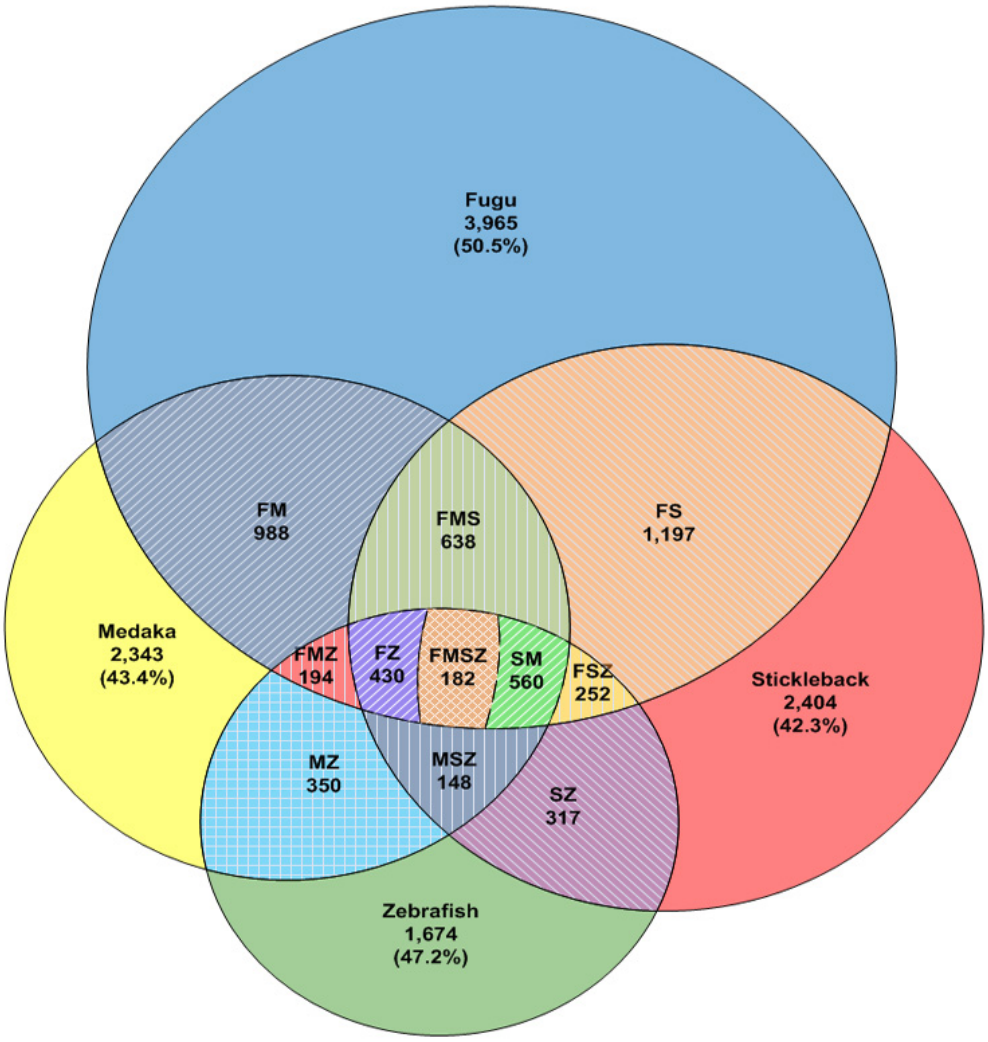


Figure 2.6: Venn diagram of shared AS genes among four teleost species

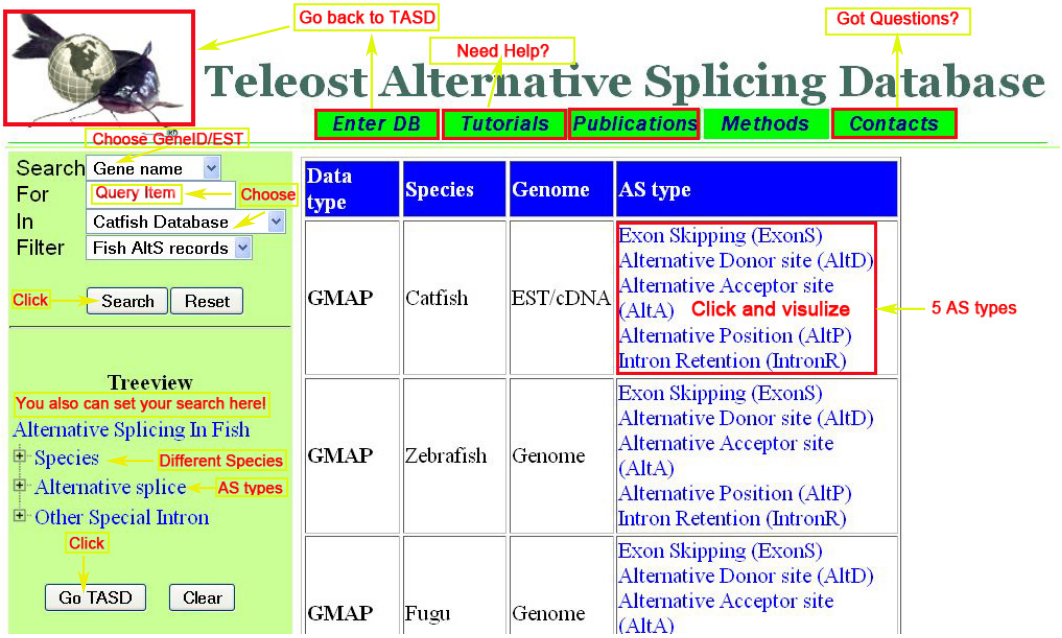


Figure 2.7: Teleost Alternative Splicing Database screenshot

Chapter 3

Profiling of gene duplication patterns of teleost genomes: Evidence for rapid lineage-specific genome expansion mediated by recent tandem duplications

3.1 Abstract

Gene duplication has had a major impact on genome evolution. Localized (or tandem) duplication resulting from unequal crossing over and whole genome duplication are believed to be the two dominant mechanisms contributing to vertebrate genome evolution. While much scrutiny has been directed toward discerning patterns indicative of whole-genome duplication events in teleost species, less attention has been paid to the continuous nature of gene duplications and their impact on the size, gene content, functional diversity, and overall architecture of teleost genomes. Here, using a Markov clustering algorithm directed approach we catalogue and analyze patterns of gene duplication in the four model teleost species with chromosomal coordinates: zebrafish, medaka, stickleback, and *Tetraodon*. Our analyses based on set size, duplication type, synonymous substitution rate (Ks), and gene ontology emphasize shared and lineage-specific patterns of genome evolution via gene duplication. Most strikingly, our analyses highlight the extraordinary duplication and retention rate of recent duplicates in zebrafish and their likely role in the structural and functional expansion of the zebrafish genome. We find that the zebrafish genome is remarkable in its large number of duplicated genes, small duplicate set size, biased Ks distribution toward minimal mutational divergence, and proportion of tandem and intra-chromosomal duplicates when compared with the other teleost model genomes. The observed gene duplication patterns have played significant roles in shaping the architecture of teleost genomes and appear to have contributed to the recent functional diversification and divergence of important physiological processes in zebrafish. We have analyzed gene duplication patterns and duplication

types among the available teleost genomes and found that a large number of genes were tandemly and intrachromosomally duplicated, suggesting their origin of independent and continuous duplication. This is particularly true for the zebrafish genome. Further analysis of the duplicated gene sets indicated that a significant portion of duplicated genes in the zebrafish genome were of recent, lineage-specific duplication events. Most strikingly, a subset of duplicated genes is enriched among the recently duplicated genes involved in immune or sensory response pathways. Such findings demonstrated the significance of continuous gene duplication as well as that of whole genome duplication in the course of genome evolution.

3.2 Introduction

Three main mechanisms are believed to generate gene duplications; unequal crossing over, retrotransposition, and chromosomal (genome) duplication [58] [54]. Of these, localized (or tandem) duplication resulting from unequal crossing over and genome duplication are believed to be the two dominant mechanisms contributing to vertebrate genome evolution [34] [103]. Much energy has been devoted to the examination and modeling of the whole genome duplication events believed to have shaped vertebrate genomes. Over four decades ago, Ohno (1970) suggested that two rounds of large-scale gene duplication had occurred early in vertebrate evolution. Sequencing analysis of Hox gene clusters from a spectrum of vertebrate species provided critical evidence in support of Ohno's hypothesis [50] [110] [88] [30] and indicated, in turn, an additional round of fish-specific genome duplication (FSGD) prior to the divergence of most teleost species [5] [131] [134] [28] [29]. Additional evidence supporting FSGD has been garnered from studies of pufferfish, *Takifugu rubripes* and *Tetraodon nigroviridis*. In these studies, hundreds of genes and gene clusters are present in duplicate in teleost fish but possessing only single copy in other vertebrates, illustrating fish-specific duplication of syntenic regions between humans and fish [56] [90] [112]. Ongoing examination of gene families across vertebrate evolution continues to provide general support for the three rounds of genome duplication (3R) hypothesis [142] [15] [153] [62] [4] [95]

in teleost fish. By contrast, far less energy has been expended in understanding the larger and, arguably, more complicated landscape of gene duplication across model fish genomes and examining how genomes have been shaped and sized by gene duplication forces. Tandem duplication, in particular, is now recognized as a powerful, fast-acting evolutionary mechanism in the generation and expansion of gene families [103], accounting for greater than 10% of human genes [123]. Tandemly-arrayed genes (TAGs) are critical zones of adaptive plasticity, forming the building blocks for sensitive immune, reproductive, and sensory responses [113] [65] [133]. However, their extent and impact on teleost genome architecture has been routinely overlooked in the search for broader genome duplication patterns. While many teleost fish species are in advanced stages of genome sequencing and assembly, only four species currently possess well-annotated genomes with chromosomal-anchored sequence information allowing extensive analysis of gene duplication-zebrafish, *Danio rerio*, medaka, *Oryzias latipes*, green spotted pufferfish, *T. nigroviridis*, and stickleback, *Gasterosteus aculeatus*. These fish, however, represent an interesting cross section of teleost diversity, with genomes differing in size from 342 Mb in pufferfish to 1.5 Gb in zebrafish, and with great variations in effective population sizes and generation intervals ranging from 7 weeks to 2 years. Differences in life history may reasonably be expected to impact patterns of gene duplication and retention. According to the neutral theory of molecular evolution [64] a new paralogous allele, if selectively neutral, has a probability of 1/2N (where N is effective population size) of being fixed in a diploid population, with fixation occurring, on average, over 4N generations. Differences in population size and generation interval among the teleost model species may also impact the extent and effectiveness of positive selection as seen previously in comparisons of duplicated genes between human and mouse [122]. Several recent studies have highlighted exceptional features of the zebrafish genome. These include reports of significantly higher rates of evolution in conserved noncoding elements [69], the largest numbers of tandemly-arrayed duplicates among all surveyed vertebrate species [103],

and the highest average duplication rate of all lineages in the vertebrate tree (9.04 duplications/million years [11]). Our own research has previously revealed a potentially related phenomenon of lower levels of alternative splicing when compared to other teleost species [79] and has explored the extensive nature of tandem duplications within some zebrafish gene families, e.g. cc chemokines [105]. Indeed, the particularities of the zebrafish genome have led many studies to use the more canonical pufferfish and medaka genomes in testing genome and gene duplication models and theories. The zebrafish genome, however, may represent some of the genome architecture of a large number of vertebrate species given its location on a portion of the tree of life with tremendous species diversity. Its family, Cyprinidae, alone accounts for over 2,400 extant species, i.e., 10% of all teleost fish or 5% of all vertebrate animals, and represents the largest family of freshwater fish. Detailed examination and comparative analysis of the nature and impact of duplications in the zebrafish genome may be particularly important, therefore, in understanding the teleost radiation and in informing genomic studies in ostariophysan fishes. To study the nature and extent of duplication among teleost species, here, we used a Markov clustering dynamic programming algorithm to arrange gene duplicates within the four model fish genomes into sets. Further analyses based on set size, duplication type, synonymous substitution rate (Ks), and gene ontology emphasize shared and lineage-specific patterns of genome evolution via duplication. Most strikingly, our analyses confirm the extraordinary duplication and retention rate of recent duplicates in zebrafish and their likely role in the expansion of the zebrafish genome.

3.3 Materials and Methods

3.3.1 Gene set and duplicated gene search

The zebrafish, medaka, stickleback, and *Tetraodon* protein sequences used in this study were obtained from Ensembl (www.ensembl.org; Ensembl Gene 63; Zv9 for zebrafish, HdrR for medaka, BROAD S1 for stickleback, and TETRAODON 8.0 for *Tetraodon*) were used for the gene duplication analysis. Sequences annotated as unknown, random, and mitochondrial

were removed, and only genes with known chromosome location were kept. For all genes with overlapping chromosomal locations, shorter genes were discarded and the longest coding form kept following similar methods used previously [123] [39]. Following filtering, there were 26,842 genes in the zebrafish genome, 18,027 genes in the medaka genome, 19,178 genes in the stickleback genome, and 14,038 genes in the *Tetraodon* genome (Table 3.1). These genes then were used for all-against-all blastp searches [2] using the BLOSUM62 matrix and the SEG filter to mask regions of low compositional complexity [143]. Next, all the gene pairs were sorted by gene name and a filter script was used to remove all the redundant pairs, including self matches and multiple matches. These unique and sorted BLAST results were used as the input of MCscan [130]. MCscan is based on a Markov cluster algorithm which retrieves multiple chromosomal regions using dynamic programming based on the similarity matrix generated from previous BLAST results. The default parameter was used ('mul (0.4343), ceil (200)') to generate the prerequisite .mcl file for MCscan. For the generated duplication sets, we examined the chromosomal locations of the family members for the following duplication type categories.

Table 3.1: Summary of gene duplications in four teleost model species

	Zebrafish	Medaka	Stickleback	*Tetraodon*
Genes	26,842	18,027	19,178	14,038
Duplication sets	3,991	2,584	2,669	2,020
Average duplication set size (gene number)	4.3	5.4	5.4	5.4
Inter-chromosomal duplication sets	3,109 (77.9%)	2,249 (87.0%)	2,262 (84.8%)	1,645 (81.4%)
Intra-chromosomal duplication sets	1,264 (31.7%)	614 (23.8%)	573 (21.5%)	477 (23.6%)
Tandem duplication sets	612 (15.3%)	260 (10.1%)	373 (14.0%)	303 (15.0%)
Mixed duplication sets	994	539	539	405

3.3.2 Duplication categories

The copies of the duplicated gene sets may reside on the same chromosome (intra-chromosomal) or on different chromosomes (inter-chromosomal). Based on the locations and arrangements of the duplicated gene copies, we classified the duplicated genes into the following three categories: 1) Tandem duplication: duplicated gene copies are located next to each other on the same chromosome within a distance of less than 10 kb; 2) Intra-chromosomal duplication (Non-tandem): duplicated gene copies are located on the same chromosome with a distance of greater than 10 kb between all set members; and 3) Inter-chromosomal duplication (Non-tandem): duplicated gene copies are located on different chromosomes. Synonymous substitution (Ks) mutation rates for gene pairs For each pair of homologs, their protein sequences were aligned with CLUSTALW [132] and their protein alignment converted to DNA alignment with PAL2NAL [129]. The Ks values were calculated using the PAML software package [148]. The Nei-Gojobori algorithm [99] was implemented in the PAML package.

3.3.3 Gene ontology calculation for gene pairs

Gene ontology enrichment was calculated using goatools [130]. The resulting data structure is based on a directed acyclic graph (DAG) which can be easily traversed from leaf to root. The over-representation and under-representation of certain GO terms were analyzed based on Fisher's exact test. Also several multiple corrections were implemented including Bonferroni, Sidak, and false discovery rate. The latest version (Jun. 6th, 2011) obo-formatted file was downloaded from Gene Ontology website (`http://geneontology.org`).

3.4 Results

3.4.1 Duplicated gene sets among four model teleost species

Unigene sets gathered from the Ensembl databases of the four teleost fish were used for self-BLAST (all vs. all) followed by Markov clustering dynamic programming utilizing chromosomal coordinates as implemented in the program MCScan [130]. As shown in Table 3.1, a total of 3,991, 2,584, 2,669, and 2,020 duplicated gene sets were identified from zebrafish, medaka, stickleback, and green spotted pufferfish (*Tetraodon*), respectively. Based on chromosomal positions and relationships, the duplication sets were divided into three non-exclusive types: tandem duplication, inter-chromosomal duplication (non-tandem) and intra-chromosomal duplication (non-tandem). Definitions for the duplication types were as follows: 1) tandem duplication: duplicated gene copies located within 10 kb of one another (pairwise); 2) Intra-chromosomal duplication (Non-tandem): duplicated gene copies located on the same chromosome with a distance of greater than 10 kb between all members; and 3) Inter-chromosomal duplication (Non-tandem): duplicated gene copies located on different chromosomes. A portion of the duplicated sets combined several duplication types (e.g., duplicate set members present in both tandem and inter-chromosomal arrangements; Table 1). Inter-chromosomal duplications were the most prevalent among the three types across all four teleost species, accounting for around 80% of duplication sets and indicating the importance of genome-level duplication events in shaping teleost genome architecture. Intra-chromosomal and tandem duplication were the second and third most prevalent types, respectively. Zebrafish had the highest percentage of sets within these latter two categories, 47%, compared with 33.9%, 35.5%, and 38.6% in medaka, stickleback, and *Tetraodon*, respectively. In addition, zebrafish differed noticeably from medaka, stickleback, and *Tetraodon* in average duplication set size, with 4.3 genes per duplication set compared to 5.4 genes per set in the three other species.

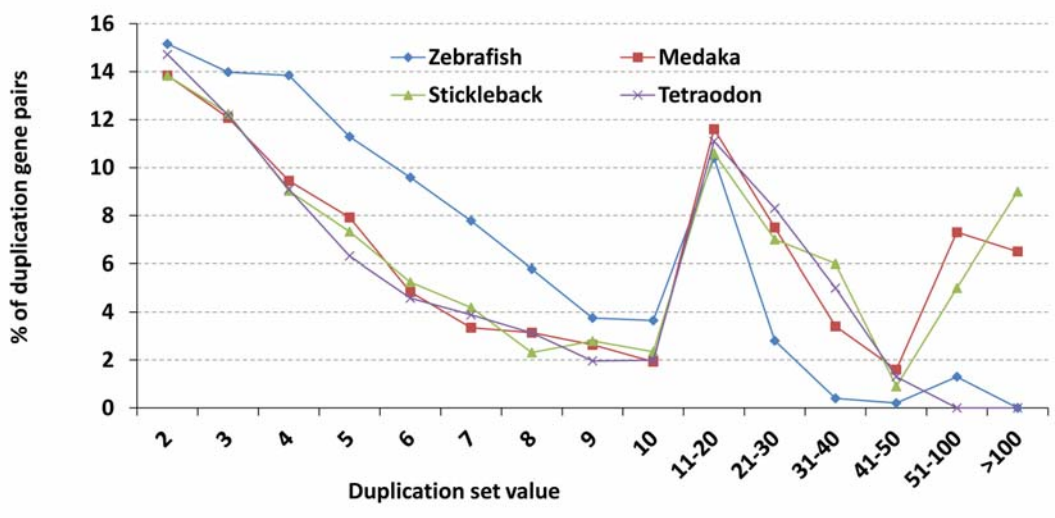


Figure 3.1: The distribution of duplicated genes from the four model teleost species across varying duplication set sizes

3.4.2 Duplication set size prevalence differs between zebrafish and other teleost species

To better understand the distribution of duplicated genes within the four model teleost species, we examined the number of genes on a percentage basis found within duplication sets of varying size. While the relationship between duplication set size and percentage of duplicated genes was similar among the four species (Figure 3.1; Table 3.2), zebrafish again was the outlier, showing a pattern of more numerous small-scale duplications (set sizes 2-10). This pattern was consistent with our observation of smaller average set size in zebrafish, as was the larger number of duplications found in set sizes greater than 20 in medaka, stickleback, and *Tetraodon*.

3.4.3 Lineage-specific patterns of duplication events among four teleost species

We next asked whether the observed prevalence of small duplication sets in zebrafish reflected a faster evolutionary rate in the species as manifested in its duplicated genes. To answer the question, we first examined the mutational distance between the duplicated genes (pairwise) of each species using *Ks*, a measure of the number of substitutions per synonymous

Table 3.2: Duplication set size distribution in four teleost species

Set size	Zebrafish	Medaka	Stickleback	*Tetraodon*
2	1309 (15.2%)	970 (13.8%)	1005 (13.8%)	810 (14.7%)
3	805 (14.0%)	564 (12.1%)	593 (12.2%)	447 (12.2%)
4	598 (13.8%)	331 (9.4%)	328 (9.0%)	250 (9.1%)
5	390 (11.3%)	222 (7.9%)	213 (7.3%)	139 (6.3%)
6	276 (9.6%)	113 (4.8%)	127 (5.2%)	84 (4.6%)
7	192 (7.8%)	67 (3.3%)	87 (4.2%)	61 (3.9%)
8	125 (5.8%)	55 (3.1%)	42 (2.3%)	43 (3.1%)
9	72 (3.8%)	41 (2.6%)	45 (2.8%)	24 (2.0%)
10	63 (3.6%)	27 (1.9%)	34 (2.3%)	22 (2.0%)
11-20	134 (10.4%)	113 (11.6%)	106 (10.6%)	84 (11.1%)
21-30	20 (2.8%)	42 (7.5%)	41 (7.0%)	37 (8.3%)
31-40	2 (0.4%)	14 (3.4%)	25 (6.0%)	16 (5.0%)
41-50	1 (0.2%)	5 (1.6%)	3 (0.9%)	3 (1.3%)
51-100	4 (1.3%)	14 (7.3%)	11 (5.0%)	0 (0.0%)
$\geq$100	0 (0.0%)	6 (6.5%)	9 (9.0%)	0 (0.0%)

site. We again noted a strikingly different *Ks* distribution in zebrafish when compared with the three other model species (Figure 3.2). Over 24.4% of duplicated genes in zebrafish had Ks values of $\leq$1.0% compared to 1.3%, 0.97%, and 0.05% of duplicated genes in medaka, stickleback and *Tetraodon*, respectively.

