Genome-wide comparative analysis of channel catfish (*Ictalurus punctatus*)

by

Yanliang Jiang

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
May 4, 2013

Keywords: Catfish, comparative analysis, whole genome, comparative map, BAC end sequences, physical map contig-specific sequences

Approved by

Zhanjiang(John) Liu, Chair,Professor of Fisheries and Allied Aquacultures
Yolanda Brady, Associate Professor of Fisheries and Allied Aquacultures
Eric Peatman, Assistant Professor of Fisheries and Allied Aquacultures
Nannan Liu, Professor of Entomology and Plant Pathology

Abstract

Comparative map is a powerful tool to transfer the genomic information from well-studied model species to non-model species whose genomic resources are limited. Channel catfish is a primary aquaculture species in United States. To better understand the genetic basis of catfish, and to improve genetic breeding and selection, plenty of genomic resources have been developed in catfish. In this dissertation study, over 40,000 BAC end sequences were generated using traditional Sanger sequencing technology. Moreover, millions of physical map contig-specific sequences were generated by using next-generation sequencing technology. Utilizing both of these two valuable genomic resources along with other existed genomic resources including genetic linkage map, BAC-based physical map, and the catfish draft genome sequences, the first genome-wide comparative analysis between catfish and zebrafish was conducted in this study.

Acknowledgments

I would like to express my deep and sincere gratitude to my major professor, Dr. Zhanjiang (John) Liu, for his guidance throughout my degree program. I also would like to express my gratitude to all of my committee members and my outside reader for their time and directions of my dissertation work. I would like to extend my appreciation to many people who helped me during my graduate education including Hong Liu, Ping Li, Huseyin Kucuktas, Ludmilla Kaltenboeck, Shaolin Wang, Parichart Ninwichian, Zhenxia Sha, Jason Abernathy,Yue Chen, Shikai Liu, Jianguo Lu, Wenqi Wang, Zunchun Zhou, Fei Chen, Hao Zhang, Donghong Niu, Tingting Feng, Fanyue Sun, Yu Zhang, Jiaren Zhang, Chao Li and Ruijia Wang. Finally, I would like to thank my parents and all my family members for their love, understanding and support.

Table of Contents

List of Tables

List of Figures

I. INTRODUCTION AND LITERATURE REVIEW

Catfish (Order *Siluriformes*), named for the predominant barbels, is a diverse group of ray-finned fish, representing more than 3,000 species, 478 genera and 36 families (Ferraris and dePinna 1999). Catfish are commonly found in freshwaters, especially abundant in South America, Africa and Asia (Bruton 1996). The general body shape of catfish is sub-cylindrical, with no scales. Most catfish are bottom-feeders. In general, they sink rather than float due to a reduced gas bladder and a heavy, bony head (Bruton 1996).

Channel catfish, *Ictalurus punctatus*, is one of most important commercial aquaculture species in the United States, accounting for 60% of all aquaculture production. It originally ranged from the Gulf States and the Mississippi Valley north to the Great Lakes and Prairie provinces of Canada (Tucker 1985). Of the Ictalurid catfishes, the channel catfish is the only spotted North American catfish with a deeply forked tail. Generally, channel catfish are irregularly spotted on the sides, but the spots will disappear in adult fish. As an important commercial species, channel catfish possess desirable qualities, such as being quite easy to spawn, satisfactory growth, relatively high feed conversion, crowding tolerance, adapting well in many commonly used culture systems and the ability of survive in a wide range of environments (Tucker 1985).

Genomic resources

As a new branch of science, genome studies, including the structural genomics and functional genomics, is trying to better understand the genetic basis of an organism, such as the linear order of the nucleotide bases, the expression of all genes as a dynamic system, how the genes interact and influence biological pathway, and the evolution role of the organism. In recent years, the high throughput sequencing technologies have drastically improved the genome studies. A lot of genomic resources have been developed in catfish:

1. Expressed Sequence Tags (ESTs)

ESTs, short sub-sequence of cDNA, are obtained by single-pass sequencing of random cDNA clones. They are partial cDNA sequences corresponding to mRNAs generated from randomly selected cDNA clones. Efficient EST analysis required the construction of normalized cDNA library. Normalized cDNA library is the cDNA library that has been equalized in representation to increase the representation of rarely expressed genes and to reduce the representation of abundantly expressed genes (Soares et al. 1994).