To determine whether the abundance of small duplicate sets in zebrafish may be explained by recent evolution (low *Ks*) of these genes, we calculated average *Ks* values for each duplicated set size in the size ranges where zebrafish has a greater percentage of duplicated genes (set size 1-10; Figure 3.3). Indeed, *Ks* values in these sets are markedly lower in zebrafish than in medaka, stickleback, and *Tetraodon*. Interestingly, while a clear positive correlation existed between duplication set size and *Ks* value in stickleback and Tetraodon, this pattern was obscured in medaka and not apparent in zebrafish. In fact, no gain in *Ks* value was observed from set size 3 through set size 10 in zebrafish.

The relationship between *Ks* and set size was even more evident when the duplicated set sizes were analyzed separately and individual pairwise Ks values were plotted (Figure 3.4). As seen previously, zebrafish has an abundance of low *Ks* (*Ks* <1) duplicate pairs at all the

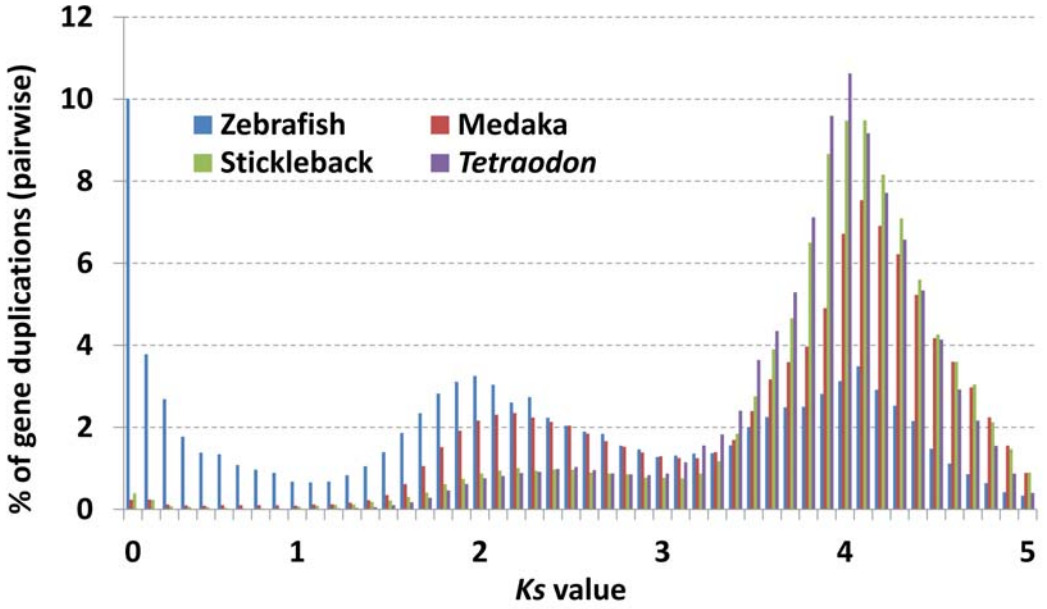


Figure 3.2: The distribution of duplicated genes (pairwise comparisons) from the four model teleost species across varying *Ks* values

studied set sizes when compared with the other three species. However, several other interesting patterns were evident in this analysis. Zebrafish and medaka maintain two roughly proportional peaks of *Ks* values (approximate mean values of Ks=2 and Ks=4), indicating two broad age (divergence level) categories of duplicated genes in these species, irrespective of duplicate set number. In contrast, a single major peak (mean Ks=4) was observed in stickleback and *Tetraodon*, with a much smaller *Ks* peak (Ks=2) appearing to generally diminish with increasing set size. The *Ks* distributions of stickleback and *Tetraodon* are particularly striking in their similarity to one another and suggest a dramatically diminished role for recent duplications in shaping these species' genomes when compared with zebrafish and medaka.

3.4.4 Tandem duplications are predominant among small, recent gene duplications in zebrafish

We next asked whether the large numbers of small, recent duplications observed in zebrafish were evenly distributed across duplication types or whether they were biased toward

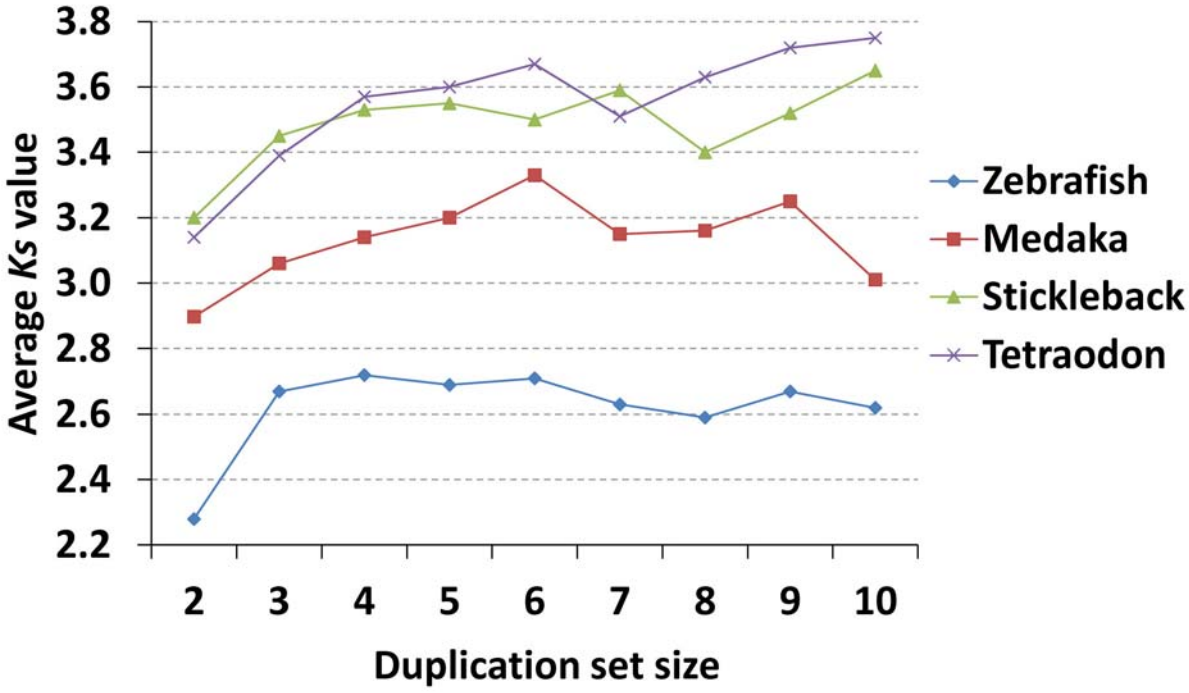


Figure 3.3: The relationship between duplication set size and average *Ks* value of duplicated genes from the four teleost species

a particular type. As seen in Figure 3.5, tandem gene duplicates had the lowest *Ks* values in each species irrespective of duplication set size. Tandem duplicates from zebrafish had the lowest *Ks* values observed in any species with little perceptible increase in mutational distance across the analyzed duplicated set sizes. Intra-chromosomal duplicates in zebrafish and medaka had intermediate *Ks* values between tandem and inter-chromosomal duplication with an upward trend correlated with increasing duplication set size. By contrast, *Ks* values for intra-chromosomal duplicates in stickleback and *Tetraodon* were virtually indistinguishable from those of inter-chromosomal duplicates in duplication sets of size ≥ 3. These patterns again point to the static nature of these genomes, with diminished retention and/or minimal levels of recent intra-chromosomal or tandem duplication activity to shape their genome architecture.

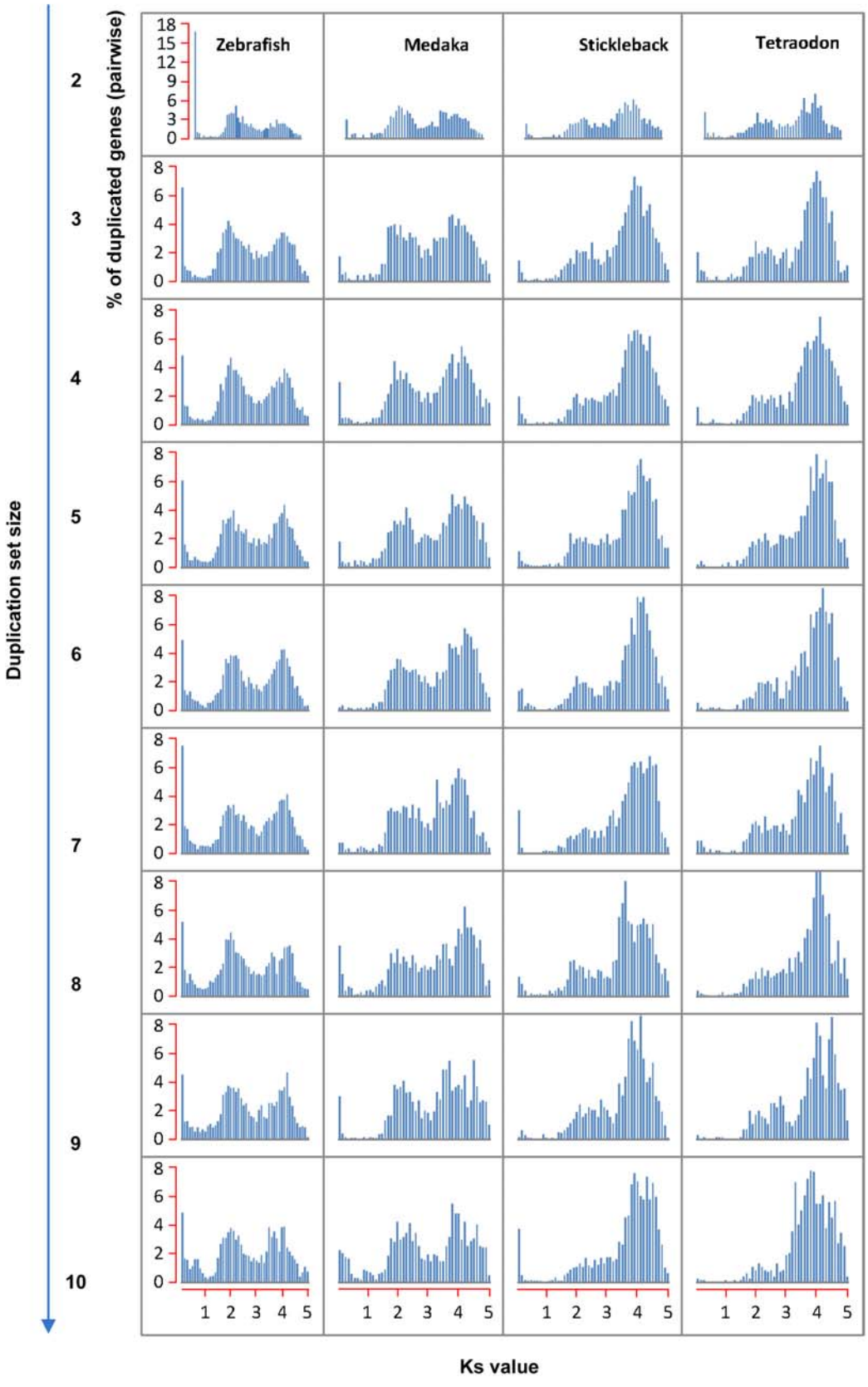


Figure 3.4: The distribution of duplicated genes (pairwise comparisons) across increasing *Ks* values for each duplication set size (2 to 10) 39

3.4.5 Functional bias of recent (low *Ks*) duplicates in zebrafish

In order to determine whether the expansion of recent, retained duplicates in zebrafish has contributed to the diversification of genes mediating particular physiological functions in the species, we carried out gene ontology analysis on the duplicated gene sets with *Ks* values ≤ 1.0. This *Ks* range comprises the duplicated set with the most striking expansion when compared with the three other teleost models (Figure 3.3 and Figure 3.5). Three GO terms were enriched among these duplicates when compared to the larger set of duplicated zebrafish genes (Table 3.3)-MHC protein complex, olfactory receptor activity, and antigen processing and presentation. Similar enrichment was not detected in the other three species, precluded in part by their small set sizes in this *Ks* range. The enriched categories, critical for immune and sensory capabilities, strongly suggest a functional bias in mechanisms of duplication and retention in zebrafish and further point to the importance of lineage-specific patterns of duplication in genome evolution and species diversification.

3.5 Discussion

Gene duplication has been described as an opportunity to explore forbidden evolutionary space [54], the idea that duplicated genes operating under temporary conditions of relaxed selection provide the raw material for evolution of new gene functions. While whole-genome duplication events are critical in shaping broader genome architecture, gene duplication, particularly tandem events, represent more recent, and potentially, adaptive signatures of evolution [39] which are expected to differ among vertebrate lineages [123] [78]. Indeed, Robinson-Rechavi and Laudet (2001), using zebrafish as their model, and others have shown evidence that evolutionary rates of duplicated genes in teleost fish far outstrip those of the mouse lineage. These differences, aside from adaptive consequences, can have profound effects on the degree of shared ancestry and synteny among vertebrate genomes. For example, only 50% of duplicated genes in zebrafish, and 70% in *Tetraodon*, have their origin in 1R/2R WGD events, compared to over 80% in mammalian, avian, and amphibian lineages. The

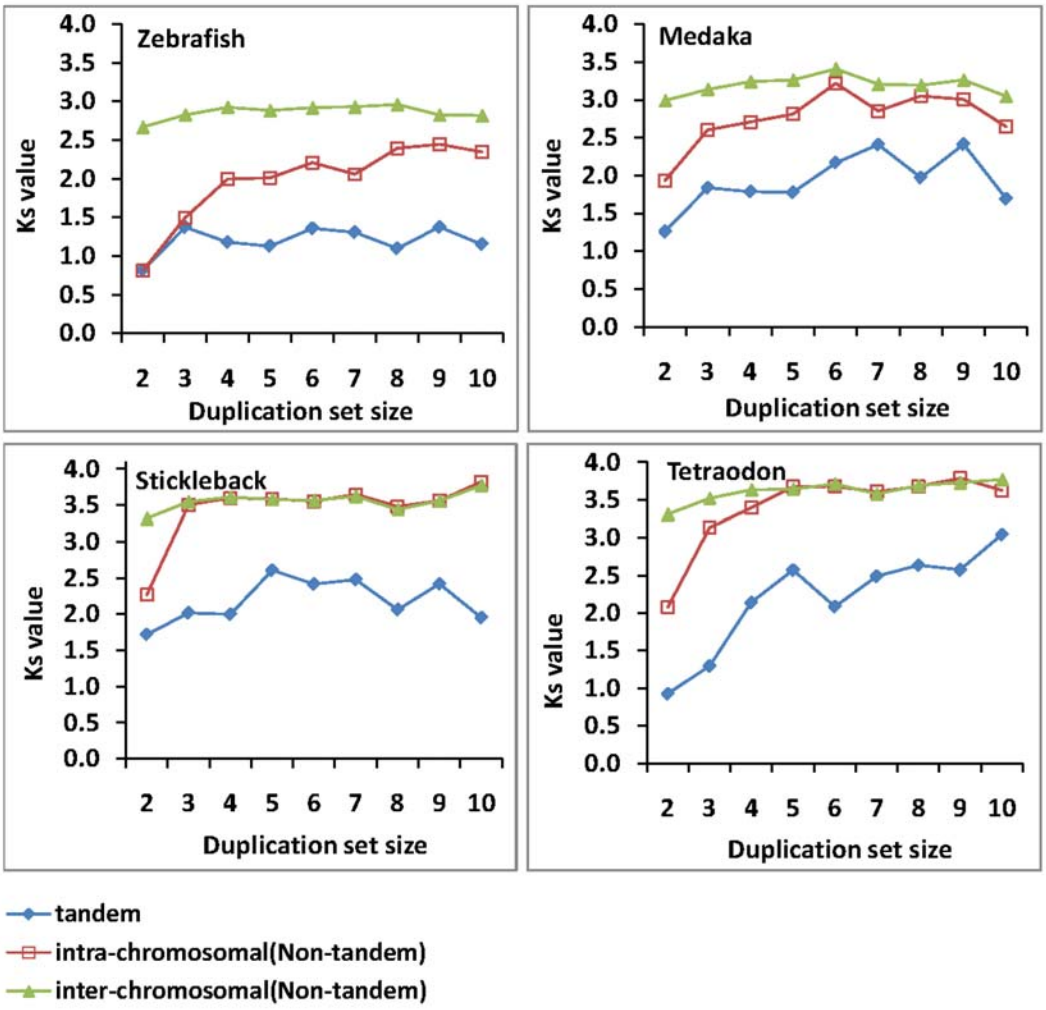


Figure 3.5: Average *Ks* values for varying duplication set sizes and among the three different duplication types in the four model teleost species

remaining fraction comes from FSGD and species-specific events [11]. Clearly, patterns of teleost gene duplication deserve closer scrutiny to better understand how this process continues to shape genome evolution. Therefore, here we examined the nature and extent of gene duplication in four model teleosts, zebrafish, medaka, stickleback and *Tetraodon*.

Table 3.3: Gene ontology enrichment in zebrafish duplicate pairs with low *Ks* values ($Ks \leq 1.0$)

GO term	GO category	GO name
GO:0042611	Cellular component	MHC protein complex
GO:0004984	Molecular function	Olfactory receptor activity
GO:0019882	Biological process	Antigen processing and presentation

Our approach divided duplicated genes into sets based on duplication type and captured larger gene families as well as smaller, recent duplications. From the onset of our analysis, zebrafish stood out from the other three model species by most measures, with a larger percentage of sets involved in tandem and intra-chromosomal arrangements and numerous small duplication sets (Table 3.1, Figure 3.1). Our analysis of the mutational distance between duplicate pairs (*Ks*) across the teleost species (Figure 3.2), however, produced the most striking illustration of different patterns of duplication and retention. Over 24% of duplicate pairs in zebrafish had *Ks* values of ≤ 1.0 compared to around 1% or less in the other three species. These results are supported by previous studies which noted high evolutionary rates and duplicate retention rates in zebrafish [69] [11]. The abundance of low *Ks* duplicate pairs in zebrafish may stem from a greater number of birth events, fewer gene loss events among young duplicates, and/or maybe the homogenization effect of gene conversion. Gene conversion can obscure the antiquity of duplicated genes as recombination between paralogous sequences minimizes divergence [54] [102], and has been observed previously among zebrafish protocadherin tandem duplicates [101]. Also a small number of paralogs may be missed or misassembled in one or more of the teloest genomes. Our approach focused on surveying the broader architecture of duplication in the teleost genomes rather than relying on cross-species phylogenetic analysis for identification of orthologous relationships. Our analyses

are limited, therefore, in distinguishing between rapid lineage specific gains in zebrafish and excessive gene loss in other teleosts for particular duplicate sets. The bias in the low *Ks* duplicate pairs in zebrafish toward tandem duplication (Figure 3.5) provides support for these being recent duplication events. Close to 65% of these zebrafish duplicate pairs with $Ks < 1.0$ are found in tandem arrangements compared with 15% of total duplicated sets (data not shown). In addition, gene ontology analysis revealed a bias in these duplicates toward physiological functions previously associated with rapid evolution and adaptation [122] [121] [26]. Indeed, the enriched categories (olfactory receptors, MHC) are well known for their rapid diversification through duplication, recombination, and gene conversion [121] [13] [3]. Taken together, our results suggest strikingly rapid evolution and high retention of recent duplicates in zebrafish in a manner likely to result in specialization of immune and sensory mechanisms. The differences observed in *Ks* distributions among the four teleost species (Figure 3.3 and Figure 3.5) raised several intriguing questions for further research: What is the effect of life history on the genome architecture of fish, and is there a link between genome size and duplication rate/retention rate in fish? Shiu et al. (2006) examined similar lineage-specific patterns when comparing human and mouse duplicates, suggesting that the larger population size and shorter generation interval in murine species could account for more effective natural selection and retention of duplicated genes. In the four investigated teleost genomes, zebrafish and medaka share similar life history patterns, generation intervals of 7-9 weeks and large effective population sizes, and similar *Ks* distributions (excluding $Ks < 1.0$). By contrast, *Tetraodon* and stickleback, generation intervals of 1-2 yrs and smaller effective population sizes, had a notable absence of young (low *Ks*) duplicates and shared remarkably similar *Ks* distributions (Figure 3.5) across their duplicated genes. These patterns of duplication rate and retention have been explored in the light of population size using genome sequence information in invertebrates [72] and previously, on a more theoretical basis [38] [81]. Previous observations of correlations between spontaneous duplication/deletion rates and effective population size and increasing retention of linked (tandem) duplicates at

intermediate population sizes appear to support such a connection between life history and duplication profiles as suggested by our data. Another pattern deserving further attention as additional teleost genomes become available is a potential association between duplication timing/retention rates and genome size. Based on the limited data available from the four model genomes here, patterns of duplication rate (especially as reflected by those pairs with $Ks \leq 1.0$) reflect genome size with zebrafish with the largest genome at 1.5 Gb, followed by medaka (700 Mb), stickleback (446 Mb) and *Tetraodon* (342 Mb). The drastically differing patterns of duplicate formation and retention as detected here and by Blomme (2006) may be reflected in evolution of non-coding elements as well [69] and, together, could contribute to significantly higher genic content and associated genome size, as observed in zebrafish [51]. Previously, we highlighted the low levels of alternative splicing detected from zebrafish (17% of mapped genes) compared with the other model teleost species [79]. By contrast, the compact genome of *Tetraodon* showed alternative splicing in 43% of mapped genes. In that study, an inverse correlation between genome size and alternative splicing was observed. Researchers have previously suggested an inverse relationship between rates of gene duplication and alternative splicing in animals [145] and, more recently, in plants [150] based on single gene or gene family investigations. Our previous analysis of alternative splicing combined with our present examination of gene duplication in the same teleost species appears to support this connection on a genome scale. Further study is warranted to investigate whether the recent duplicates of zebrafish can provide the functional repertoire generated through alternative splicing in other, smaller teleost genomes. Our findings indicate that varying rates of gene duplication and retention can have a dramatic impact on the ancestry and architecture of teleost genomes and contribute to functional diversification and divergence of important physiological processes. These patterns may be reflective of differences in life history across the teleost radiation and may ultimately influence genic content and genome size. Further analyses of the genomes of additional, key teleosts (i.e. catfish, carp) in the near future will allow us to test these theoretical relationships and analyze the particularities

of the zebrafish genome in the context of more recently diverged species. In Brown's paper, the Copy number variation elements (CNVE) appeared to be consistent with extensive population substructuring (i.e., local adaption) among zebrafish population, with 4,199 (69%) of the identified CNVEs unique to one strain and only 457 (7.5%) CNVEs are common to all four groups [18]. Given this large amount of genome variation, analysis of genomes from additional zebrafish populations may reveal differing gene numbers within a given duplication set. This would be of great interest in helping to establish the rate of gene birth in zebrafish. However, only the reference genome sequences were available for the present analysis. We still feel the work is an important foundation for guiding analysis of duplication numbers in individuals and strains as sequencing costs rapidly allow such analyses.