The primary function of ESTs is to rapidly identify genes. A small collection of ESTs in a species without any genome information can result in the identification of a large number of genes (Liu 2007). ESTs provide a foundation for full-length cDNA analysis (Chen et al. 2010), and the identification of genetic markers such as microsatellite markers and SNP markers (Wang et al. 2010). Moreover, ESTs serve as a resource for high-density microarray platform, alternative splicing and differential

polyadenylation studies, and comparative mapping. Large-scale EST projects have been carried out in several teleost species, such as rainbow trout (Rexroad et al. 2003), Fugu (Clark et al. 2003), and salmon (Rise et al. 2004). So far, nearly 500,000 ESTs from catfish have been generated, which will greatly benefit the catfish introgression breeding program and whole genome association studies (Li et al. 2007; Wang et al. 2010).

Moreover, ESTs are good anchors for comparative mapping because of the easy assessment of the homology of particular sequences among species. Especially, the microsatellite markers associated with ESTs are favorable for comparative mapping, because they are located near genes, as well as their flank sequences are highly conserved.

2. Full-length cDNA sequences

Full-length cDNAs are a useful tool for correct annotation and clustering of the genomic sequences. Moreover, full-length cDNA serve as an important resource to analyze genome structure and genome function (Satoh et al. 2007; Hurowitz et al. 2007), to produce gene models that establish accurate exon-intron boundaries (Hayashizaki, 2003; Castelli et al. 2004), to predict protein sequences (Harhay et al. 2005) to provide vital information about alternative splice forms of gene products (Xin et al. 2008) and aid in discrimination between alternative splicing and gene duplication and pseudogenes (Harhay et al. 2005). Comparing full-length cDNAs to genome provides insights into evolution and gene regulation. Several studies in agriculture species have produced full-length cDNA sets: Harhay et al. (2005) have generated a total of 954 bovine full-length cDNA sequences to create predicted bovine protein sequences to support

bovine genome assembly and functional genomic studies; a total of 560 Atlantic salmon full-length cDNAs have been generated by Andreassen et al. (2009) to correct annotation and clustering of a forthcoming whole genome sequence; Chen et al. (2010) generated a total of 1,767 full-length cDNAs from several catfish cDNA libraries, to serve as a important resource for the ongoing catfish whole genome sequencing and gene-based studies of function and evolution in teleost fishes.

3. BAC-based physical map

Physical map, a map showing physical locations on a DNA molecular, represents the linear order of genes or DNA fragments containing the order of overlapping cloned DNA fragments, usually obtained by restriction endonuclease digestion (Liu, 1998). BAC clones are the suitable building blocks for constructing a physical map. The most frequent approach to construct the BAC-based physical map is DNA fingerprinting, using one-enzyme and agarose gel system (Marra et al. 1997), or the more efficient multiple-enzyme with fluorescence-labeling system (Luo et al. 2003). The purpose of fingerprinting is to identify the overlapped clones so that they can be grouped into one contig.

The BAC-based physical map, which consists of a set of ordered overlapped BAC clones, is important for the understanding of genome structure and organization, and for position-based cloning of economically important genes (Liu 2007). It provides the basis not only for clone-by-clone whole genome sequencing, but also for the orientation and scaffolding of next-generation-sequencing-based whole genome sequencing assembly. Because of the importance to genome study, a physical map has been constructed in

many species including aquaculture species, such as Atlantic salmon (Ng et al. 2005), Nile tilapia (Katagiri et al. 2005), rainbow trout (Palti et al. 2009), common carp (Xu et al. 2011), Asian sea bass (Xia et al. 2010) and Zhikong scallop (Zhang et al. 2011). Xu et al. (2007) constructed a BAC-based physical map of channel catfish genome, using 34, 580 BAC clones. A total of 3307 contigs were assembled with an E-value cutoff of 1e-20. The combined contig size for all contigs was 0.965 Gb, approximately the genome size of channel catfish (Xu et al. 2007).

4. BAC end sequences (BESs)

BAC library, a large-insert genomic library made by using a specific vector called bacterial artificial chromosome, is usually used for clone-by-clone whole genome sequencing project and physical mapping. BESs are often generated by direct sequencing of BAC clones using sequencing primers designed based on the vector sequences near the border of the insert, usually using SP6 and T7 sequencing primers. The inserts from BAC clone are usually prepared by partial restriction digest of genomic DNA, with average size of 100 ~ 200 kb. Assuming the restriction sites are randomly distributed, then the BESs represents random genomic sequences (Liu 2007).

For those species without whole genome sequences, BESs are important genomic resources which can be used for many purposes, such as virtual mapping genes to BAC-based fingerprint physical map, assessing the repeat structure of the genome, mining microsatellite markers. In addition, BESs can be exploited for comparative analysis, such as identification of evolutionarily conserved syntenies, characterization of

genome composition and architecture. Fujiyama et al. (2002) constructed a first-generation of a human-chimpanzee comparative map by aligning chimpanzee BESs to human genome sequences, which allowed for putative orthologs to be identified. A similar approach was extensively used in mammals for construction of comparative maps, such as human-mouse (Gregory et al. 2002), human-cattle (Larkin et al. 2003), human-horse (Leeb et al. 2006), human-porcine (Meyers et al. 2005) comparative maps. Moreover, the conserved synteny with the sequenced genome can improve the genome assembly of species whose genome is not available yet by detecting errors, correcting inverted segments, and filling gaps (Zimin et al. 2009). So far, over 60,000 BESs have been generated from channel catfish BAC libraries (Xu et al. 2006; Liu et al. 2009), providing a vital genome resource for comparative genome mapping, as well as scaffolding of whole genome sequencing assembly.

5. Genetic linkage map

A genetic linkage map is a linear arrangement of genes or genetic markers on chromosomes, representing the tendency of certain alleles or loci to be inherited together (Liu 1998). A linkage map is a powerful tool in structural genome studies, functional genome studies, evolutionary studies, and especially QTL mapping for marker associated selection (MAS) in many species. It is created based on the frequencies of recombination of the genetic markers during meiosis. The higher recombination rate between two markers, the farther apart they are supposed to be. A genetic marker is a fragment of DNA associated with a certain location on the genome that can be used to measure the genome variations. The commonly used genetic markers are: allozymes,

which are the various forms of a enzyme that encoded by different alleles at the same locus; mitochondria DNA (mtDNA); Restriction Fragment Length Polymorphism (RFLP), which is based on the differences of DNA fragment length after digestion of the DNA sample by restriction enzyme; Random Amplified Polymorphism DNA (RAPD), which is a PCR-based multi-loci DNA fingerprint technology; Amplified Fragment Length Polymorphism (AFLP), which combines the advantages of RFLP and RAPD and selectively amplifies a set of genomic restriction fragments using PCR; Microsatellites which are based on the variation of copy number of repeats; and Single Nucleotide Polymorphism (SNP) which is the most abundant form of polymorphism in the genome caused by point mutation at a certain locus .

Microsatellite markers are considered as most informative markers because of the high heterozygosity and polymorphism information content (PIC) value. Microsatellites can be type I markers if they are developed from ESTs, which make them important markers because they are associated with genes with known functions. Kucuktas et al. (2009) constructed a catfish linkage map using EST-derived microsatellites and SNPs marker, which mapped a total of 331 markers to 29 linkage groups with an average marker interval of 6.0 cM. However, designing microsatellite primers from ESTs needs to take into consideration many situations such as the presence of introns, the exon-intron boundaries, and UTR regions. Microsatellites are mostly type II markers if they are developed from BESs or other genomic DNA sequences, because of more than 90% represented in non-coding region. In catfish, a high resolution linkage map has been constructed by Ninwichian et al. (2012) using BES-derived microsatellite marker.

A total of 2099 BES-derived microsatellite markers have been mapped into 29 linkage groups, with an average marker interval of 1.4 cM.