Chapter 4

MPI-Velvet Next Generation Sequence Assembler

4.1 Abstract

Next generation sequence is undoubtedly the most important technology in biology, especially for the non-model species. Next generation sequence technological advances have dramatically improved sequencing throughput and quality. Like Illuminas Genome Analyzer produces a significant larger volume of sequence data than traditional sanger sequencing. Compared to just a few years ago, it is now much easier and cheaper to sequence entire genomes. Because of the rapid improvements in cost and quality of sequencing data, de novo sequencing and assembly is possible not only in large sequencing centers, but also in small labs. However, when you get the genome sequence information from sequencing company, there are several questions will be prompted immediately. The first question is which alignment algorithm or which assembler you want to choose for your sequencing project? The second question is how to optimize your alignment algorithm based on both high speed and high accuracy. To answer these two questions, this project addressed the Message Passing Interface (MPI) version assembler software, MPI-Velvet. It can process high coverage data sets and quickly reconstruct the underlying sequences.

4.2 Introduction

Very short reads are not well suited to this traditional approach. Because of their length, they must be produced in large quantities and at greater coverage depths than traditional Sanger sequencing projects. The sheer number of reads makes the overlap graph, with one node per read, extremely large and lengthy to compute. With long reads, repeats in the data

are disambiguated by careful metrics over long overlaps that distinguish repeat matches from real overlaps, using, for example, high-quality base disagreements. With short reads, and correspondingly short overlaps to judge from, many reads in repeats will have only a single or no base differences. This leads to many more ambiguous connections in the assembly.

The fundamental data structure in the de Bruijn graph is based on kmers, not reads, thus high redundancy is naturally handled by the graph without affecting the number of nodes. In addition, each repeat is present only once in the graph with explicit links to the different start and end points. Depending on available information, it can be either resolvable or not, but it is readily identifiable. Mis-assembly errors are therefore more easily avoided than with overlap graphs. Finally, searches for overlaps are simplified, as overlapping reads are mapped onto the same arcs and can easily be followed simultaneously.

Despite the attractiveness of the de Bruijn graph data structure for short read assemblies, it has not been used extensively in current production-based assembly methods. Chaisson et al. [22] and Sundquist et al. [127] suggested ways of using these graphs specifically for short read assembly (100−200 bp), but not for very short reads (25−50 bp). More recently, programs such as SSAKE [140], SHARCGS [32], and VCAKE [57] implicitly use this framework, but at a local level. With the advent of highly cost effective very short reads, de Bruijn graph-based methods will grow in utility. However, it is necessary to develop efficient and robust methods to manage experimental errors and repeats.

Velvet, that manipulates these de Bruijn graphs efficiently to both eliminate errors and resolve repeats. These two tasks are done separately: first, the error correction algorithm merges sequences that belong together, then the repeat solver separates paths sharing local overlaps. Velvet is capable of assembling bacterial genomes,with N50 contig lengths of up to 50 kb, and simulations on 5 Mb regions of large mammalian genomes, with contigs of $\sim$3 kb [152].

4.3 Message Passing Interface (MPI)

MPI is a language-independent communications protocol used to program parallel computers. Both point-to-point and collective communication are supported. MPI is a message-passing application programmer interface, together with protocol and semantic specifications for how its features must behave in any implementation [80]. MPI's goals are high performance, scalability, and portability. MPI remains the dominant model used in high-performance computing today [128].

MPI is not sanctioned by any major standards body; nevertheless, it has become a *de facto* standard for communication among processes that model a parallel program running on a distributed memory system. Actual distributed memory supercomputers such as computer clusters often run such programs. The principal MPI1 model has no shared memory concept, and MPI2 has only a limited distributed shared memory concept. Nonetheless, MPI programs are regularly run on shared memory computers. Designing programs around the MPI model (contrary to explicit shared memory models) has advantages over NUMA architectures since MPI encourages memory locality.

Most MPI implementations consist of a specific set of routines (i.e., an API) directly callable from Fortran, C and C++ and from any language capable of interfacing with such libraries (such as C, Java or Python). The advantages of MPI over older message passing libraries are portability (because MPI has been implemented for almost every distributed memory architecture) and speed (because each implementation is in principle optimized for the hardware upon which it runs).

MPI uses Language Independent Specifications (LIS) for calls and language bindings. The first MPI standard specified ANSI C and Fortran-77 bindings together with the LIS. The draft was presented at Supercomputing 1994 (November 1994) [44] and finalized soon thereafter. About 128 functions constitute the MPI?1.3 standard which was released as the final end of the MPI1 series in 2008.

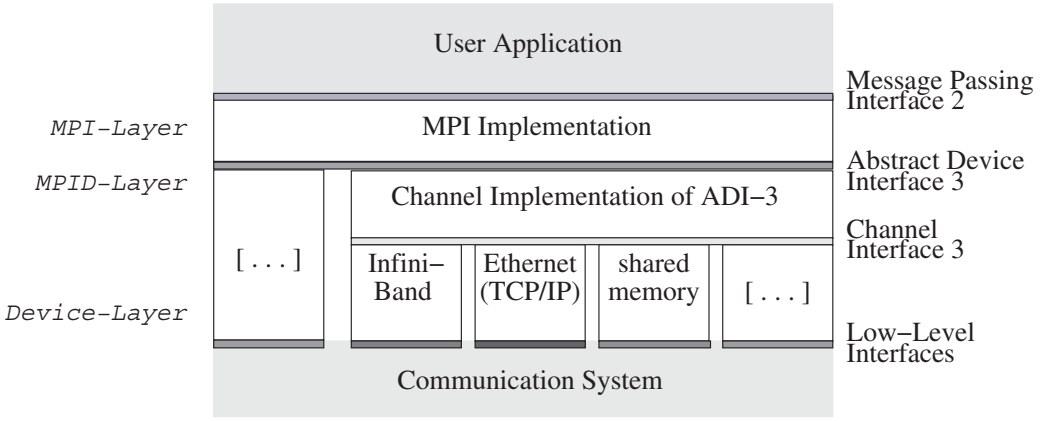


Figure 4.1: Hierarchical Structure of MPICH2

At present, the standard has several popular versions: version 1.3 (shortly called MPI1), which emphasizes message passing and has a static runtime environment, and MPI2.2 (MPI2), which includes new features such as parallel I/O, dynamic process management and remote memory operations [45]. MPI2's LIS specifies over 500 functions and provides language bindings for ANSI C, ANSI Fortran (Fortran90), and ANSI C++. Object interoperability was also added to allow for easier mixed-language message passing programming. A side-effect of MPI2 standardization (completed in 1996) was clarification of the MPI1 standard, creating the MPI1.2.

4.4 MPICH2

MPICH - one of the most popular MPI implementations - were developed at the Argonne National Laboratory [80]. The early MPICH version supports the MPI-1 standard. MPICH2 - a successor of MPICH - not only provides support for the MPI-1 standard, but also facilitates the new MPI-2 standard, which specifies functionalities like one-sided communication, dynamic process management, and MPI I/O [40]. Compared with the implementation of MPICH, MPICH2 was completely redesigned and developed to achieve high performance, maximum flexibility, and good portability.

Fig. 4.1 shows the hierarchical structure of the MPICH2 implementation, where there are four distinct layers of interfaces to make the MPICH2 design portable and flexible. The four layers, from top to bottom, are the message passing interface 2 (MPI-2), the abstract device interface (ADI3), the channel interface (CH3), and the low-level interface. ADI3 - the third generation of the abstract device interface - in the hierarchical structure (see Fig. 4.1) allows MPICH2 to be easily ported from one platform to another. Since it is non-trivial to implement ADI3 as a full-featured abstract device interface with many functions, the CH3 layer simply implements a dozen functions in ADI3 [74].

As shown in Fig. 4.1, the TCP socket Channel, the shared memory access (SHMEM) channel, and the remote direct memory access (RDMA) channel are all implemented in the layer of CH3 to facilitate the ease of porting MPICH2 on various platforms. Note that each one of the aforementioned channels implements the CH3 interface for a corresponding communication architecture like TCP sockets, SHMEM, and RDMA. Unlike an ADI3 device, a channel is easy to implement since one only has to implement a dozen functions relevant for with the channel interface.

To address the issues of message snooping in the message passing environments on clusters, I seek to implement a standard MPI mechanism with confidentiality services to counter snooping threats in MPI programs running on a cluster connected an unsecured network. More specifically, I aim to implement cryptographic algorithms in the TCP socket channel in the CH3 layer of MPICH2.

4.5 Next generation sequence technology

Today's commercial DNA sequencing platforms include the Genome Sequencer from Roche 454 Life Sciences, the Solexa Genome Analyzer from Illumina, the SOLiD System from Applied Biosystems, the Heliscope from Helicos, and the commercialized Polonator.

These platforms have been well reviewed [84], [85], [126], [106]. A distinguishing characteristic of these platforms is that they do not rely on Sanger chemistry [118] as did first-generation machines including the Applied Biosystems Prism 3730 and the Molecular Dynamics MegaBACE. The second-generation machines are characterized by highly parallel operation, higher yield, simpler operation, much lower cost per read, and (unfortunately) shorter reads. Today's machines are commonly referred to as short read sequencers or next-generation sequencers (NGS) though their successors may be on the horizon. Pacific Biosciences machines [36] might produce reads longer than first-generation machines. First generation reads were commonly 500 bp to 1000 bp long. Today's NGS reads are in the 400 bp range (from 454 machines), the 100 bp range (from the Solexa and SOLID machines), or less. Shorter reads deliver less information per read, confounding the computational problem of assembling chromosome-size sequences. Assembly of shorter reads requires higher coverage, in part to satisfy minimum detectable overlap criteria. High coverage increases complexity and intensifies computational issues related to large data sets.

4.6 Assembly software overview

4.6.1 What is Assembly?

An assembly is a hierarchical data structure that maps the sequence data to a putative reconstruction of the target. It groups reads into contigs and contigs into scaffolds. Contigs provide a multiple sequence alignment of reads plus the consensus sequence. The scaffolds, sometimes called supercontigs or metacontigs, define the contig order and orientation and the sizes of the gaps between contigs. Scaffold topology may be a simple path or a network. Most assemblers output, in addition, a set of unassembled or partially assembled reads. The most widely accepted data file format for an assembly is FASTA, wherein contig consensus sequence can be represented by strings of the characters A, C, G, T, plus possibly other characters with special meaning. Dashes, for instance, can represent extra bases omitted from the consensus but present in a minority of the underlying reads. Scaffold consensus

sequence may have N's in the gaps between contigs. The number of consecutive N's may indicate the gap length estimate based on spanning paired ends.

Assemblies are measured by the size and accuracy of their contigs and scaffolds. Assembly size is usually given by statistics including maximum length, average length, combined total length, and N50. The contig N50 is the length of the smallest contig in the set that contains the fewest (largest) contigs whose combined length represents at least 50 % of the assembly. The N50 statistics for different assemblies are not comparable unless each is calculated using the same combined length value. Assembly accuracy is difficult to measure. Some inherent measure of accuracy is provided by the degrees of mate-constraint satisfaction and violation [109]. Alignment to reference sequences is useful whenever trusted references exist.

4.6.2 Current assembly software

Along with the next-generation sequencing (NGS) technologies, it gave rise to a new generation of assembly algorithm and software. There are some new class of assembler which are used to deal with the short, non-perfect reads generated form Illumina/Solexa, and ABI SOLiD platforms. Such as SSAKE [140], Velvet [152], VCAKE [57], ALLPATHS [20], Euler-SR [23], SHARCGS [32], Celera Assembler [31], CABOG [82], SOAP [71], Edena [48], ABySS [124], QSRA [19] and MIRA [27]. All the short read assembler is to combine small DNA sequence fragments into longer contigs. The common feather of these assembler is to build the DNA contigs based on fragments overlap. However, the algorithms of these assemblers are different. So, the runtime and output differences are significant different even using the same parameters and same input sequence information. How to optimize your alignment algorithm for both high speed and low error rate is very important for NGS assembly ananlysis.

4.6.3 The challenge of assembly

DNA sequencing technologies share the fundamental limitation that read lengths are much shorter than even the smallest genomes. WGS overcomes this limitation by over-sampling the target genome with short reads from random positions. Assembly software reconstructs the target sequence. Assembly software is challenged by repeat sequences in the target. Genomic regions that share perfect repeats can be indistinguishable, especially if the repeats are longer than the reads. For repeats that are inexact, high-stringency alignment can separate the repeat copies. Careful repeat separation involves correlating reads by patterns in the different base calls they may have. Repeat separation is assisted by high coverage but confounded by high sequencing error. For repeats whose fidelity exceeds that of the reads, repeat resolution depends on "spanners", that is, single reads that span a repeat instance with sufficient unique sequence on either side of the repeat. Repeats longer than the reads can be resolved by spanning paired ends, but the analysis is more complicated. Complete resolution usually requires two resources: pairs that straddle the repeat with each end in unique sequence, and pairs with exactly one end in the repeat. The limit of repeat resolution can be explored for finished genomes under some strict assumptions. For instance, it was shown that the theoretical best assembly of the E. coli genome from 20 bp unpaired reads would put 10 percent of bases in contigs of 10 Kbp or longer given infinite coverage and error-free reads [141].

The limit calculation is not straightforward for reads with sequencing error, paired-end reads, or unfinished genomes. Careful estimates of repeat resolution involve the ratio of read length (or paired-end separation) to repeat length, repeat fidelity, read accuracy, and read coverage. In regard to NGS data, shorter reads have less power to resolve genomic repeats but higher coverage increases the chance of spanning short repeats. Repeat resolution is made more difficult by sequencing error. Software must tolerate imperfect sequence alignments to avoid missing true joins. Error tolerance leads to false positive joins. This is a problem especially with reads from inexact (polymorphic) repeats. False positive joins can

induce chimeric assemblies. In practice, tolerance for sequencing error makes it difficult to resolve a wide range of genomic phenomena: polymorphic repeats, polymorphic differences between non-clonal asexual individuals, polymorphic differences between non-inbred sexual individuals, and polymorphic haplotypes from one non-inbred individual. If the sequencing platforms ever generate error-free reads at high coverage, assembly software might be able to operate at 100% stringency.

WGS assembly is confounded by non-uniform coverage of the target. Coverage variation is introduced by chance, by variation in cellular copy number between source DNA molecules, and by compositional bias of sequencing technologies. Very low coverage induces gaps in assemblies. Coverage variability invalidates coveragebased statistical tests, and undermines coverage-based diagnostics designed to detect over-collapsed repeats. WGS assembly is complicated by the computational complexity of processing larger volumes of data. For efficiency, all assembly software relies to some extent on the notion of a K-mer. This is a sequence of K base calls, where K is any positive integer. In most implementations, only consecutive bases are used. Intuitively, reads with high sequence similarity must share K-mers in their overlapping regions, and shared K-mers are generally easier to find than overlaps. Fast detection of shared K-mer content vastly reduces the computational cost of assembly, especially compared to all-against-all pairwise sequence alignment. A tradeoff of K-mer based algorithms is lower sensitivity, thus missing some true overlaps. The probability that a true overlap spans shared K-mers depends on the value of K, the length of the overlap, and the rate of error in the reads. An appropriate value of K should be large enough that most false overlaps don't share K-mers by chance, and small enough that most true overlaps do share K-mers. The choice should be robust to variation in read coverage and accuracy. WGS assembly algorithms, and their implementations, are typically complex. Assembly operation can require high-performance computing platforms for large genomes. Algorithmic success can depend on pragmatic engineering and heuristics, that is, empirically

derived rules of thumb. Heuristics help overcome convoluted repeat patterns in real genomes, random and systematic error in real data, and the physical limitations of real computers.

4.7 Next generation sequence assemly algorithm

4.7.1 Overlap-layout-consensus

Intuitively, finding read overlaps should be the primary idea behind assembly. The overlap-layoutconsensus framework is merely a three-step process [92]. First, the overlaps between reads are computed. Second, the order and orientation of reads in the assembly are determined. Third, the consensus allows the determination of the nucleotide at each position in the contigs. Assemblers implementing this idea are numerous, and were crafted to overcome the assembly hurdles in Sanger-technology projects before high-throughput systems were developed. These software are tailored for the assembly of long reads, such as Sanger reads. They include the Celera assembler [92] and Arachne [9]. Afterwards, the paradigm was adapted to the Roche/454 system. The Roche/454 sequencer is distributed with Newbler [87], as shown in Table 4.1 Also, EDENA is an overlap-layout-consensus assembler that can assemble short reads (35 bases) [48].

4.7.2 Greedy assemblers

Using overlaps between reads to build contigs is intuitive. Algorithms can be built upon this idea. A greedy algorithm iteratively grows contigs by choosing best overlaps first. Implementations of the greedy algorithm were the first to be introduced for the short read technologies: SSAKE [140], VCAKE [57], and SHARGCS [32].

4.7.3 Assembly with de Bruijn graphs

The introduction of the de Bruijn graph for the AP is motivated by redundant information in reads: a large depth of coverage implies that a lot of overlaps between reads occurs.

In presence of such redundant information, de Bruijn assemblers can solve the assembly problem with a memory usage bounded by the genome length [152].

Table 4.1: Next Generation Sequence Assembly Summary

Assembler	Algorithm	Parallel	Language
Euler-SR	De Bruijin Algorithm	Yes	C++
ALLPaths	De Bruijin Algorithm	Yes	C++
SOAP	De Bruijin Algorithm	Yes	C++
Velvet	De Bruijin Algorithm	No	C
ABySS	De Bruijin Algorithm	Yes	C++
QSRA	De Bruijin Algorithm	No	C++
RAY	De Bruijin Algorithm	Yes	C++
CABOG	Overlap-layout Consensus	No	C
Newbler	Overlap-layout Consensus	No	—
Shorty	Overlap-layout Consensus	No	—
Edena	Overlap-layout Consensus	No	—
MIRA	Overlap-layout Consensus	No	—
SHARCGS	Greedy Algorithm	NO	Perl
SSAKE	Greedy Algorithm	NO	Perl
VCAKE	Greedy Algorithm	NO	Perl

4.8 Materials and Methods

4.9 Results

4.9.1 Evaluation Environment

Table 4.2: The test platform

	Cluster
CPU	Intel Xeon X3430
Memory	2GB
OS	Ubuntu 10.04 (32 bit)
Network	1000Mbps
SSD	Intle X-25M 80GB
RAID 0	4 WD50000AAKS

As shown in Table 4.2, single Machine Configuration: HP ProLiant ML110G6 brand computer with a 2.4Ghz, 4 core, Intel Xeon processor, 2GB of RAM, and 160GB HDD,

running Ubuntu 10.04 (32 bit version). There is no cluster or network overhead. The machine is functioning as a local, stand-alone unit, and is considered to be a comodity machine due to it's common hardware and affordable price. Cluster Configuration: 12 machines with hardware identical to the single machine using a Gigabit switch, comprise the twelve node cluster used for experiments. File Sizes: 25MB, 50MB, 100MB, 150MB, 200MB, and 250MB. Data Collected: All data collected for use in figures and graphs are an average of three experimental trials and are expressed in seconds. We choose an Intel X-25M 80GB solid state disk, and a SATA Raid tower with four WD50000AAKS disks as a RAID 0 array.The MPICH2-1.0.7 is chosen as the message passing interface (MPI) in the cluster.

4.10 Project Description

MPI-Velvet is implemented in C and uses the MPI protocol for communication between nodes. The latency and bandwidth of the cluster network can have a significant impact on the performance of a parallel application and require special consideration. Each communications message is given a unique identifier. The sending process does not wait for an immediate response, but rather saves the current state of the operation, keyed by the message ID, and continues to process other operations. When a response is received for a particular message, the saved state information is retrieved by using the message ID and the original task continues. This system allows many simultaneous operations on each cluster node, effectively hiding the latency of the network link. As the messages passed between cluster nodes tend to be very short, the messages are collected into larger, 1 kB packets to minimize the impact of communication overhead.The goal of this project is to improve the next generation sequence assembly efficency using the MPI package. This research project investigates the MPI version Velvet sequence assembly within a trusted cluster. The performance of MPI-Velvet was tested based on different input data size and different computing nodes.

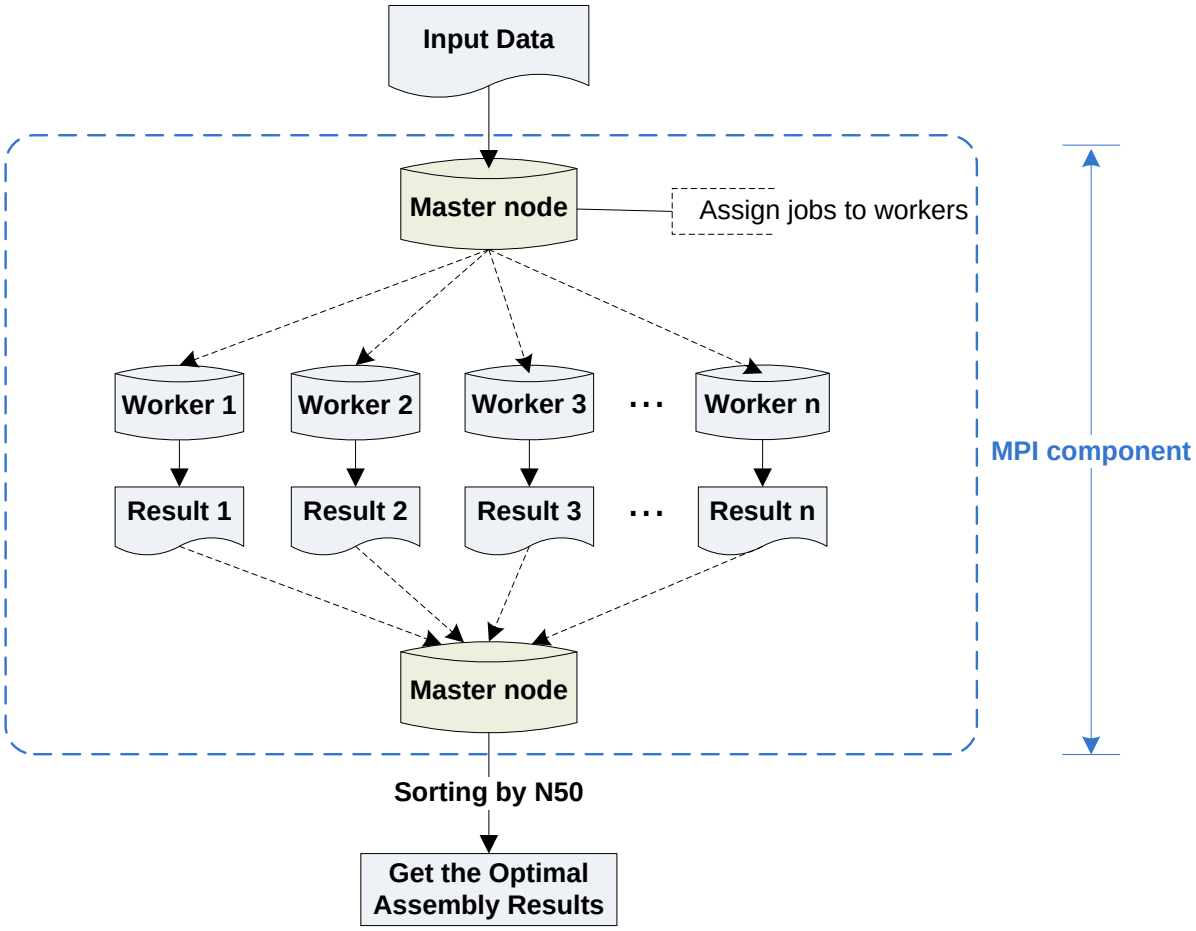


Figure 4.2: MPI-Velvet data workfolw

4.11 Materials and Methods

The MPI-Velvet was built based on the original Velvet assembly software. Velvet assembly software was implemented with C language. So, it was the more efficent approach using MPI component to acheive the parallel assembly software. The data workflow is showed in Fig. 4.2. In summary, there was Four main steps to finish the data analysis using MPI-Velvet. **Step 1:** Master node to assign the job to the workers; **Step 2:** Workers node analysis data seperately; **Step 3:** Send the results back to the Master node; **Step 4:** Get the optimal assembly result by sorting the statistical results.

4.12 Algorithm

Algorithm 1 MPI-Velvet algorithm

SET *results* be the current set of Velvet results
SET F = $\{f_1, f_2, ...\}$ be the set of database fragments
SET **Unsearched** $\subseteq$ **F** be the set of unsearched database fragments
SET **Unassigned** $\subseteq$ **F** be the set of unassigned database fragments
SET **W** = $\{w_1, w_2, ...\}$ be the set of participating workers
SET $\mathbf{D}_i \subseteq \mathbf{W}$ be the set of workers that have fragment f_i on local storage
SET **Distributed** = $\{\mathbf{D}_1, \mathbf{D}_2, ...\}$ be the set of **D** for each fragment
Require: $|\mathbf{W}| \neq 0$
Ensure: $|\mathbf{Unsearched}| = 0$
Unsearched $\leftarrow$ **F**
Unassigned $\leftarrow$ **F**
Broadcast queries to workers
while $|\mathbf{Unsearched}| \neq 0$ **do**
  Receive a *message* from a worker w_j
  if *message* is a state request **then**
    if $|\mathbf{Unsearched}| = 0$ **then**
      Send worker w_j the state SEARCH_COMPLETE
    else
      Send worker w_j the state SEARCH_FRAGMENT
    end if
  else if *message* is a fragment request **then**
    Find f_i such that $f_i \in Unassigned$
    if $|\mathbf{D}_i| = 0$ **then**
      Add w_i to $\mathbf{D}_i$
    end if
    Remove f_i from **Unassigned**
    Send fragment assignment f_i to worker w_j
  else if *message* is a set of search results for fragment f_i **then**
    Merge *message* with *results*
    Remove f_i from **Unsearched**
  end if
end while

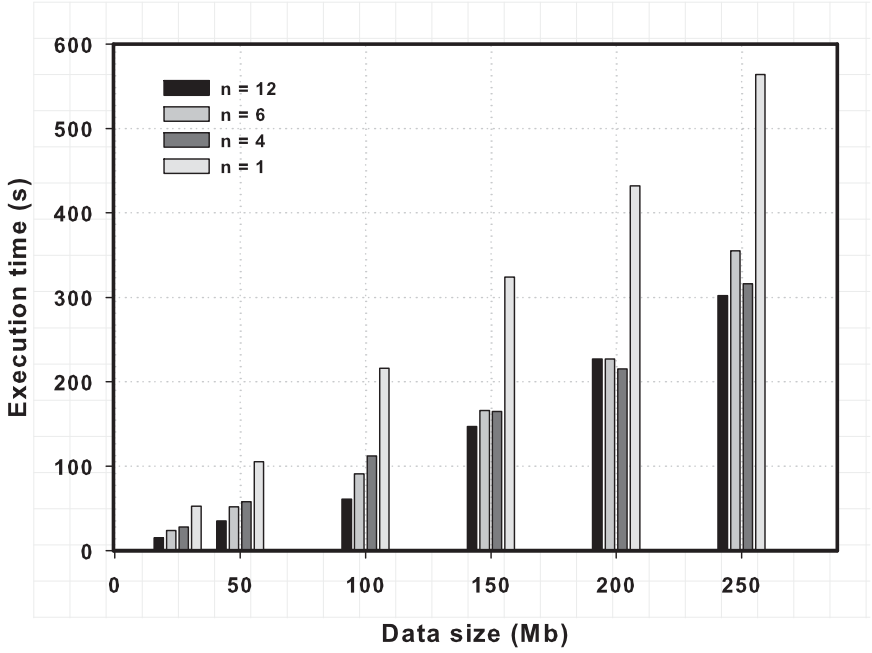


Figure 4.3: The execution time comparison between sequential Velvet and MPI-Velvet

4.13 Results

4.13.1 Performance comparison between sequential velvet and MPI-velvet

Fig. 4.3 shows the runtime between sequential Velvet and MPI-Velvet with different computing nodes including n=4, n=6, and n=12. Clearly, the execution time increased along with the computing node decrement. The best speedup performance was achieved by 12 computing nodes, which can reach 3.5X compared with single node. Meanwhile, the speedup performance was stable when n = 4 and n = 6 respectively, which is around 2X compared with single node (see Table 4.3). In terms of different data size, such as 25MB, 50MB, 100MB, 150MB, 200MB, and 250MB, the speedup performance was stable when n=4, which is around 2X faster than single computing node. The speedup performance demonstrated big variation using the input data size was different when n=12. As the Fig. 4.4 showed that, when the input data size was smaller than 150 MB, the speedup reached 3.5X at most compared with single computing node. The speedup for the MPI-Vevlet was evaluated

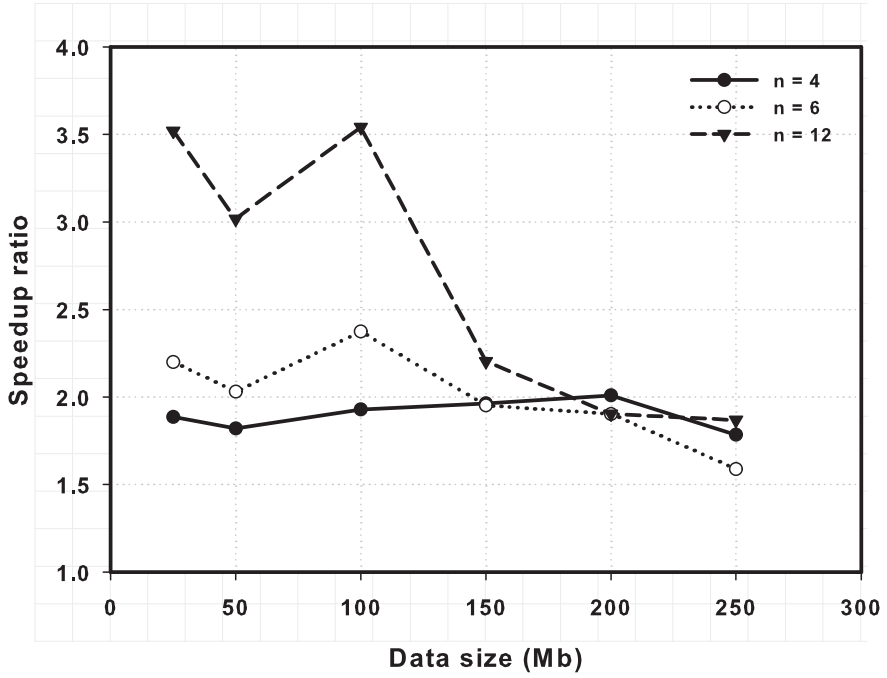


Figure 4.4: The speedup comparisons of different MPI-Velvet nodes

using the Eq. 4.1.

$$Speedup = \frac{T_{singlenode}}{T_{mpinode}} \tag{4.1}$$

Table 4.3: The execution time and speedup comparisons between sequential Velvet and MPI-Velvet using different datasets

Data size(MB)	n=1(s)	n=4(s)	speedup	n=6(s)	speedup	n=12(s)	speedup
25	52.8	28	1.9	24	2.2	15	3.5
50	105.6	58	1.8	52	2.0	35	3.0
100	216	112	1.9	91	2.4	61	3.5
150	324	165	2.0	166	2.0	147	2.2
200	432	215	2.0	227	1.9	227	1.9
250	564	316	1.8	355	1.6	302	1.9

4.13.2 Performance comparison among HDD, SSD, and RAID

The performance was tested on different devices including Hard Disk Drive (HDD), Solid State Disk (SSD), and RAID. The I/O performance and CPU usage using HDD, SSD and RAID was showed separately in Fig. 4.5, Fig. 4.6, and Fig. 4.7. There was no big

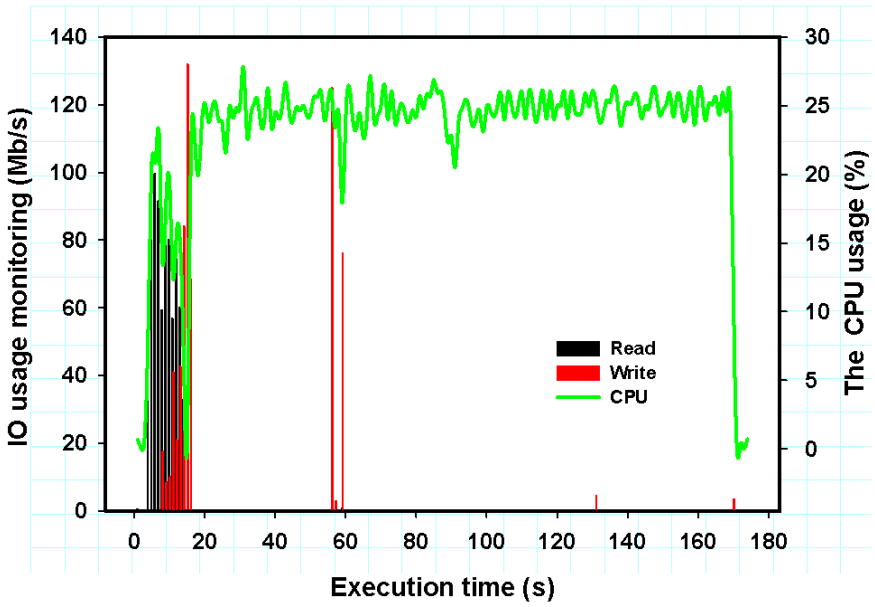


Figure 4.5: Velvet I/O activities of HDD

difference among HDD, SSD, and RAID in terms of the execution time. However, there was large variation among these three devices based on the I/O performance pattern. As Fig. 4.6 and Fig. 4.7 shown that there was a lot of the read and write activities at the beginning. However, there was little read activity and longer write time in SSD (See Fig. 4.6).

Table 4.4: The execution time and speed ratio comparisons various computing nodes using different datasets

Node	50MB	$Ratio_{(\frac{100MB}{50MB})}$	100MB	$Ratio_{(\frac{200MB}{100MB})}$	200MB	$Ratio_{(\frac{200MB}{50MB})}$
4	58s	1.9	112s	2.0	220s	3.8
6	52s	1.8	91s	2.5	227s	4.4
8	48s	1.8	85s	2.4	205s	4.3
12	35s	1.7	61s	3.7	227s	6.5

4.13.3 Execution time comparison among different computing node with various input data size

The execution time comparison among various data size in different computing nodes was tested in this project. The result shows that the execution time decreased along with the increment of computing node when the data size was 50MB and 100MB. However, when the

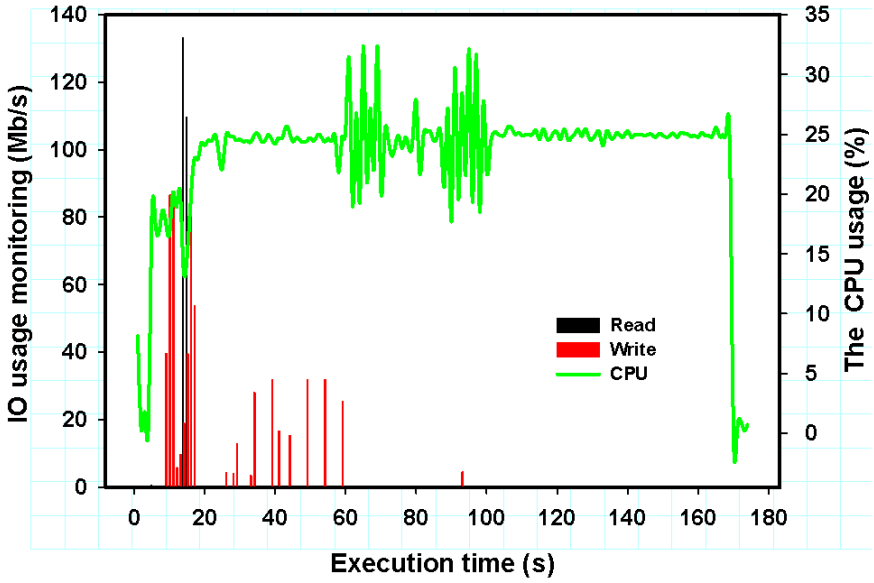


Figure 4.6: Velvet I/O activities of SSD

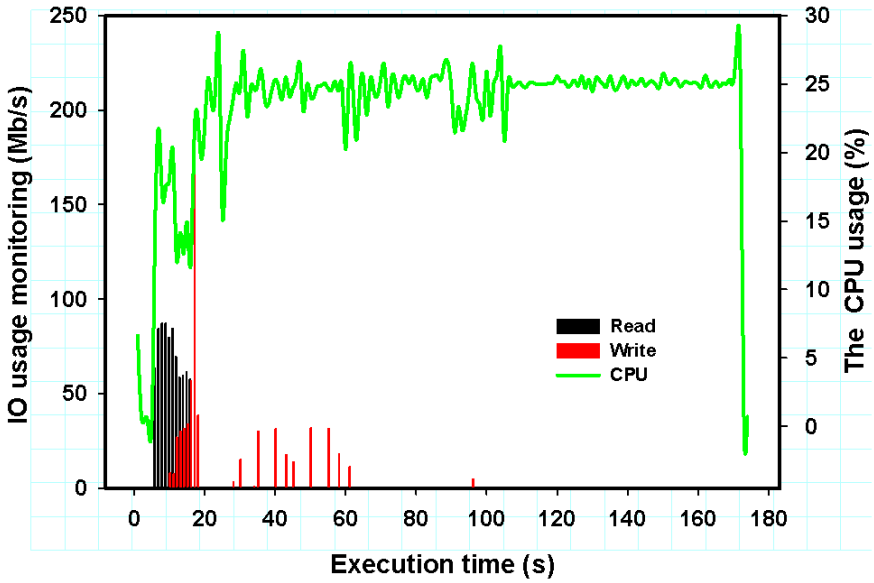


Figure 4.7: Velvet I/O activities of RAID

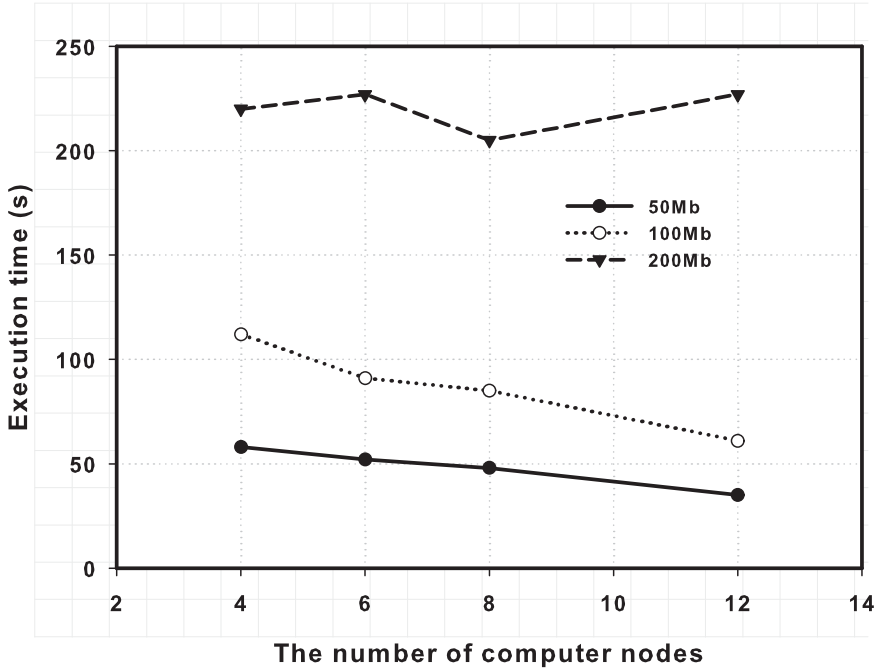


Figure 4.8: Impact of the number of computing nodes on execution time. Data size is set to 50, 100, and 200 MB, respectively.

data size reached to 200MB, the execution time was very stable in different computing node (See Fig. 4.8). Also, the data size we tested 2 times in depth. Interestedly, (See Table 4.4) the execution time ratio of 100MB/50MB was 1.7 to 1.9, which was very consistent among different computing node. While, the ratio of 200MB/100MB 2.0, 2.5, 2.4 and 3.7 in 4 node, 6 node, 8 node and 12 node respectively. It indicted that the execution time increased a lot along with the computing node increment when the input data size became larger.

4.13.4 Speedup comparison among different computing node with various input data size

The speedup comparison among various data size in different computing nodes was also tested in this project. The result showed that the speedup performance was the best when the input data size was 100MB. Also the speedup performance incresed proportionally with

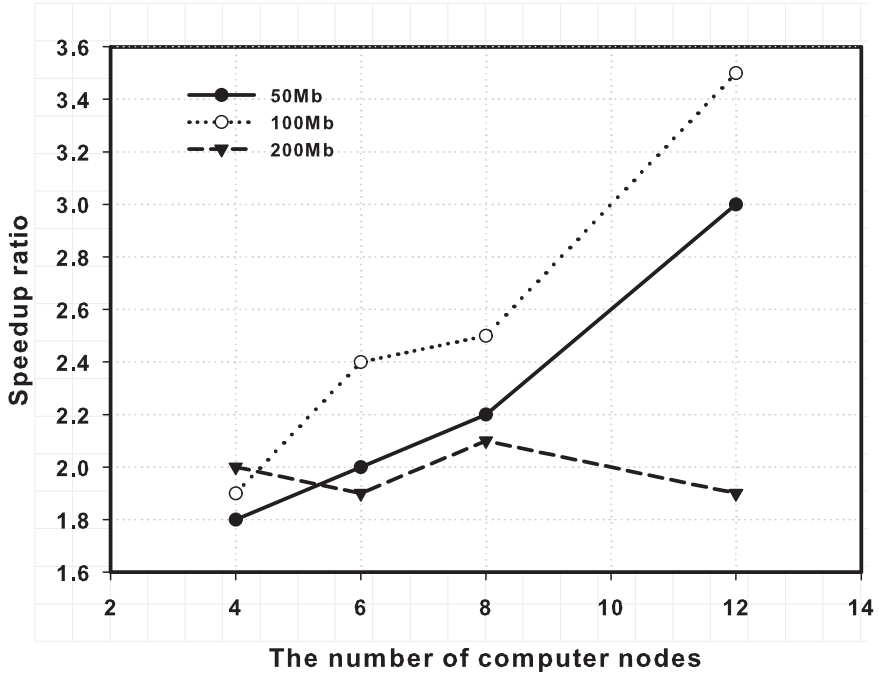


Figure 4.9: The speedup ratio comparison among various datasize in different computing nodes

the computing node when the input data size was 50MB and 100MB. However, the speedup performance maintained stable stage when the input data size was 200MB (See Fig. 4.9).