Next-generation sequencing and assembly

In recent years, the development and progress of high throughput sequencing technologies, also referred to as “next-generation” or “third-generation” sequencing, including the 454 sequencing technology, the Illumina technology, ABI SOLiD sequencing technology and PacBio strobe sequencing technology, has significantly improved sequencing throughput and quality, and reduced costs per base. The 454 sequencing technology yields longer reads, on average, than the Illumina sequencing reads, which offers advantage for *de novo* assembly. However, 454 reads have a higher indel error rate (Margulies et al. 2005) when compared to Illumina reads. On the other hand, although Illumina reads are much shorter, they provide a higher depth of coverage compared to 454 reads. Several studies demonstrated that using the complementary nature of both 454 and Illumina reads may take full advantage of those two types of data for *de novo* assembly and overcome each other’s shortcomings (Reinhardt et al.2009; Aury et al. 2008).

The next-generation sequencing technologies have greatly revolutionized the genomic research, allowing whole genome sequencing of many organisms, including prokaryotes and eukaryotes feasible. Several prokaryotic genomes were fully sequenced utilizing the next-generation sequencing technologies, e.g. a rice pathogen, *Pseudomonas sringae* genome was completely sequenced and its *de novo* assembly is completed using low-coverage 454 and Illumina reads (Reinhardt et al. 2009). Similarly,

Nagarajan et al. (2010) assembled a complete microbial *Geobactersul furreducens* genome using both 454 and Illumina sequencing technology.

In contrast to prokaryotes, the *de novo* assembly in eukaryotes is particularly difficult due to the short reads generated and the complexity of the large eukaryotic genomes. Several *de novo* assembly software packages have been developed for assembling genomes in parallel to emerging next-generation sequencing platforms. Two basic graph based algorithms exist for assemblers: First one is based on overlap-layout-consensus graphs, such as Newbler (Margulies et al. 2005), MIRA (Chevreux et al. 1999). In these packages, the algorithm computes all pair-wise overlaps between reads to build an overlap graph. Then, the overlap graph is used to compute a layout of reads and consensus sequence of contigs. The second algorithm is based on de-Bruijn graphs, such as the algorithm used in Velvet (Zerbino et al. 2008), ABySS (Simpson et al. 2009), and many other packages in which the reads are broken into smaller sequences of DNA, called as K-mers, where K denotes the length of these sequences bases (Miller et al. 2010). The de-Bruijn graph is built based on the overlaps of length K-1 between these K-mers other than the actual reads. The assembly algorithms and their implementations become typically complex with large volumes of sequence data.

The next-generation sequencing technology has greatly improved the structural genomic and functional genomic studies in catfish. Several transcriptome sequencings have been conducted, for instance, for single nucleotide polymorphisms (SNP) marker mining (Liu et al. 2011) and for transciptomic response analysis after bacterial infection

(Li et al. 2012; Sun et al. 2012). The catfish whole-genome sequencing project and related projects are either ongoing or done (Jiang et al. 2011; unpublished data), using next-generation sequencing technology.

Comparative genomics

Comparative genomics is a study focused on the relationship of genome structure and function across different biological species. It provides a powerful way to transfer the genomic information from one genome-sequences-available species to the species whose whole genome sequence data is not yet available, and allows identifying the potential syntenies and therefore to understand the genome organization and how the genome is remodeled during evolution. Comparative genomics exploits both differences and similarities in regulatory regions, mRNA, and genes among different species to understand how nature selection acts upon those elements, the phylogenetic relationship, and the mechanism of genome evolution. The DNA sequences that are responsible for the similarities between different species are conserved through time, while the DNA sequences that are responsible for the differences will be divergent (Hardison 2003).

Based on the foundation, there are two types of comparative analyses (Liu 2007). One is based on the orthologous genes, called clusters of orthologous groups, in which, proteins from all completed genomes are compared. Both orthologs and paralogs will be identified and clustered into groups. Another one is based on the alignment of the genomes to identify the syntenies. Comparing the genome sequences from different species can reveal the rearrangement event. Comparative genomics has been significantly advanced in recent years and become increasingly valuable, because whole

genome sequences become available for more and more species, and more genomic resources have been developed which can be used for comparative analysis.