4.13.5 I/O bandwidth monitoring with different computing node using same input data size

In order to monitor the read/write bandwidth with different computing node, the I/O performance was recorded when the MPI-Velvet computing performance was tested using the same input data size (100MB), as shown in Fig. 4.10, Fig. 4.11, and Fig. 4.12. The computing node was n = 4, n = 6, and n = 12 respectively. These monitoring results illustrated that when the computing node was small such as n = 4 and n = 6, there existed densely reading activity along with the whole procedure. While, there was much less reading activity when n = 12.

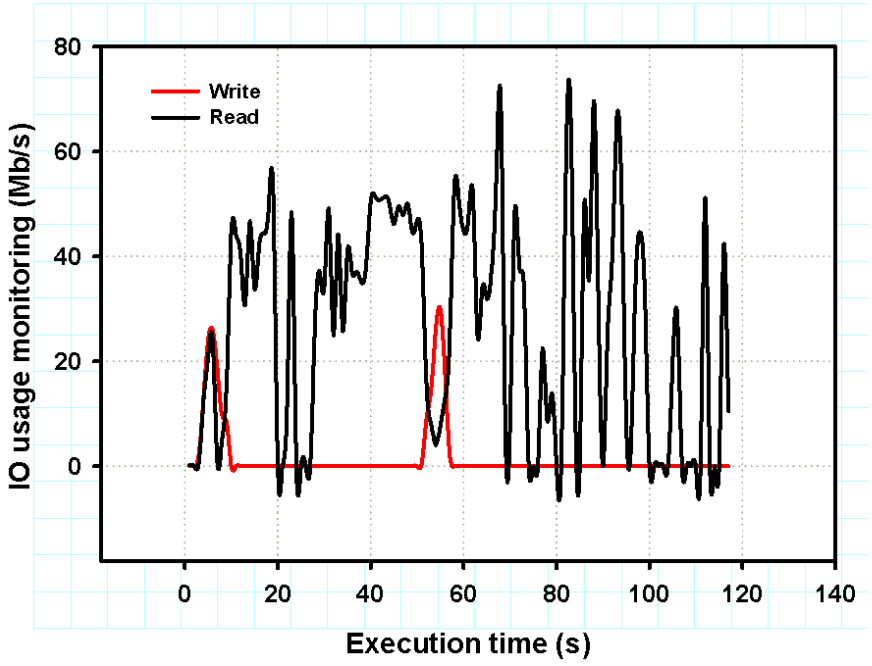


Figure 4.10: MPI-Velvet I/O bandwidth monitoring with 4 computing nodes

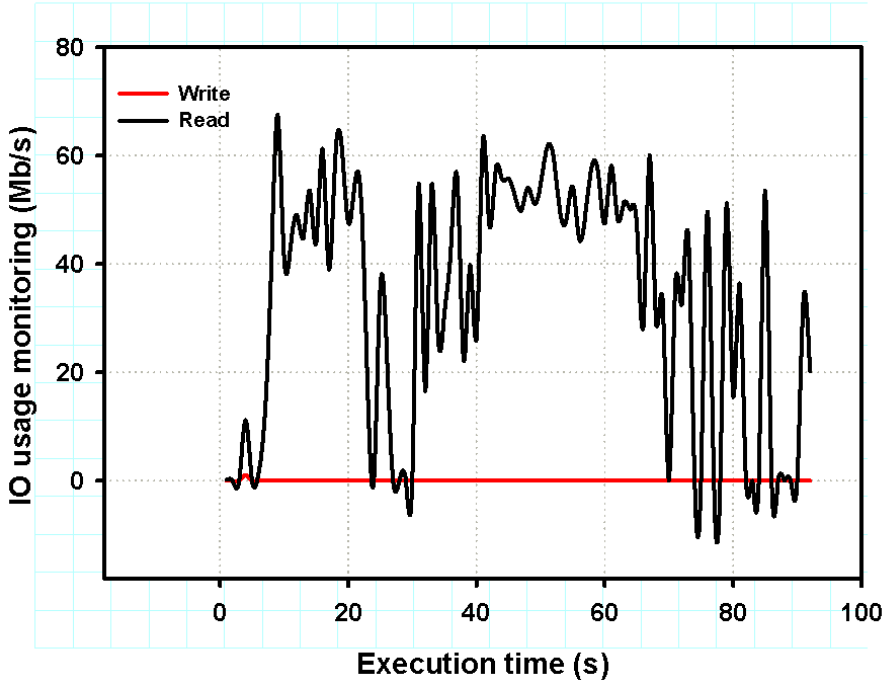


Figure 4.11: MPI-Velvet I/O bandwidth monitoring with 6 computing nodes

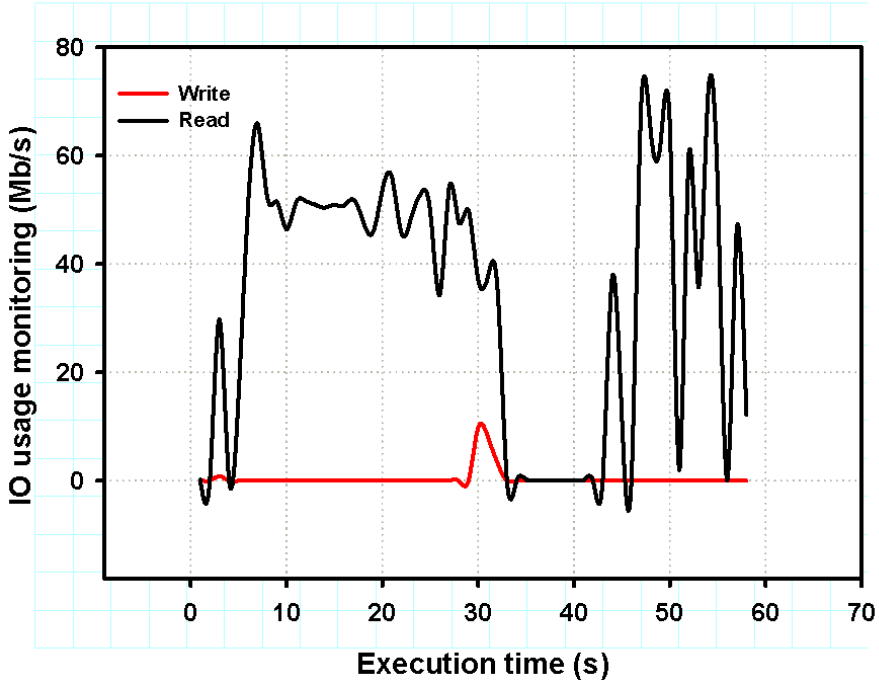


Figure 4.12: MPI-Velvet I/O bandwidth monitoring with 12 computing nodes

4.13.6 I/O bandwidth monitoring with different input data size using same computing node

In order to monitor the read/write bandwidth with different input data size, the I/O performance was recorded when the MPI-Velvet computing performance was tested using the same computing node (n = 6), as shown in Fig. 4.13, Fig. 4.14, and Fig. 4.15. The input data size was 50MB, 100MB, and 200MB respectively. These monitoring results illustrated that when the input data size was small such as 50MB and 100MB, there existed less reading activity along with the whole procedure. While, there was densely reading activity when the input data size reached 200MB.

4.14 Conclusion

- The best speedup performance was achieved by 12 computing nodes, which can reach 3.5X compared with single node. Meanwhile, the speedup performance was stable

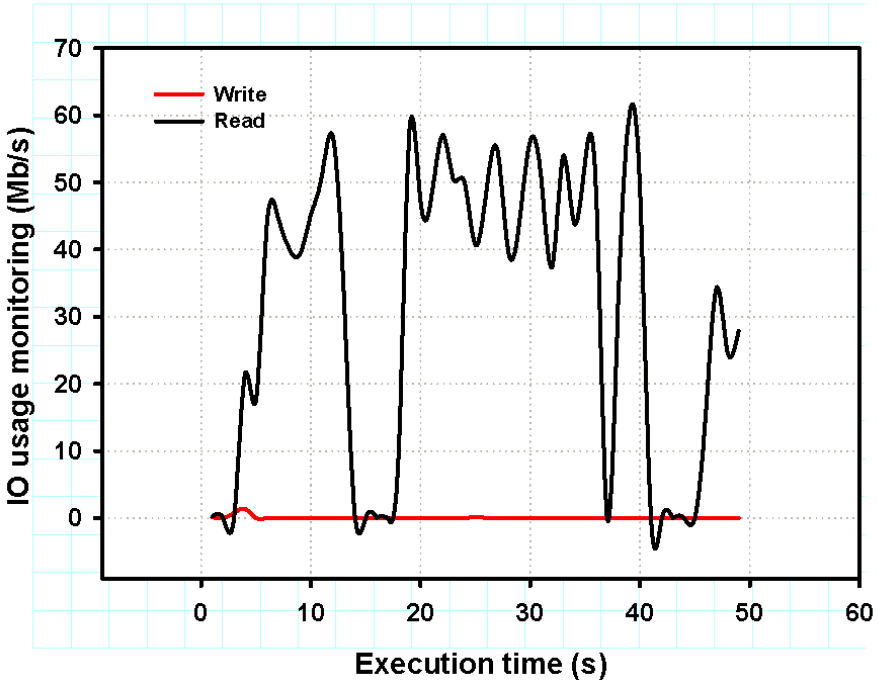


Figure 4.13: MPI-Velvet I/O bandwidth monitoring with 50MB data size

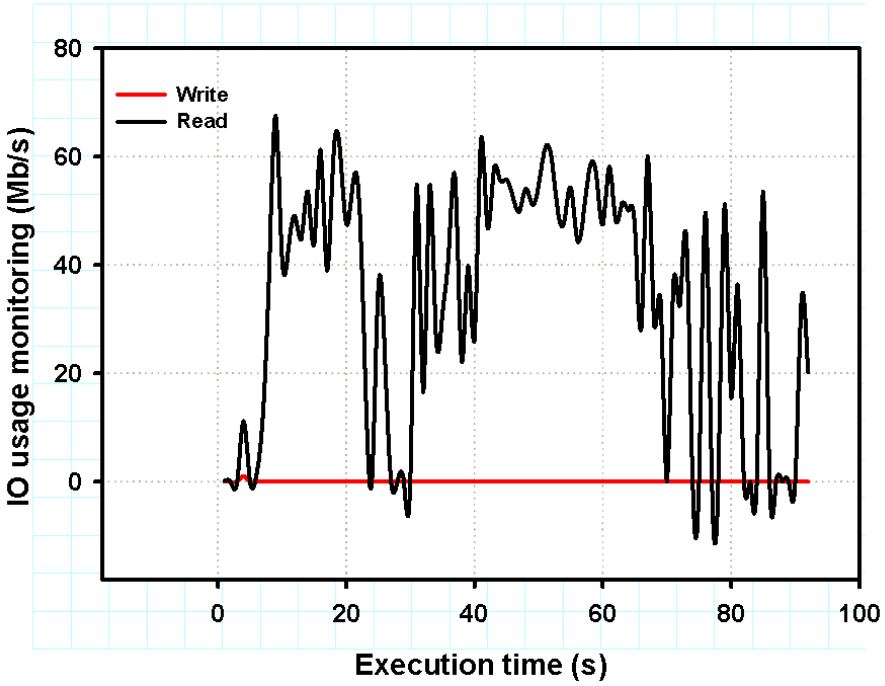


Figure 4.14: MPI-Velvet I/O bandwidth monitoring with 100MB data size

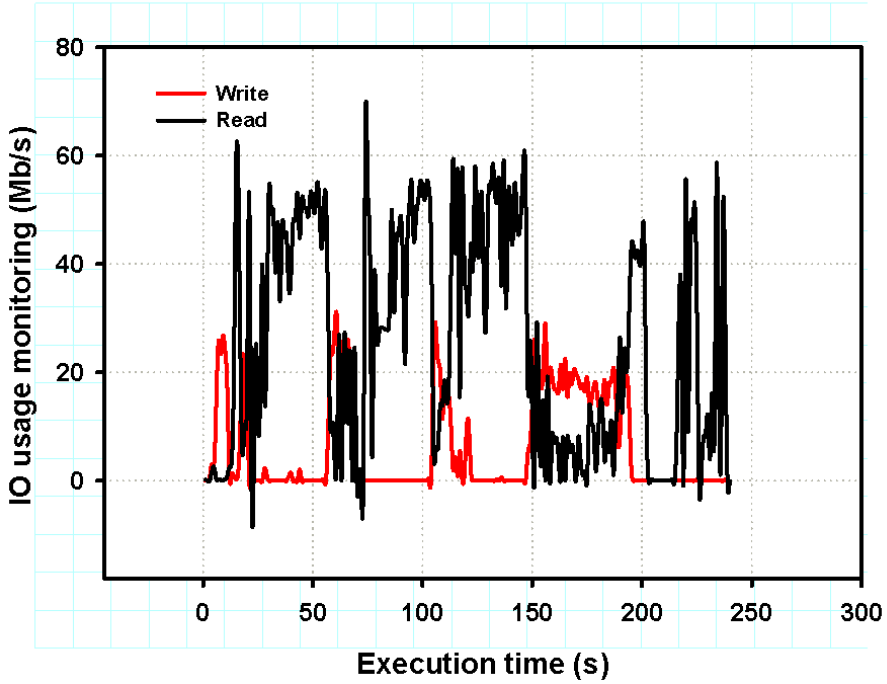


Figure 4.15: MPI-Velvet I/O bandwidth monitoring with 200MB data size

when n = 4 and n = 6 respectively, which is around 2X compared with single node. The speedup performance demonstrated big variation using the input data size was different when n=12. When the input data size was smaller than 150 MB, the speedup reached 3.5X at most compared with single computing node.

- The speedup comparison among various data size in different computing nodes was also tested in this project. The result showed that the speedup performance was the best when the input data size was 100MB. Also the speedup performance incresed proportionally with the computing node when the input data size was 50MB and 100MB. However, the speedup performance maintained stable stage when the input data size was 200MB.

4.15 Future Work

- Hadoop implementation for MPI-Velvet. Similarly, the hadoop version for Velvet can be implemented using the same algorithm. The only difference between VelvetDoop and MPI-Velvet is the programming environment.

- Compare MPI-Velvet with VelvetDoop. A series of test will be run to compare the MPI-Velvet and VelvetDoop. The performance can be evaluated using these two different parallel programs.

- Use VelvetDoop as a hadoop benchmark to evaluate performance of hadoop clusters. After the VelvetDoop is generated, the reliability and scalability performance will be tested using a series of input sequence data in our local cluster system.

Chapter 5

The catfish genome database cBARBEL: an informatic platform for genome biology of ictalurid catfish

5.1 Abstract

The catfish genome database, cBARBEL(abbreviated from catfish Breeder And Researcher Bioinformatics Entry Location) is an online open-access database for genome biology of ictalurid catfish (*Ictalurus spp.*). It serves as a comprehensive, integrative platform for all aspects of catfish genetics, genomics and related data resources. cBARBEL provides BLAST-based, fuzzy and specific search functions, visualization of catfish linkage, physical and integrated maps, a catfish EST contig viewer with SNP information overlay, and GBrowse-based organization of catfish genomic data based on sequence similarity with zebrafish chromosomes. Subsections of the database are tightly related, allowing a user with a sequence or search string of interest to navigate seamlessly from one area to another. As catfish genome sequencing proceeds and ongoing quantitative trait loci (QTL) projects bear fruit, cBARBEL will allow rapid data integration and dissemination within the catfish research community and to interested stakeholders. cBARBEL can be accessed at http://catfishgenome.org.

5.2 Introduction

Catfish (*Ictalurus spp.*) is an important aquaculture species in the United States, accounting for over 60% of domestic aquaculture production. While channel catfish (*Ictalurus punctatus*) accounts for the large majority of farm-raised catfish, increasing numbers of channel catfish female X blue catfish (*Ictalurus furcatus*) male hybrids are being cultured. Facing

rising feed costs and stiff international competition, catfish producers require improvement in fish production and performance traits such as disease resistance, growth rate and feed conversion efficiency to maintain profitability. Efficient utilization of the natural diversity of trait phenotypes already present in different species, strains and hybrids of catfish for selection of superior broodstock, requires the identification of genetic underpinnings of trait differences. Toward this end and eventual marker-assisted selection (MAS), significant genome resources have been developed in catfish. These include over a half million expressed sequence tags (ESTs) [21] [59] [60] [66] [138] [70], a large number of genome sequences generated from bacterial artificial chromosome (BAC ends [73], genetic linkage maps [76] [136] [67], genome physical maps [111] [147], tens of thousands of microsatellite markers [120] [7], hundreds of thousands of singlenucleotide polymorphisms (SNPs) [139], over 10,000 full-length cDNAs (flcDNA; [24]) and an alternative splicing database [79]. Additionally, USDA NIFA funding has been secured to allow catfish whole-genome sequencing and the development of high-density SNP chips for catfish. With these and many more genome-oriented projects from catfish currently underway, the catfish research community needed a central repository for storing and integrating genomic data and a bioinformatic entry location for public access to currently inaccessible specialized data sets. To meet this need, we have created a catfish genome database, cBARBEL, the Catfish Breeder and Researcher Bioinformatic Entry Location, a title that makes use of the distinctive whisker-like organs that give catfish their name. cBARBEL represents one of the first comprehensive bioinformatic databases for an aquaculture species, although genome sequencing is planned or proceeding in close to a dozen different species from this fast-growing sector. cBARBEL provides wide-ranging query functions to facilitate user access to a host of catfish genome resources and integrates a variety of previously scattered data types. Here, we present an overview of cBARBEL search tools, platforms and functions connecting catfish EST, fl-cDNA, SNP, BAC-end sequence (BES), molecular marker, linkage map and physical map data. cBARBEL can be accessed at `http://catfishgenome.org/`.

5.3 Materials and Methods

Several software packages were used in the construction of cBARBEL database prerequisites including:

- the operating system, Ubuntu 9.10 Linux system (`http://www.ubuntu.com/`);
- Apache web server version 2.0 (`http://www.apache.org/`);
- MySQL database management system, version 5.1 (`http://www.mysql.com/`);
- Selective Perl modules and configuration of Bioperl and PHP (`www.perl.org`; `www.bioperl.org`; `http://php.net/`).

The Generic Genome Browser (GBrowse) package (v1.7), a component of the Generic Model Organism Project (GMOD), was utilized for display of the catfish physical map, EST contig viewer and zebrafish comparative GBrowse. Another GMOD component, CMap [149], was used to display and align genetic and physical (FPC) maps.

5.4 Results

5.4.1 cBARBEL database schematic

The cBARBEL database schematic is shown in Figure 5.1. allowing visualization of potential data connections. The integration of existing and forthcoming catfish genome resources provided by the cBARBEL platform should speed research progress. Utilizing cBARBEL, for example, a user searching for a given gene in catfish is able to search with a similar sequence from a related species, such as zebrafish, identify matching catfish ESTs, identify the corresponding EST contig, identify SNP markers within that contig, visualize linkage map position of these markers and relate linkage and physical map locations through a comparative map system, all by navigating through a series of intuitive links. To do a similar search without cBARBEL would require interrogating a series of separate public and

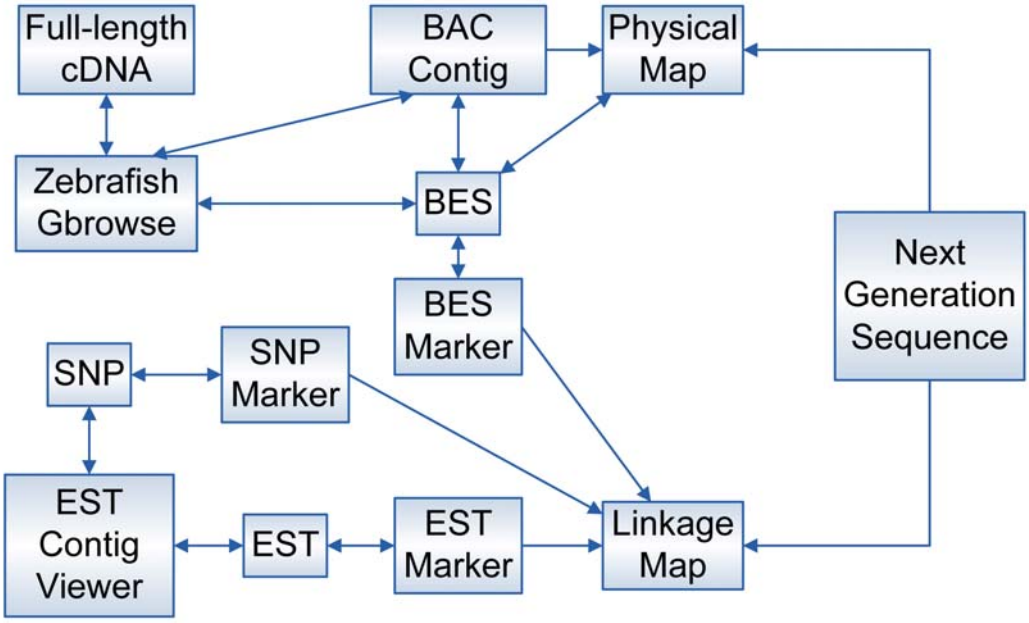


Figure 5.1: The cBARBEL database schematic

local databases, relating disparate nomenclatures, would not include visualization and could take well over 30 min for an individual query.

cBARBEL currently is organized around three components: nucleotide sequence, genetic markers and maps. These components are brought together by search tools and multi-directional links. Features of each of these components are described briefly below

5.4.2 Sequence search function

Sequence similarity searches of catfish database cBARBEL database provides the catfish BLAST search function. It can be used to search all or subset of the catfish sequences including catfish EST, catfish BES, Full-length cDNA, and catfish all (Figure 5.2). The search results can link further not only NCBI server, but also can link to the corresponding location on physical map (BES) / EST contig viewer (EST) and zebrafish GBrowse.

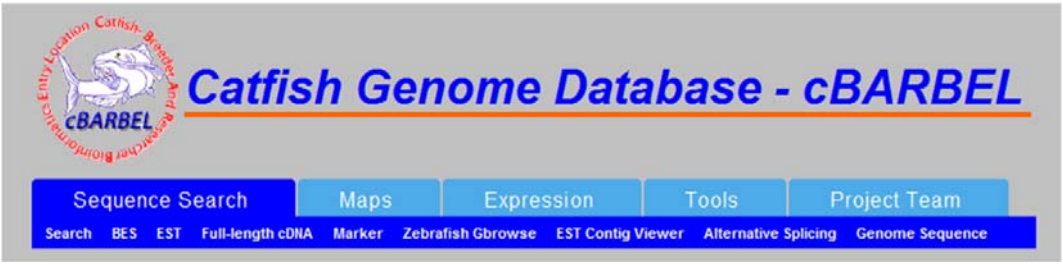


Figure 5.2: The navigation of cBARBEL database allow user to easily switch different sections. The entries on the second line can let user to move directly to a specific cBARBEL page

5.4.3 Specific search function of catfish database

BLAST searches can be carried out to query all or a subset of catfish sequences including catfish ESTs, catfish BES, full-length cDNA and catfish all (Figure 5.3). Search results provide further links to NCBI records and location of the hit on the catfish physical map (BES), EST contig viewer (EST) and zebrafish GBrowse chromosomes (All).

The specific search function can provide data access using a variety of search terms (Figure 5.3). Catfish ESTs can be queried using GenBank accession numbers, marker names for those ESTs containing a SNP or microsatellite marker, and EST contig ID. Search results provide sequence links as well as deeper connections to the EST contig viewer, zebrafish GBrowse comparative alignments and linkage map position, where appropriate. Similar searches can be carried out for BES and fl-cDNA. Those looking for a marker of interest can use fuzzy or specific search terms. For example, a search such as 'AUEST' returns all mark information for EST microsatellite markers generated at Auburn University. The resulting marker table contains the accession number of the relevant sequence, primer information, marker status (polymorphic, not polymorphic, untested), linkage map position, physical map position (BES) and EST-contig (EST).

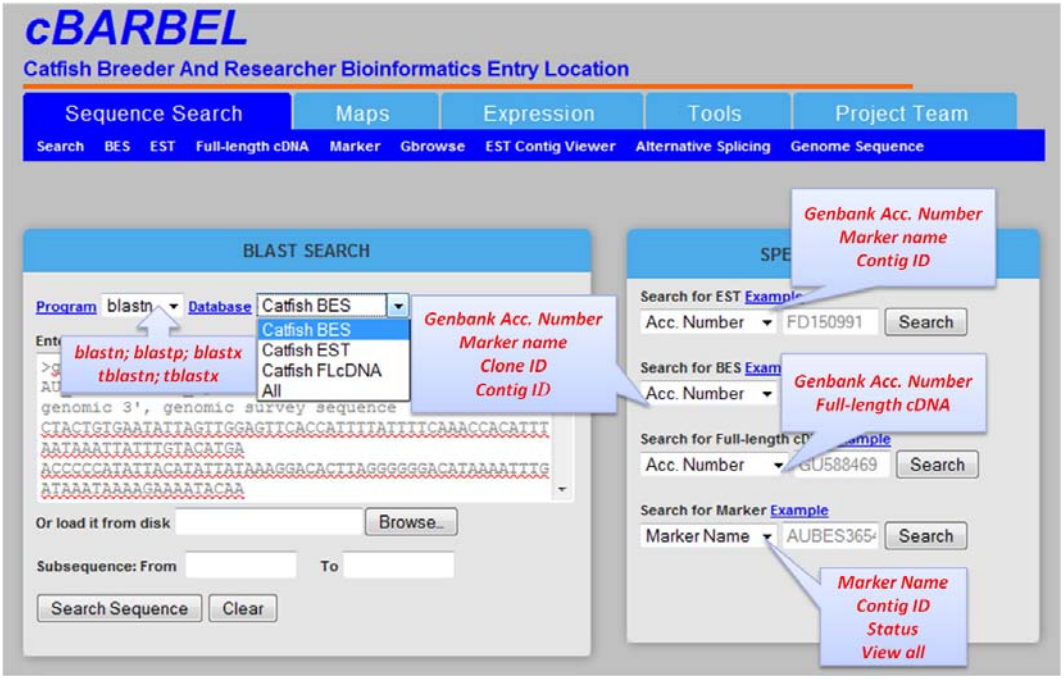


Figure 5.3: cBARBEL database sequence search function. It includes the catfish BLAST search function and the specific search function. Four different subjects were provided in BLAST search function, including catfish EST, catfish BES, Full-length cDNA, and catfish all. The specific search function can be used including EST specific search, BES specific search, Full-length cDNA specific search, and marker specific search

5.4.4 Zebrafish GBrowse genomic viewer versus catfish genomic dataset

GBrowse is a GMOD tool that displays features of the genome aligned to a genomic sequence [125]. GBrowse is easily customized to allow a variety of data tracks and third-party data types to be visualized. Catfish whole-genome sequencing is underway, but, in the interim, we have aligned a number of catfish genome data types to the genome sequence of zebrafish, the closest evolutionarily related species with an available sequenced genome (Figure 5.4). The alignments (based on tblastx similarity) help to organize catfish data on a genome scale, and harnessing synteny between the two species has proven useful in gene isolation and QTL fine-mapping studies. As catfish genome assembly proceeds, this comparative approach should also prove useful in scaffolding catfish supercontigs. By default, cBARBEL presents a view of catfish EST contigs, catfish singleton ESTs, catfish BES, and catfish fl-cDNA aligned to zebrafish chromosomal sequences (Figure 5.4). For each of these tracks, clicking on a feature provides a related link to the NCBI server or a link to a local copy of the sequence information. Users can locate a specific region of interest by entering a specific sequence range or by dragging the selection box to specify a chromosomal region. Additionally, cBARBEL BLAST and specific search outputs include zebrafish GBrowse links to allow connectivity to other database sectors.

5.4.5 Catfish EST contig viewer

The GBrowse package was also utilized to create a catfish EST contig viewer displaying the alignment of individual ESTs on the contig consensus sequence. A SNP track was also added allowing visualization of SNP density and position within the EST contig (Figure 5.5). For the EST track, clicking on an individual feature provides a related link to NCBI-based sequence information. Clicking on SNP entries allows the user to navigate to further SNP information in the AutoSNP program including SNP allele frequency, type, and position. For the contig track, a link is provided to local sequence information. As with the zebrafish

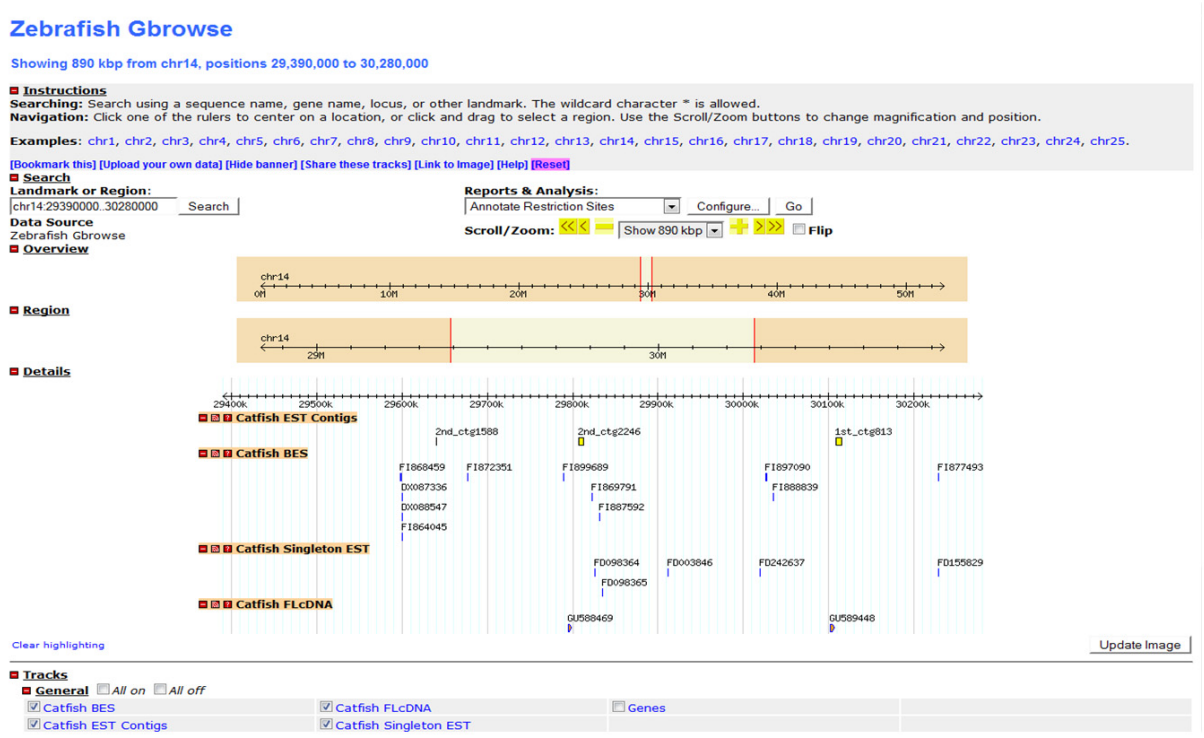


Figure 5.4: Catfish genomic sequence view of the region of zebrafish chromosome. The tracks includes Catfish EST Contigs, Catfish Singleton EST, Catfish BES, and Catfish FLcDNA. For each of these tracks, a click on a feature provides a related link to NCBI server or link to the sequence information

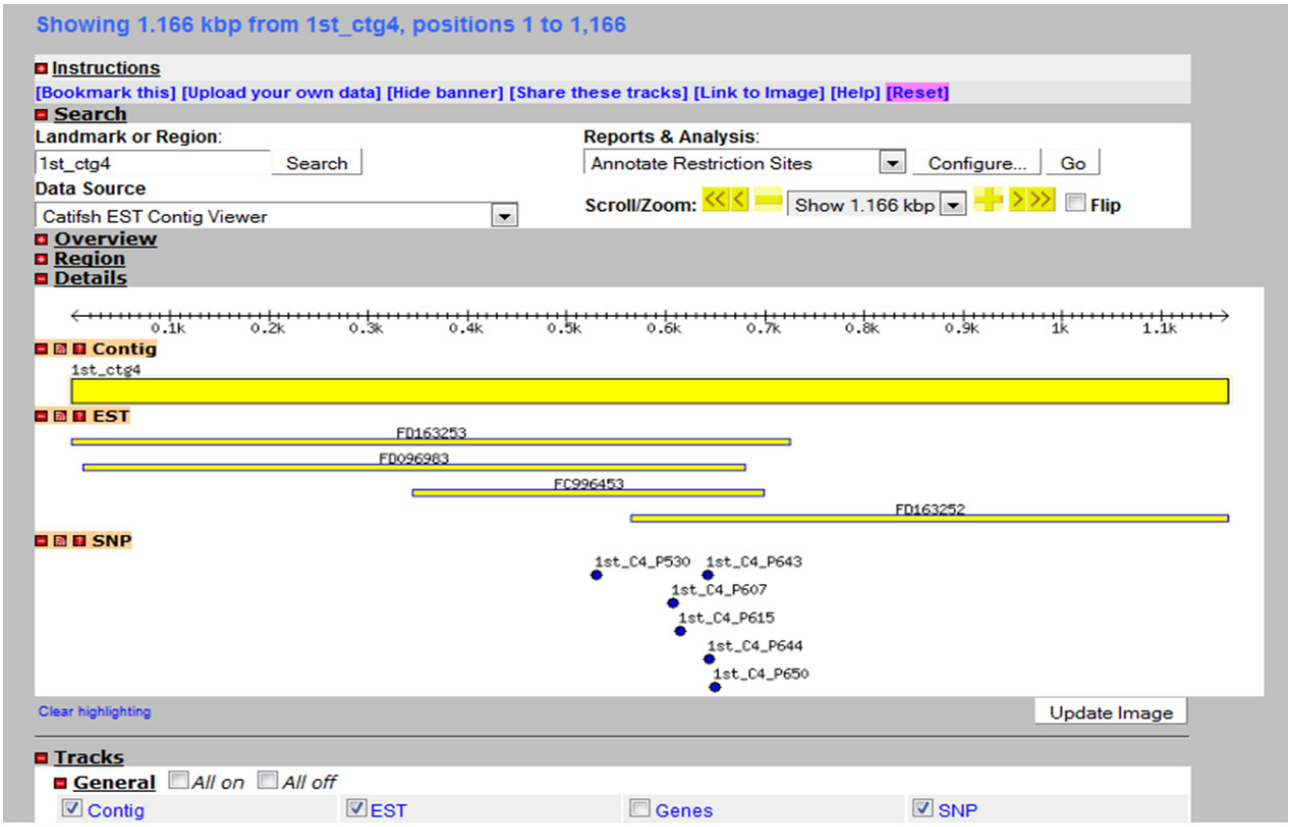


Figure 5.5: Catfish EST contig viewer. The tracks include Contig, EST, and SNP. For each of these tracks, a click on a feature provides a related link to NCBI server, link to the SNP viewer, or link to the sequence information

GBrowse view, cBARBEL BLAST and specific search outputs include links to the EST contig viewer.

5.4.6 Physical map and linkage map

We previously reported the construction of a fingerprint contig (FPC) BAC-based physical map with 3307 contigs spanning the catfish genome [147], and are currently engaged in efforts to integrate the physical map with catfish linkage maps using BES-associated microsatellites. This map was previously displayed using a Java-based program, WebFPC, which did not allow efficient search or data integration options. For example, BES associated with BAC clones could not be searched or visualized with WebFPC. To remedy these

issues, we adapted GBrowse to display both BAC contigs and BES information in a searchable, integrated format (Figure 5.6). BAC clones are displayed using their FPC position within a given contig. Custom scripts were developed to indicate the presence of BES within a clone using blue boxes and to allow BES sequence retrieval via NCBI link by clicking on the desired clone. As with other sections, cBARBEL BLAST and specific search outputs include physical map links where appropriate.

A linkage map of catfish has been constructed based on genotyping of EST-based microsatellites, SNP markers and BES-based microsatellite markers on backcross hybrid (channel X blue) families [67]. Efforts are ongoing to increase marker density using BES-associated microsatellites and SNP markers. Marker information for the 29 linkage groups is displayed in table format with marker name and map position (cM).

5.4.7 Catfish CMap - map integration

Integration of catfish linkage and physical maps is ongoing, based largely on the mapping of BES microsatellites from physical map contigs onto the catfish linkage map. While details of integration can be gathered in table format in other cBARBEL sections, we used CMap to provide visualization of LG-level map integration (Figure 5.7). Links are provided to each of the 29 linkage groups arrayed alongside corresponding physical map contigs using predetermined settings. A link is also provided to allow users to access numerous different display settings including the choice to view graphic representations of linkage groups alone. As with the linkage maps, data is continually updated as additional markers are genotyped.

5.4.8 Analysis tools and data mining tools

Under tools, cBARBEL provides the NCBI BLAST server, which the user can do the sequence similarity search. Also, it provides the domain search function, which can link to the domain search server (`http://smart.embl-heidelberg.de/`), The user can find the conserved domain region comparing other species. Users also can do the sequence alignment

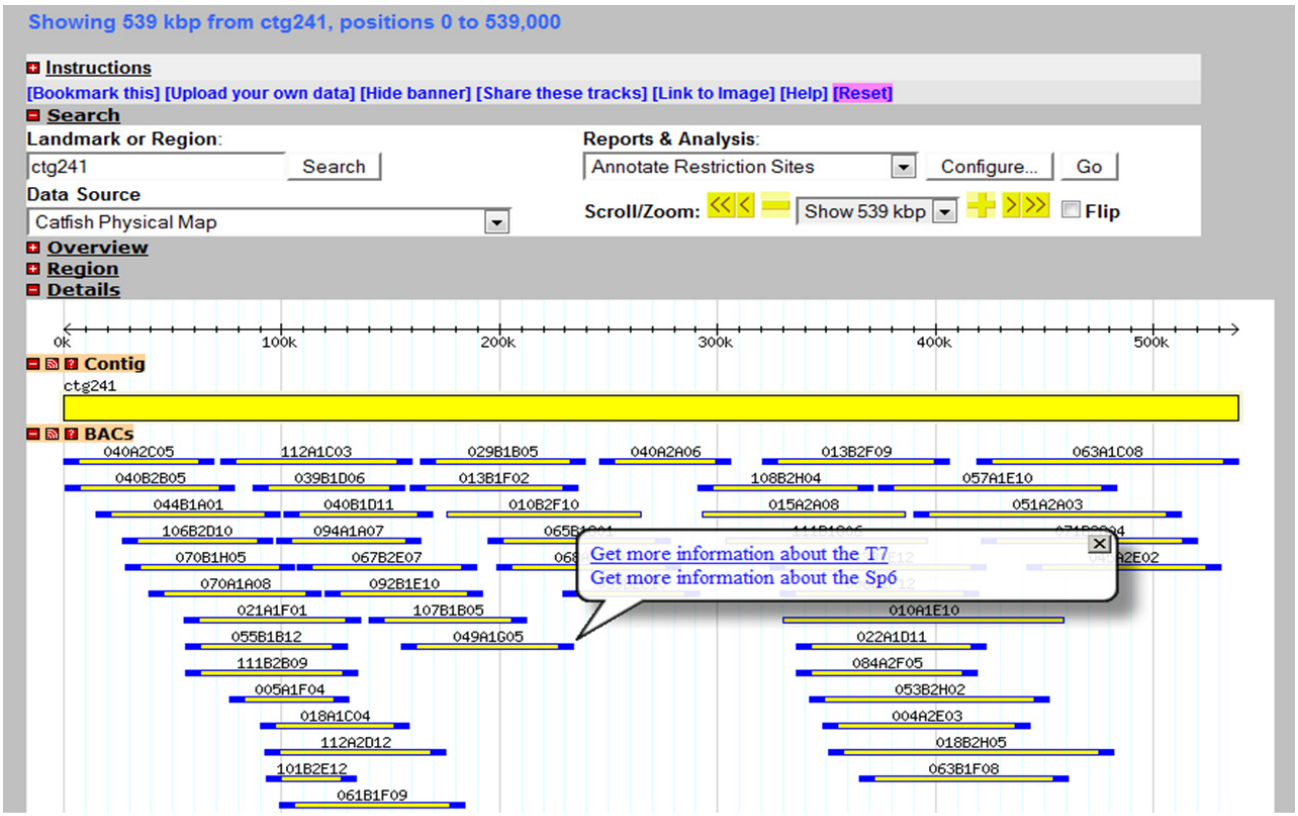


Figure 5.6: Catfish physical map. This physical map presents the catfish BACs aligned to the corresponding contig sequence. It also provides the BES links to NCBI server. The blue ends represent the BES

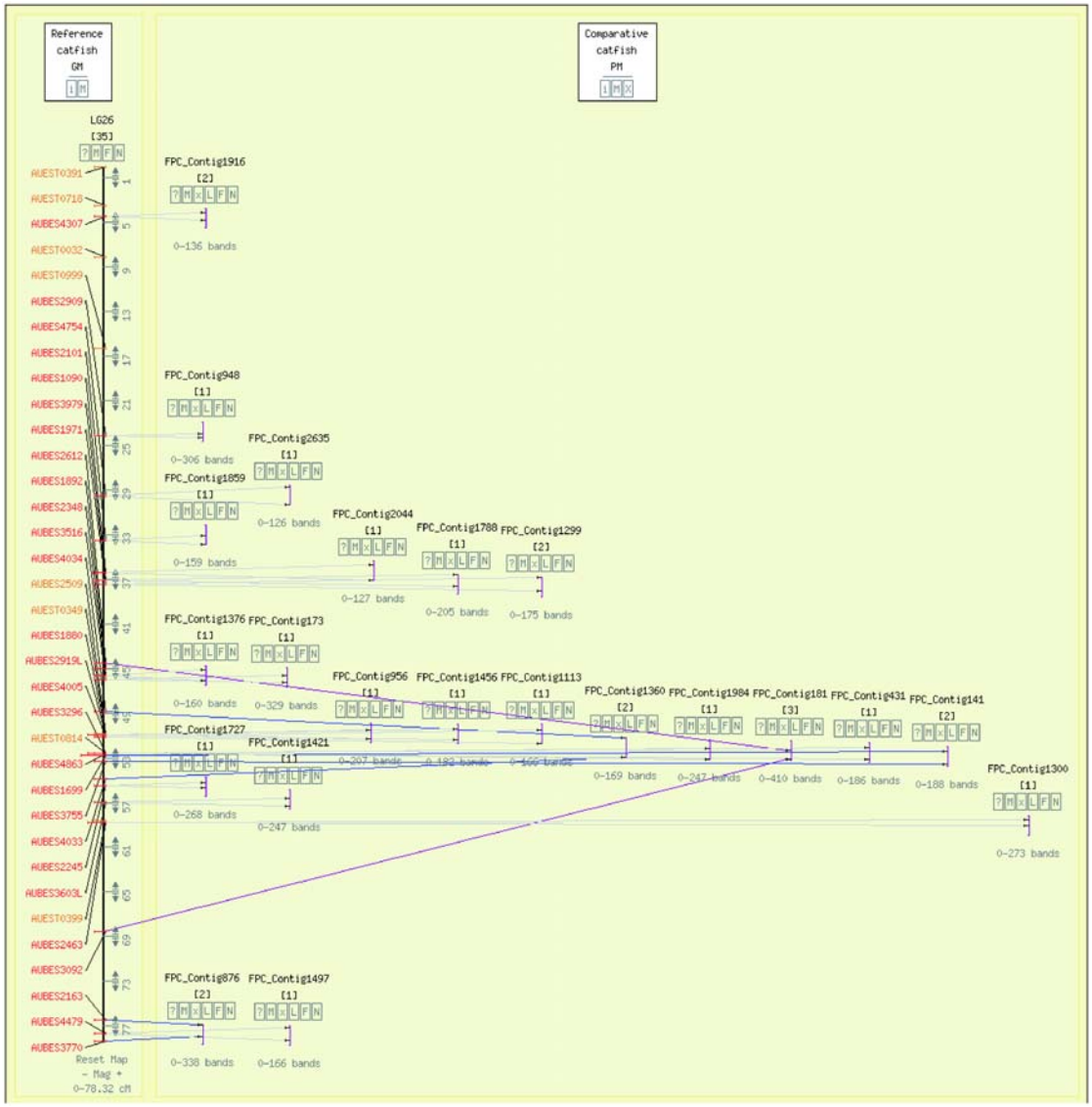


Figure 5.7: CMap-based visualization of early integration of catfish linkage and physical maps through mapping of BES microsatellites. Catfish linkage group 26 (LG26) is provided as an example on the left, and BES-associated contigs are shown on the right. CMap allows numerous options for viewing and comparing the catfish linkage and physical maps.

analysis using the alignment tool, which links to the clustalw server (http://www.ebi.ac.uk/tools/clustalw2/index.html). Moreover, the cBARBEL tool subsection provides the basic translation server, which links to the DNA translation server (http://www.vivo.colostate.edu/molkit/translate/). Along with these convenient tools, user can analysis the genetic and genomic sequence more efficient and also can get different level analysis output according user's demand.