Although available genomic resources are scarce for non-model fish species, great efforts have been made in comparative genomics. For instance, Shimizu et al. (2006) sequenced medaka one chromosome, Linkage group 22, and compared it with fugu genome, which revealed unique features of the medaka genome. A pilot comparative genomic analysis of European sea bass with five model teleosts was performed, and revealed a high level of evolutionarily conserved regions (Chistiakov et al. 2008). Several comparative analyses related to studies on catfish have been reported, based on the linkage map using EST-derived microsatellite markers (Kucuktas et al. 2009), BESs (Liu et al. 2009), or the next-generation-sequencing data spanning around 1 Mb genome region (Jiang et al. 2011). However, these studies only provide either framework of comparative analysis or on a small portion of genomic region. In this dissertation, the comparative genomics analysis between channel catfish and other teleosts has been conducted on a whole-genome scale, using all available catfish BESs, physical map contig-specific sequences, as well as the draft whole genome sequences.

II. RESEARCH OBJECTIVES

The first objective of this dissertation work is to generate more catfish genomic resources, including BAC end sequences (BES) and physical map contig-specific sequences. The physical map contig in this dissertation refers to the contig in the BAC-based physical map, while the genome contig refers to the contig generated by *de novo* assembly of whole genome sequencing. The BESs allows mining of microsatellite markers, which can be used for the construction of genetic linkage map. Both BESs and the physical map contig-specific sequence can serve as sequence tag to anchor the genome contigs, group them into the linkage map, and be used for the comparative analysis.

The second objective is to conduct the genome-wide comparative analysis of catfish with zebrafish. Utilizing the BESs and physical map contig-specific sequences, along with all other available catfish genomic resources including the BAC-based physical map, genetic linkage map, and the draft whole genome sequences, this study will identify the homologous genes in zebrafish, identify the conserved microsyntenies, identify the homologous chromosome in zebrafish corresponding to each catfish linkage group, and construct the comparative map.

III. GENERATION OF CATFISH BAC END SEQUENCES

Abstract

Comparative mapping is a powerful tool to transfer genomic information from sequenced genomes to closely related species for which whole genome sequence data are not yet available. However, such an approach is still very limited in catfish, the most important aquaculture species in the United States. This project was initiated to generate additional BAC end sequences and to demonstrate their applications in comparative mapping in catfish. In this study, 43,000 BAC end sequences were generated and used for comparative genome analysis in catfish. Using these and the additional 20,000 existing BAC end sequences as a resource along with linkage mapping and an existing physical map, conserved syntenic regions were identified between the catfish and zebrafish genomes. A total of 10,943 catfish BAC end sequences (17.3%) had significant BLAST hits to the zebrafish genome (cutoff value ≤ e-5), of which 3,221 were unique gene hits, providing a platform for comparative mapping based on locations of these genes in catfish and zebrafish. Genetic linkage mapping of microsatellites associated with contigs allowed identification of large conserved genomic segments and construction of super scaffolds. BAC end sequences and their associated polymorphic markers are great resources for comparative genome analysis in catfish. Highly conserved chromosomal regions were identified to exist between catfish and zebrafish. However, it appears that the level of conservation at local genomic regions is high while

a high level of chromosomal shuffling and rearrangements exist between catfish and zebrafish genomes. Orthologous regions established through comparative analysis should facilitate both structural and functional genome analysis in catfish.

Introduction

Comparative mapping is a powerful tool to transfer genomic information from sequenced genomes to closely related species for which whole genome sequence data are not yet available. Such an approach was initially demonstrated by Fujiyama *et al*. (Fujiyama et al. 2002) for the construction of the human-chimpanzee comparative map. In these closely related primate species, approximately 98% chimpanzee BAC end sequences (BES) had significant BLAST hits to the human genome sequence allowing putative orthologues to be identified (Fujiyama et al. 2002). A similar approach was used for the construction of the human-mouse comparative map (Gregory et al. 2002). Subsequently, this approach was extensively used in mammals including construction of the human-cattle, the human-horse, and the human-porcine comparative maps (Larkin et al. 2003; Leeb et al. 2006; Meyers et al. 2005). Most recently, this approach was utilized one step further for the construction of the comparative genome contig (CGC)-based physical map of the sheep genome (Dalrymple et al. 2007), where CGC is established based on anchorage of the sheep BES onto the genome sequences of dog, cow, and human. These successes depended on a high percentage of BLAST hits and/or high levels of genome collinearity.