5.5 Conclusion

cBARBEL is intended to be a central database for catfish researchers. In addition to the sequence search function, catfish map, GBrowse function, and tools described above, a number of community resources are also available either as the data or the links to other sites. These include other catfish websites and research group, a list of upcoming meeting related to catfish researchers. All of the data in cBARBEL is free available. User can contact with Auburn cBARBEL project team or email the correspond author to request specific subset of the data.

cBARBEL is continuously updated to include new available data. These new data types are incorporated and linked to the existing data when appropriate. Some new data will soon be added to cBARBEL include Catfish Next generation sequence data generated by 454 and Illumina sequencer company.

Chapter 6

Conclusion

Alternative splicing (AS) is a mechanism by which the coding diversity of the genome can be greatly increased. Rates of AS are known to vary according to the complexity of eukaryotic species, potentially explaining the tremendous phenotypic diversity among species with similar numbers of coding genes. Little is known, however, about the nature or rate of AS in teleost fish. We report here the characteristics of AS in teleost fish and classification and frequency of five canonical AS types. We conducted both same-species and cross-species analysis utilizing the Genome Mapping and Alignment Program (GMAP) and an AS pipeline (ASpipe) to study AS in four genome-enabled species (*Danio rerio*, *Oryzias latipes*, *Gasterosteus aculeatus*, *Takifugu rubripes*) and one species lacking a complete genome sequence, Ictalurus punctatus. AS frequency was lowest in the highly duplicated genome of zebrafish (17% of mapped genes). The compact genome of the pufferfish showed the highest occurrence of AS (43% of mapped genes). An inverse correlation between AS frequency and genome size was consistent across all analyzed species. Cross-species comparisons utilizing zebrafish as the reference genome allowed the identification of additional putative AS genes not revealed by zebrafish transcripts. Approximately 50% of AS genes identified by same-species comparisons were shared among two or more species. A searchable website, the Teleost Alternative Splicing Database, was created to allow easy identification and visualization of AS transcripts in the studied teleost genomes. Our results and associated database should further our understanding of alternative splicing as an important functional and evolutionary mechanism in the genomes of teleost fish.

Gene duplication has had a major impact on genome evolution. Localized (or tandem) duplication resulting from unequal crossing over and whole genome duplication are believed

to be the two dominant mechanisms contributing to vertebrate genome evolution. While much scrutiny has been directed toward discerning patterns indicative of whole-genome duplication events in teleost species, less attention has been paid to the continuous nature of gene duplications and their impact on the size, gene content, functional diversity, and overall architecture of teleost genomes. Here, using a Markov clustering algorithm directed approach we catalogue and analyze patterns of gene duplication in the four model teleost species with chromosomal coordinates: zebrafish, medaka, stickleback, and Tetraodon. Our analyses based on set size, duplication type, synonymous substitution rate (Ks), and gene ontology emphasize shared and lineage-specific patterns of genome evolution via gene duplication. Most strikingly, our analyses highlight the extraordinary duplication and retention rate of recent duplicates in zebrafish and their likely role in the structural and functional expansion of the zebrafish genome. We find that the zebrafish genome is remarkable in its large number of duplicated genes, small duplicate set size, biased Ks distribution toward minimal mutational divergence, and proportion of tandem and intra-chromosomal duplicates when compared with the other teleost model genomes. The observed gene duplication patterns have played significant roles in shaping the architecture of teleost genomes and appear to have contributed to the recent functional diversification and divergence of important physiological processes in zebrafish. We have analyzed gene duplication patterns and duplication types among the available teleost genomes and found that a large number of genes were tandemly and intrachromosomally duplicated, suggesting their origin of independent and continuous duplication. This is particularly true for the zebrafish genome. Further analysis of the duplicated gene sets indicated that a significant portion of duplicated genes in the zebrafish genome were of recent, lineage-specific duplication events. Most strikingly, a subset of duplicated genes is enriched among the recently duplicated genes involved in immune or sensory response pathways. Such findings demonstrated the significance of continuous gene duplication as well as that of whole genome duplication in the course of genome evolution.

Next generation sequence is undoubtedly the most important technology in biology, especially for the non-model species. Next generation sequence technological advances have dramatically improved sequencing throughput and quality. Like Illuminas Genome Analyzer produces a significant larger volume of sequence data than traditional sanger sequencing. Compared to just a few years ago, it is now much easier and cheaper to sequence entire genomes. Because of the rapid improvements in cost and quality of sequencing data, de novo sequencing and assembly is possible not only in large sequencing centers, but also in small labs. However, when you get the genome sequence information from sequencing company, there are several questions will be prompted immediately. The first question is which alignment algorithm or which assembler you want to choose for your sequencing project? The second question is how to optimize your alignment algorithm based on both high speed and high accuracy. To answer these two questions, this project addressed the Message Passing Interface (MPI) version assembler software, MPI-Velvet. It can process high coverage data sets and quickly reconstruct the underlying sequences.

The catfish genome database, cBARBEL(abbreviated from catfish Breeder And Researcher Bioinformatics Entry Location) is an online open-access database for genome biology of ictalurid catfish (*Ictalurus spp.*). It serves as a comprehensive, integrative platform for all aspects of catfish genetics, genomics and related data resources. cBARBEL provides BLAST-based, fuzzy and specific search functions, visualization of catfish linkage, physical and integrated maps, a catfish EST contig viewer with SNP information overlay, and GBrowse-based organization of catfish genomic data based on sequence similarity with zebrafish chromosomes. Subsections of the database are tightly related, allowing a user with a sequence or search string of interest to navigate seamlessly from one area to another. As catfish genome sequencing proceeds and ongoing quantitative trait loci (QTL) projects bear fruit, cBARBEL will allow rapid data integration and dissemination within the catfish research community and to interested stakeholders. cBARBEL can be accessed at `http://catfishgenome.org`.

Bibliography

[1] C. Adessi, G. Matton, G. Ayala, G. Turcatti, J. J. Mermod, P. Mayer, and E. Kawashima. Solid phase DNA amplification: characterisation of primer attachment and amplification mechanisms. *Nucleic Acids Res.*, 28(20):E87, Oct 2000.

[2] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25:3389–3402, Sep 1997.

[3] C. Amadou, R. M. Younger, S. Sims, L. H. Matthews, J. Rogers, A. Kumanovics, A. Ziegler, S. Beck, and K. F. Lindahl. Co-duplication of olfactory receptor and MHC class I genes in the mouse major histocompatibility complex. *Hum. Mol. Genet.*, 12:3025–3040, Nov 2003.

[4] A. Amores, J. Catchen, A. Ferrara, Q. Fontenot, and J. H. Postlethwait. Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. *Genetics*, 188(4):799–808, Aug 2011.

[5] A. Amores, A. Force, Y. L. Yan, L. Joly, C. Amemiya, A. Fritz, R. K. Ho, J. Langeland, V. Prince, Y. L. Wang, M. Westerfield, M. Ekker, and J. H. Postlethwait. Zebrafish hox clusters and vertebrate genome evolution. *Science*, 282:1711–1714, Nov 1998.

[6] S. Aparicio, J. Chapman, E. Stupka, N. Putnam, J. M. Chia, P. Dehal, A. Christoffels, S. Rash, S. Hoon, A. Smit, M. D. Gelpke, J. Roach, T. Oh, I. Y. Ho, M. Wong, C. Detter, F. Verhoef, P. Predki, A. Tay, S. Lucas, P. Richardson, S. F. Smith, M. S. Clark, Y. J. Edwards, N. Doggett, A. Zharkikh, S. V. Tavtigian, D. Pruss, M. Barnstead, C. Evans, H. Baden, J. Powell, G. Glusman, L. Rowen, L. Hood, Y. H. Tan, G. Elgar, T. Hawkins, B. Venkatesh, D. Rokhsar, and S. Brenner. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science*, 297(5585):1301–1310, Aug 2002.

[7] P. Baoprasertkul, E. Peatman, B. Somridhivej, and Z. Liu. Toll-like receptor 3 and TICAM genes in catfish: species-specific expression profiles following infection with Edwardsiella ictaluri. *Immunogenetics*, 58(10):817–830, Oct 2006.

[8] S. Batzoglou, D. B. Jaffe, K. Stanley, J. Butler, S. Gnerre, E. Mauceli, B. Berger, J. P. Mesirov, and E. S. Lander. ARACHNE: a whole-genome shotgun assembler. *Genome Res.*, 12(1):177–189, Jan 2002.

[9] S. Batzoglou, D. B. Jaffe, K. Stanley, J. Butler, S. Gnerre, E. Mauceli, B. Berger, J. P. Mesirov, and E. S. Lander. ARACHNE: a whole-genome shotgun assembler. *Genome Res.*, 12:177–189, Jan 2002.

[10] D. L. Black. Mechanisms of alternative pre-messenger RNA splicing. *Annu. Rev. Biochem.*, 72:291–336, 2003.

[11] T. Blomme, K. Vandepoele, S. De Bodt, C. Simillion, S. Maere, and Y. Van de Peer. The gain and loss of genes during 600 million years of vertebrate evolution. *Genome Biol.*, 7:R43, 2006.

[12] N. Blow. DNA sequencing: generation next-next. *Nature*, 5:267–274, 2008.

[13] J. Bohme and K. Hogstrand. Timing and effects of template number for gene conversion of major histocompatibility complex genes in the mouse. *Hereditas*, 127:11–18, 1997.

[14] S. Boue, I. Letunic, and P. Bork. Alternative splicing and evolution. *Bioessays*, 25(11):1031–1034, Nov 2003.

[15] I. Braasch, M. Schartl, and J. N. Volff. Evolution of pigment synthesis pathways by gene and genome duplication in fish. *BMC Evol. Biol.*, 7:74, 2007.

[16] D. Brett, J. Hanke, G. Lehmann, S. Haase, S. Delbruck, S. Krueger, J. Reich, and P. Bork. EST comparison indicates 38 of human mRNAs contain possible alternative splice forms. *FEBS Lett.*, 474(1):83–86, May 2000.

[17] Ron Brightwell, David S. Greenberg, Brian J. Matt, and George I. Davida. Barriers to creating a secure mpi, 1997.

[18] K. H. Brown, K. P. Dobrinski, A. S. Lee, O. Gokcumen, R. E. Mills, X. Shi, W. W. Chong, J. Y. Chen, P. Yoo, S. David, S. M. Peterson, T. Raj, K. W. Choy, B. E. Stranger, R. E. Williamson, L. I. Zon, J. L. Freeman, and C. Lee. Extensive genetic diversity and substructuring among zebrafish strains revealed through copy number variant analysis. *Proc. Natl. Acad. Sci. U.S.A.*, 109(2):529–534, Jan 2012.

[19] D. W. Bryant, W. K. Wong, and T. C. Mockler. QSRA: a quality-value guided de novo short read assembler. *BMC Bioinformatics*, 10:69, 2009.

[20] J. Butler, I. MacCallum, M. Kleber, I. A. Shlyakhter, M. K. Belmonte, E. S. Lander, C. Nusbaum, and D. B. Jaffe. ALLPATHS: de novo assembly of whole-genome shotgun microreads. *Genome Res.*, 18:810–820, May 2008.

[21] D. Cao, A. Kocabas, Z. Ju, A. Karsi, P. Li, A. Patterson, and Z. Liu. Transcriptome of channel catfish (Ictalurus punctatus): initial analysis of genes and expression profiles of the head kidney. *Anim. Genet.*, 32(4):169–188, Aug 2001.

[22] M. Chaisson, P. Pevzner, and H. Tang. Fragment assembly with short reads. *Bioinformatics*, 20:2067–2074, Sep 2004.

[23] M. J. Chaisson and P. A. Pevzner. Short read fragment assembly of bacterial genomes. *Genome Res.*, 18:324–330, Feb 2008.

[24] F. Chen, Y. Lee, Y. Jiang, S. Wang, E. Peatman, J. Abernathy, H. Liu, S. Liu, H. Kucuktas, C. Ke, and Z. Liu. Identification and characterization of full-length cDNAs in channel catfish (Ictalurus punctatus) and blue catfish (Ictalurus furcatus). *PLoS ONE*, 5(7):e11546, 2010.

[25] F. C. Chen, C. J. Chen, J. Y. Ho, and T. J. Chuang. Identification and evolutionary analysis of novel exons and alternative splicing events using cross-species EST-to-genome comparisons in human, mouse and rat. *BMC Bioinformatics*, 7:136, 2006.

[26] M. Chen, Z. Peng, and S. He. Olfactory receptor gene family evolution in stickleback and medaka fishes. *Sci China Life Sci*, 53(2):257–266, Feb 2010.

[27] B. Chevreux, T. Pfisterer, B. Drescher, A. J. Driesel, W. E. Muller, T. Wetter, and S. Suhai. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.*, 14:1147–1159, Jun 2004.

[28] A. Christoffels, E. G. Koh, J. M. Chia, S. Brenner, S. Aparicio, and B. Venkatesh. Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Mol. Biol. Evol.*, 21:1146–1151, Jun 2004.

[29] K. D. Crow, P. F. Stadler, V. J. Lynch, C. Amemiya, and G. P. Wagner. The "fish-specific" Hox cluster duplication is coincident with the origin of teleosts. *Mol. Biol. Evol.*, 23:121–136, Jan 2006.

[30] P. Dehal and J. L. Boore. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.*, 3:e314, Oct 2005.

[31] G. Denisov, B. Walenz, A. L. Halpern, J. Miller, N. Axelrod, S. Levy, and G. Sutton. Consensus generation and variant detection by Celera Assembler. *Bioinformatics*, 24:1035–1040, Apr 2008.

[32] J. C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing. *Genome Res.*, 17:1697–1706, Nov 2007.

[33] D. Dressman, H. Yan, G. Traverso, K. W. Kinzler, and B. Vogelstein. Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc. Natl. Acad. Sci. U.S.A.*, 100(15):8817–8822, Jul 2003.

[34] E. E. Eichler and D. Sankoff. Structural dynamics of eukaryotic chromosome evolution. *Science*, 301:793–797, Aug 2003.

[35] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner,

K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner. Real-time DNA sequencing from single polymerase molecules. *Science*, 323(5910):133–138, Jan 2009.

[36] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. Dewinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner. Real-time DNA sequencing from single polymerase molecules. *Science*, 323:133–138, Jan 2009.

[37] M. Fedurco, A. Romieu, S. Williams, I. Lawrence, and G. Turcatti. BTA, a novel reagent for DNA attachment on glass and efficient generation of solid-phase amplified DNA colonies. *Nucleic Acids Res.*, 34(3):e22, 2006.

[38] A. Force, M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151:1531–1545, Apr 1999.

[39] R. Friedman and A. L. Hughes. The temporal distribution of gene duplication events in a set of highly conserved human gene families. *Mol. Biol. Evol.*, 20:154–161, Jan 2003.

[40] R. Grabner, F. Mietke, and W. Rehm. Implementing an mpich-2 channel device over vapi on infiniband. In *Parallel and Distributed Processing Symposium, 2004. Proceedings. 18th International*, pages 184–, April 2004.

[41] Ren?? Grabner, Frank Mietke, and Wolfgang Rehm. Implementing an mpich-2 channel device over vapi on infiniband. *Parallel and Distributed Processing Symposium, International*, 9:184a, 2004.

[42] B. R. Graveley. Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.*, 17(2):100–107, Feb 2001.

[43] William Gropp, Ewing Lusk, Nathan Doss, and Anthony Skjellum. A high-performance, portable implementation of the mpi message passing interface standard. *Parallel Comput.*, 22(6):789–828, 1996.

[44] William Gropp, Ewing Lusk, and Anthony Skjellum. *Using MPI: Portable Parallel Programming with the Message-Passing Interface*. MIT Press, Cambridge, MA, 1994.

[45] William Gropp, Ewing Lusk, and Anthony Skjellum. *Using MPI: Portable Parallel Programming with the Message Passing Interface,* 2nd edition. MIT Press, Cambridge, MA, 1999.

[46] T. D. Harris, P. R. Buzby, H. Babcock, E. Beer, J. Bowers, I. Braslavsky, M. Causey, J. Colonell, J. Dimeo, J. W. Efcavitch, E. Giladi, J. Gill, J. Healy, M. Jarosz, D. Lapen, K. Moulton, S. R. Quake, K. Steinmann, E. Thayer, A. Tyurina, R. Ward, H. Weiss, and Z. Xie. Single-molecule DNA sequencing of a viral genome. *Science*, 320(5872):106–109, Apr 2008.

[47] C. He, L. Chen, M. Simmons, P. Li, S. Kim, and Z. J. Liu. Putative SNP discovery in interspecific hybrids of catfish by comparative EST analysis. *Anim. Genet.*, 34(6):445–448, Dec 2003.

[48] D. Hernandez, P. Francois, L. Farinelli, M. Osteras, and J. Schrenzel. De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Res.*, 18:802–809, May 2008.

[49] M. Hiller, K. Huse, M. Platzer, and R. Backofen. Creation and disruption of protein features by alternative splicing – a novel mechanism to modulate function. *Genome Biol.*, 6(7):R58, 2005.

[50] P. W. Holland, J. Garcia-Fernandez, N. A. Williams, and A. Sidow. Gene duplications and the origins of vertebrate development. *Dev. Suppl.*, pages 125–133, 1994.

[51] Y. Hou and S. Lin. Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for dinoflagellate genomes. *PLoS ONE*, 4:e6978, 2009.

[52] J. N. Housby and E. M. Southern. Fidelity of DNA ligation: a novel experimental approach based on the polymerisation of libraries of oligonucleotides. *Nucleic Acids Res.*, 26(18):4259–4266, Sep 1998.

[53] N. A. Hukriede, L. Joly, M. Tsang, J. Miles, P. Tellis, J. A. Epstein, W. B. Barbazuk, F. N. Li, B. Paw, J. H. Postlethwait, T. J. Hudson, L. I. Zon, J. D. McPherson, M. Chevrette, I. B. Dawid, S. L. Johnson, and M. Ekker. Radiation hybrid mapping of the zebrafish genome. *Proc. Natl. Acad. Sci. U.S.A.*, 96(17):9745–9750, Aug 1999.

[54] M. Hurles. Gene duplication: the genomic trade in spare parts. *PLoS Biol.*, 2:E206, Jul 2004.

[55] D. B. Jaffe, J. Butler, S. Gnerre, E. Mauceli, K. Lindblad-Toh, J. P. Mesirov, M. C. Zody, and E. S. Lander. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.*, 13(1):91–96, Jan 2003.

[56] O. Jaillon, J. M. Aury, F. Brunet, J. L. Petit, N. Stange-Thomann, E. Mauceli, L. Bouneau, C. Fischer, C. Ozouf-Costaz, A. Bernot, S. Nicaud, D. Jaffe, S. Fisher, G. Lutfalla, C. Dossat, B. Segurens, C. Dasilva, M. Salanoubat, M. Levy, N. Boudet,

S. Castellano, V. Anthouard, C. Jubin, V. Castelli, M. Katinka, B. Vacherie, C. Biemont, Z. Skalli, L. Cattolico, J. Poulain, V. De Berardinis, C. Cruaud, S. Duprat, P. Brottier, J. P. Coutanceau, J. Gouzy, G. Parra, G. Lardier, C. Chapple, K. J. McKernan, P. McEwan, S. Bosak, M. Kellis, J. N. Volff, R. Guigo, M. C. Zody, J. Mesirov, K. Lindblad-Toh, B. Birren, C. Nusbaum, D. Kahn, M. Robinson-Rechavi, V. Laudet, V. Schachter, F. Quetier, W. Saurin, C. Scarpelli, P. Wincker, E. S. Lander, J. Weissenbach, and H. Roest Crollius. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, 431:946–957, Oct 2004.

[57] W. R. Jeck, J. A. Reinhardt, D. A. Baltrus, M. T. Hickenbotham, V. Magrini, E. R. Mardis, J. L. Dangl, and C. D. Jones. Extending assembly of short DNA sequences to handle error. *Bioinformatics*, 23:2942–2944, Nov 2007.

[58] Jianzhi and Zhang. Evolution by gene duplication: an update. *Trends in Ecology and Evolution*, 18(6):292 – 298, 2003.

[59] Z. Ju, A. Karsi, A. Kocabas, A. Patterson, P. Li, D. Cao, R. Dunham, and Z. Liu. Transcriptome analysis of channel catfish (Ictalurus punctatus): genes and expression profile from the brain. *Gene*, 261(2):373–382, Dec 2000.

[60] A. Karsi, D. Cao, P. Li, A. Patterson, A. Kocabas, J. Feng, Z. Ju, K. D. Mickett, and Z. Liu. Transcriptome analysis of channel catfish (Ictalurus punctatus): initial analysis of gene expression and microsatellite-containing cDNAs in the skin. *Gene*, 285(1-2):157–168, Feb 2002.

[61] M. Kasahara, K. Naruse, S. Sasaki, Y. Nakatani, W. Qu, B. Ahsan, T. Yamada, Y. Nagayasu, K. Doi, Y. Kasai, T. Jindo, D. Kobayashi, A. Shimada, A. Toyoda, Y. Kuroki, A. Fujiyama, T. Sasaki, A. Shimizu, S. Asakawa, N. Shimizu, S. Hashimoto, J. Yang, Y. Lee, K. Matsushima, S. Sugano, M. Sakaizumi, T. Narita, K. Ohishi, S. Haga, F. Ohta, H. Nomoto, K. Nogata, T. Morishita, T. Endo, T. Shin-I, H. Takeda, S. Morishita, and Y. Kohara. The medaka draft genome and insights into vertebrate genome evolution. *Nature*, 447(7145):714–719, Jun 2007.

[62] K. S. Kassahn, V. T. Dang, S. J. Wilkins, A. C. Perkins, and M. A. Ragan. Evolution of gene function and regulatory control after whole-genome duplication: comparative analyses in vertebrates. *Genome Res.*, 19(8):1404–1418, Aug 2009.

[63] N. Kim, A. V. Alekseyenko, M. Roy, and C. Lee. The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species. *Nucleic Acids Res.*, 35(Database issue):D93–98, Jan 2007.

[64] M. Kimura. *The neutral theory of molecular evolution.* Cambridge University Press, Cambridge, London, 1983.

[65] D. J. Kliebenstein. A role for gene duplication and natural variation of gene expression in the evolution of metabolism. *PLoS ONE*, 3:e1838, 2008.

[66] A. M. Kocabas, P. Li, D. Cao, A. Karsi, C. He, A. Patterson, Z. Ju, R. A. Dunham, and Z. Liu. Expression profile of the channel catfish spleen: analysis of genes involved in immune functions. *Mar. Biotechnol.*, 4(6):526–536, Dec 2002.

[67] H. Kucuktas, S. Wang, P. Li, C. He, P. Xu, Z. Sha, H. Liu, Y. Jiang, P. Baoprasertkul, B. Somridhivej, Y. Wang, J. Abernathy, X. Guo, L. Liu, W. Muir, and Z. Liu. Construction of genetic linkage maps and comparative genome analysis of catfish using gene-associated markers. *Genetics*, 181(4):1649–1660, Apr 2009.

[68] E. S. Lander and M. S. Waterman. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, 2(3):231–239, Apr 1988.

[69] A. P. Lee, S. Y. Kerk, Y. Y. Tan, S. Brenner, and B. Venkatesh. Ancient vertebrate conserved noncoding elements have been evolving rapidly in teleost fishes. *Mol. Biol. Evol.*, 28:1205–1215, Mar 2011.

[70] P. Li, E. Peatman, S. Wang, J. Feng, C. He, P. Baoprasertkul, P. Xu, H. Kucuktas, S. Nandi, B. Somridhivej, J. Serapion, M. Simmons, C. Turan, L. Liu, W. Muir, R. Dunham, Y. Brady, J. Grizzle, and Z. Liu. Towards the ictalurid catfish transcriptome: generation and analysis of 31,215 catfish ESTs. *BMC Genomics*, 8:177, 2007.

[71] R. Li, C. Yu, Y. Li, T. W. Lam, S. M. Yiu, K. Kristiansen, and J. Wang. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25:1966–1967, Aug 2009.

[72] K. J. Lipinski, J. C. Farslow, K. A. Fitzpatrick, M. Lynch, V. Katju, and U. Bergthorsson. High spontaneous rate of gene duplication in Caenorhabditis elegans. *Curr. Biol.*, 21:306–310, Feb 2011.

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

[74] J. Liu, W. Jiang, P. Wyckoff, D.K. Panda, D. Ashton, D. Buntinas, W. Gropp, and B. Toonen. Design and implementation of mpich2 over infiniband with rdma support. In *Parallel and Distributed Processing Symposium, 2004. Proceedings. 18th International*, pages 16–, April 2004.

[75] S. Liu, Z. Zhou, J. Lu, F. Sun, S. Wang, H. Liu, Y. Jiang, H. Kucuktas, L. Kaltenboeck, E. Peatman, and Z. Liu. Generation of genome-scale gene-associated SNPs in catfish for the construction of a high-density SNP array. *BMC Genomics*, 12:53, 2011.

[76] Z. Liu. A review of catfish genomics: progress and perspectives. *Comp. Funct. Genomics*, 4(2):259–265, 2003.

[77] Z. Liu, A. Karsi, P. Li, D. Cao, and R. Dunham. An AFLP-based genetic linkage map of channel catfish (Ictalurus punctatus) constructed by using an interspecific hybrid resource family. *Genetics*, 165(2):687–694, Oct 2003.

[78] Y. H. Loh, A. Christoffels, S. Brenner, W. Hunziker, and B. Venkatesh. Extensive expansion of the claudin gene family in the teleost fish, Fugu rubripes. *Genome Res.*, 14(7):1248–1257, Jul 2004.

[79] J. Lu, E. Peatman, W. Wang, Q. Yang, J. Abernathy, S. Wang, H. Kucuktas, and Z. Liu. Alternative splicing in teleost fish genomes: same-species and cross-species analysis and comparisons. *Mol. Genet. Genomics*, 283:531–539, Jun 2010.

[80] Ewing Lusk, Nathan Doss, and Anthony Skjellum. A high-performance, portable implementation of the mpi message passing interface standard. *Parallel Computing*, 22:789–828, 1996.

[81] M. Lynch and J. S. Conery. The evolutionary fate and consequences of duplicate genes. *Science*, 290:1151–1155, Nov 2000.

[82] I. Maccallum, D. Przybylski, S. Gnerre, J. Burton, I. Shlyakhter, A. Gnirke, J. Malek, K. McKernan, S. Ranade, T. P. Shea, L. Williams, S. Young, C. Nusbaum, and D. B. Jaffe. ALLPATHS 2: small genomes assembled accurately and with high continuity from short paired reads. *Genome Biol.*, 10:R103, 2009.

[83] T. Maniatis and B. Tasic. Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature*, 418(6894):236–243, Jul 2002.

[84] E. Mardis. The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24(3):133–141, February 2008.

[85] Elaine R Mardis. The impact of next-generation sequencing technology on genetics. *Trends Genet.*, 24(3):133–41, 2008.

[86] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, Sep 2005.

[87] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W.

Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437:376–380, Sep 2005.

[88] A. McLysaght, K. Hokamp, and K. H. Wolfe. Extensive genomic duplication during early chordate evolution. *Nat. Genet.*, 31:200–204, Jun 2002.

[89] M. L. Metzker. Sequencing technologies - the next generation. *Nat. Rev. Genet.*, 11(1):31–46, Jan 2010.

[90] A. Meyer and Y. Van de Peer. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *Bioessays*, 27:937–945, Sep 2005.

[91] J. R. Miller, S. Koren, and G. Sutton. Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6):315–327, Jun 2010.

[92] J. R. Miller, S. Koren, and G. Sutton. Assembly algorithms for next-generation sequencing data. *Genomics*, Mar 2010.

[93] A. A. Mironov, J. W. Fickett, and M. S. Gelfand. Frequent alternative splicing of human genes. *Genome Res.*, 9(12):1288–1293, Dec 1999.

[94] B. Modrek, A. Resch, C. Grasso, and C. Lee. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.*, 29(13):2850–2859, Jul 2001.

[95] H. K. Moghadam, M. M. Ferguson, and R. G. Danzmann. Whole genome duplication: challenges and considerations associated with sequence orthology assignment in Salmoninae. *J. Fish Biol.*, 79(3):561–574, Sep 2011.

[96] E. W. Myers. Toward simplifying and accurately formulating fragment assembly. *J. Comput. Biol.*, 2(2):275–290, 1995.

[97] E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. Reinert, K. A. Remington, E. L. Anson, R. A. Bolanos, H. H. Chou, C. M. Jordan, A. L. Halpern, S. Lonardi, E. M. Beasley, R. C. Brandon, L. Chen, P. J. Dunn, Z. Lai, Y. Liang, D. R. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G. M. Rubin, M. D. Adams, and J. C. Venter. A whole-genome assembly of Drosophila. *Science*, 287(5461):2196–2204, Mar 2000.

[98] P. Navratilova, D. Fredman, B. Lenhard, and T. S. Becker. Regulatory divergence of the duplicated chromosomal loci sox11a/b by subpartitioning and sequence evolution of enhancers in zebrafish. *Mol. Genet. Genomics*, 283(2):171–184, Feb 2010.

[99] M. Nei and T. Gojobori. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.*, 3:418–426, Sep 1986.

[100] P. Nigumann, K. Redik, K. Matlik, and M. Speek. Many human genes are transcribed from the antisense promoter of L1 retrotransposon. *Genomics*, 79(5):628–634, May 2002.

[101] J. P. Noonan, J. Grimwood, J. Danke, J. Schmutz, M. Dickson, C. T. Amemiya, and R. M. Myers. Coelacanth genome sequence reveals the evolutionary history of vertebrate genes. *Genome Res.*, 14(12):2397–2405, Dec 2004.

[102] T Ohta. Gene Conversion and Evolution of Gene Families: An Overview. *Genes*, 1:349–356, Sep 2010.

[103] D. Pan and L. Zhang. Tandemly arrayed genes in vertebrate genomes. *Comp. Funct. Genomics*, page 545269, 2008.

[104] J. J. Pastor, I. Lingard, G. Bhalay, and M. Bradley. Ion-extraction ladder sequencing from bead-based libraries. *J Comb Chem*, 5(2):85–90, 2003.

[105] E. Peatman and Z. Liu. CC chemokines in zebrafish: evidence for extensive intrachromosomal gene duplications. *Genomics*, 88(3):381–385, Sep 2006.

[106] E. Pettersson, J. Lundeberg, and A. Ahmadian. Generations of sequencing technologies. *Genomics*, 93:105–111, Feb 2009.

[107] P. A. Pevzner, P. A. Pevzner, H. Tang, and G. Tesler. De novo repeat classification and fragment assembly. *Genome Res.*, 14(9):1786–1796, Sep 2004.

[108] P. A. Pevzner, H. Tang, and M. S. Waterman. An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. U.S.A.*, 98(17):9748–9753, Aug 2001.

[109] A. M. Phillippy, M. C. Schatz, and M. Pop. Genome assembly forensics: finding the elusive mis-assembly. *Genome Biol.*, 9:R55, 2008.

[110] J. H. Postlethwait, Y. L. Yan, M. A. Gates, S. Horne, A. Amores, A. Brownlie, A. Donovan, E. S. Egan, A. Force, Z. Gong, C. Goutel, A. Fritz, R. Kelsh, E. Knapik, E. Liao, B. Paw, D. Ransom, A. Singer, M. Thomson, T. S. Abduljabbar, P. Yelick, D. Beier, J. S. Joly, D. Larhammar, F. Rosa, M. Westerfield, L. I. Zon, S. L. Johnson, and W. S. Talbot. Vertebrate genome evolution and the zebrafish gene map. *Nat. Genet.*, 18(4):345–349, Apr 1998.

[111] S. M. Quiniou, G. C. Waldbieser, and M. V. Duke. A first generation BAC-based physical map of the channel catfish genome. *BMC Genomics*, 8:40, 2007.

[112] V. Ravi and B. Venkatesh. Rapidly evolving fish genomes and teleost diversity. *Curr. Opin. Genet. Dev.*, 18:544–550, Dec 2008.

[113] C. Rizzon, L. Ponger, and B. S. Gaut. Striking similarities in the genomic distribution of tandemly arrayed genes in Arabidopsis and rice. *PLoS Comput. Biol.*, 2:e115, Sep 2006.

[114] M. Robinson-Rechavi, O. Marchand, H. Escriva, and V. Laudet. An ancestral whole-genome duplication may not have been responsible for the abundance of duplicated fish genes. *Curr. Biol.*, 11(12):R458–459, Jun 2001.

[115] J. M. Rothberg and J. H. Leamon. The development and impact of 454 sequencing. *Nat. Biotechnol.*, 26(10):1117–1124, Oct 2008.

[116] A. Sahoo and S. H. Im. Interleukin and interleukin receptor diversity: role of alternative splicing. *Int. Rev. Immunol.*, 29(1):77–109, 2010.

[117] M. Sammeth, S. Foissac, and R. Guigo. A general definition and nomenclature for alternative splicing events. *PLoS Comput. Biol.*, 4(8):e1000147, 2008.

[118] F. Sanger, S. Nicklen, and A. R. Coulson. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, 74:5463–5467, Dec 1977.

[119] F. Santini, L. J. Harmon, G. Carnevale, and M. E. Alfaro. Did genome duplication drive the origin of teleosts? A comparative study of diversification in ray-finned fishes. *BMC Evol. Biol.*, 9:194, 2009.

[120] J. Serapion, H. Kucuktas, J. Feng, and Z. Liu. Bioinformatic mining of type I microsatellites from expressed sequence tags of channel catfish (Ictalurus punctatus). *Mar. Biotechnol.*, 6(4):364–377, 2004.

[121] D. Sharon, G. Glusman, Y. Pilpel, M. Khen, F. Gruetzner, T. Haaf, and D. Lancet. Primate evolution of an olfactory receptor cluster: diversification by gene conversion and recent emergence of pseudogenes. *Genomics*, 61:24–36, Oct 1999.

[122] S. H. Shiu, J. K. Byrnes, R. Pan, P. Zhang, and W. H. Li. Role of positive selection in the retention of duplicate genes in mammalian genomes. *Proc. Natl. Acad. Sci. U.S.A.*, 103:2232–2236, Feb 2006.

[123] V. Shoja and L. Zhang. A roadmap of tandemly arrayed genes in the genomes of human, mouse, and rat. *Mol. Biol. Evol.*, 23:2134–2141, Nov 2006.

[124] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. Jones, and I. Birol. ABySS: a parallel assembler for short read sequence data. *Genome Res.*, 19:1117–1123, Jun 2009.

[125] L. D. Stein, C. Mungall, S. Shu, M. Caudy, M. Mangone, A. Day, E. Nickerson, J. E. Stajich, T. W. Harris, A. Arva, and S. Lewis. The generic genome browser: a building block for a model organism system database. *Genome Res.*, 12(10):1599–1610, Oct 2002.

[126] R. L. Strausberg, S. Levy, and Y. H. Rogers. Emerging DNA sequencing technologies for human genomic medicine. *Drug Discov. Today*, 13:569–577, Jul 2008.

[127] A. Sundquist, M. Ronaghi, H. Tang, P. Pevzner, and S. Batzoglou. Whole-genome sequencing and assembly with high-throughput, short-read technologies. *PLoS ONE*, 2:e484, 2007.

[128] Sayantan Sur, Matthew J. Koop, and Dhabaleswar K. Panda. High-performance and scalable mpi over infiniband with reduced memory usage: an in-depth performance analysis. In *Proceedings of the 2006 ACM/IEEE conference on Supercomputing*, SC '06, New York, NY, USA, 2006. ACM.

[129] M. Suyama, D. Torrents, and P. Bork. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.*, 34:W609–612, Jul 2006.

[130] H. Tang, X. Wang, J. E. Bowers, R. Ming, M. Alam, and A. H. Paterson. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.*, 18:1944–1954, Dec 2008.

[131] J. S. Taylor, Y. Van de Peer, and A. Meyer. Revisiting recent challenges to the ancient fish-specific genome duplication hypothesis. *Curr. Biol.*, 11:R1005–1008, Dec 2001.

[132] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22:4673–4680, Nov 1994.

[133] L. M. van der Aa, J. P. Levraud, M. Yahmi, E. Lauret, V. Briolat, P. Herbomel, A. Benmansour, and P. Boudinot. A large new subset of TRIM genes highly diversified by duplication and positive selection in teleost fish. *BMC Biol.*, 7:7, 2009.

[134] K. Vandepoele, W. De Vos, J. S. Taylor, A. Meyer, and Y. Van de Peer. Major events in the genome evolution of vertebrates: paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc. Natl. Acad. Sci. U.S.A.*, 101:1638–1643, Feb 2004.

[135] A. E. Vinogradov. Genome size and GC-percent in vertebrates as determined by flow cytometry: the triangular relationship. *Cytometry*, 31(2):100–109, Feb 1998.

[136] G. C. Waldbieser, B. G. Bosworth, D. J. Nonneman, and W. R. Wolters. A microsatellite-based genetic linkage map for channel catfish, Ictalurus punctatus. *Genetics*, 158(2):727–734, Jun 2001.

[137] B. B. Wang, M. O'Toole, V. Brendel, and N. D. Young. Cross-species EST alignments reveal novel and conserved alternative splicing events in legumes. *BMC Plant Biol.*, 8:17, 2008.

[138] S. Wang, E. Peatman, J. Abernathy, G. Waldbieser, E. Lindquist, P. Richardson, S. Lucas, M. Wang, P. Li, J. Thimmapuram, L. Liu, D. Vullaganti, H. Kucuktas, C. Murdock, B. C. Small, M. Wilson, H. Liu, Y. Jiang, Y. Lee, F. Chen, J. Lu, W. Wang, P. Xu, B. Somridhivej, P. Baoprasertkul, J. Quilang, Z. Sha, B. Bao, Y. Wang, Q. Wang, T. Takano, S. Nandi, S. Liu, L. Wong, L. Kaltenboeck, S. Quiniou, E. Bengten, N. Miller, J. Trant, D. Rokhsar, Z. Liu, J. Ainsworth, I. Altinok, C. R. Arias, J. A. Bader, A. L. Bilodeau, C. Bird, J. Bogerd, B. G. Bosworth, R. C. Bruch,

K. Burnett, J. T. Caprio, J. Chappell, N. Chatakondi, G. Chinchar, W. W. Dickhoff, R. T. DiGiulio, C. Duan, M. V. Duke, R. A. Dunham, S. Gabel, T. A. Giambernardi, W. L. Gray, E. D. Green, L. A. Hanson, M. Hardman, C. He, J. Hikima, A. Hutson, L. Jaso-Friedmann, Z. Ju, A. Karsi, K. Kelley, D. Kingsley, C. Kleinholz, P. H. Klesius, A. Kocabas, W. K. Lee, M. Lennard, W. Litaker, G. W. Litman, C. J. Lobb, G. Luker, B. G. Magor, T. J. McConnel, W. Muir, E. Noga, K. Nusbaum, D. D. Ourth, V. Panangala, R. Patino, B. C. Peterson, R. Phelps, K. P. Plant, J. H. Postlethwait, H. E. Quintero, D. Rodriguez, H. L. Saunders, B. Scheffler, T. Schwedler, R. A. Shelby, W. Simc, C. A. Shoemaker, L. Tang, J. Terhune, R. L. Thune, T. R. Tiersch, G. W. Warr, T. Welker, M. Westerfield, K. L. Willett, K. Williams, R. Winn, C. Wu, D. Xu, R. Yant, H. Y. Yeh, Y. Zohar, and J. Zou. Assembly of 500,000 inter-specific catfish expressed sequence tags and large scale gene-associated marker development for whole genome association studies. *Genome Biol.*, 11(1):R8, 2010.

[139] S. Wang, Z. Sha, T. S. Sonstegard, H. Liu, P. Xu, B. Somridhivej, E. Peatman, H. Kucuktas, and Z. Liu. Quality assessment parameters for EST-derived SNPs from catfish. *BMC Genomics*, 9:450, 2008.

[140] R. L. Warren, G. G. Sutton, S. J. Jones, and R. A. Holt. Assembling millions of short DNA sequences using SSAKE. *Bioinformatics*, 23:500–501, Feb 2007.

[141] N. Whiteford, N. Haslam, G. Weber, A. Prugel-Bennett, J. W. Essex, P. L. Roach, M. Bradley, and C. Neylon. An analysis of the feasibility of short read sequencing. *Nucleic Acids Res.*, 33:e171, 2005.

[142] C. Winkler, M. Schafer, J. Duschl, M. Schartl, and J. N. Volff. Functional divergence of two zebrafish midkine growth factors following fish-specific gene duplication. *Genome Res.*, 13(6A):1067–1081, Jun 2003.

[143] JC. Wootton and S. Federhen. Statistics of local complexity in amino acid sequences and sequence databases. *Computer and Chemistry.*, 17:149–163, Jun 1993.

[144] T. D. Wu and C. K. Watanabe. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9):1859–1875, May 2005.

[145] Y. Xing and C. Lee. Alternative splicing and RNA selection pressure–evolutionary consequences for eukaryotic genomes. *Nat. Rev. Genet.*, 7(7):499–509, Jul 2006.

[146] P. Xu, S. Wang, L. Liu, E. Peatman, B. Somridhivej, J. Thimmapuram, G. Gong, and Z. Liu. Channel catfish BAC-end sequences for marker development and assessment of syntenic conservation with other fish species. *Anim. Genet.*, 37(4):321–326, Aug 2006.

[147] P. Xu, S. Wang, L. Liu, J. Thorsen, H. Kucuktas, and Z. Liu. A BAC-based physical map of the channel catfish genome. *Genomics*, 90(3):380–388, Sep 2007.

[148] Z. Yang. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, 13:555–556, Oct 1997.

[149] K. Youens-Clark, B. Faga, I. V. Yap, L. Stein, and D. Ware. CMap 1.01: a comparative mapping application for the Internet. *Bioinformatics*, 25(22):3040–3042, Nov 2009.

[150] Y. Yuan, J. D. Chung, X. Fu, V. E. Johnson, P. Ranjan, S. L. Booth, S. A. Harding, and C. J. Tsai. Alternative splicing and gene duplication differentially shaped the regulation of isochorismate synthase in Populus and Arabidopsis. *Proc. Natl. Acad. Sci. U.S.A.*, 106(51):22020–22025, Dec 2009.

[151] D. R. Zerbino and E. Birney. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, 18(5):821–829, May 2008.

[152] D. R. Zerbino, G. K. McEwen, E. H. Margulies, and E. Birney. Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler. *PLoS ONE*, 4:e8407, 2009.

[153] X. Zhou, Q. Li, H. Lu, H. Chen, Y. Guo, H. Cheng, and R. Zhou. Fish specific duplication of Dmrt2: characterization of zebrafish Dmrt2b. *Biochimie*, 90:878–887, Jun 2008.