Five teleost fish genomes have been fully sequenced (http://www.ensembl.org/index.html) including zebrafish (*Danio rerio*, from the order Cypriniformes), Japanese pufferfish (*Fugu rubripes*, from the order Tetraodontiformes), green spotted pufferfish (*Tetraodon nigroviridis*, from the order Tetraodontiformes), medaka (*Oryzias latipes*, from the order Beloniformes), and three-spined stickleback (*Gasterosteus aculeatus*, from the order Gasterosteiformes), while whole genome sequencing is also underway for tilapia (http://www.cichidgenome.org; http://www.broad.mit.edu/science/projects/mammals-models/vertebrates-invertebrates/tilapia/tilapia-genome-sequencing-project/). The availability of these whole genome sequences lends great opportunities for comparative genome analysis. Recently, major genomic resources have been developed from a number of fish species such as Atlantic salmon (*Salmo salar*) (Moen et al. 2008; Ng et al. 2005; Rise et al. 2004), rainbow trout (*Oncorhynchus mykiss*) (Guyomard et al. 2006; Rexroad et al. 2003), tilapia (Katagiri et al. 2005; Lee et al. 2005), gilthead sea bream (*Sparusauratus*) (Franch et al. 2006; Sarropoulou et al. 2005; Sarropoulou et al. 2007; Senger et al. 2006), and European sea bass (*Dicentrarchus labrax*) (Chistiakov et al. 2005; Whitaker et al. 2006), and channel catfish (*Ictalurus punctatus*) (for a review, see Liu 2003, 2008).

Catfish is the major aquaculture species in the United States. It is one of the six species included in the US National Animal Genome Project NRSP-8. A number of genomic resources have been developed in catfish including a large number of molecular markers (He et al. 2003; Serapion et al. 2004; Somridhivej et al. 2008; Xu et

al. 2006), genetic linkage maps (Kucuktas et al. 2009; Liu et al. 2003; Waldbieser et al. 2001), several hundred thousands of ESTs (Cao et al. 2001; Ju et al. 2000; Karsi et al. 2002; Kocabas et al. 2002; Li et al. 2007; wang et al. 2010), microarray platforms (Ju et al. 2002; Li and Waldbieser 2006; Liu et al. 2008; Peatman et al. 2007, 2008), BAC libraries (Quinious et al. 2003; Wang et al. 2007), and BAC-based physical maps (Quinious et al. 2007; Xu et al. 2007). To enable BAC end sequence-based comparative genome analysis, we previously reported generation of 20,366 BES in catfish (Xu et al. 2006). In spite of the great value of those BES for the characterization of genome repeat structures (Nandi et al. 2007) and for the identification of microsatellite markers, our previous comparative genome analysis using BES revealed very limited conservation between the catfish and zebrafish genomes. Of the 141 mate-paired BES with genes on both ends of the BAC inserts, only 34 (24.1%) were found in nearby genomic locations in the zebrafish genome, suggesting high levels of chromosomal rearrangements (Wang et al. 2007). Such findings were in strong contrast to the situations found between medaka-sea bream, *Tetraodon*-sea bream, medaka-stickleback, *Tetraodon*-medaka, stickleback-sea bream, *Tetraodon*-stickleback genome comparisons where almost complete genome collinearities were found (Sarropoulou et al. 2008). We speculated that our earlier inability to discover a large amount of genome collinearity between catfish and zebrafish could be a result of the low numbers of BES and the lack of a physical map. Therefore, in this study, we extended our efforts in BAC end sequencing and generated additional 43,021 BES, bringing the total to 63,387 (25,676 mate-paired). Using these catfish BES and the BAC contig-based

physical map (Xu et al. 2007), genetic linkage mapping of BAC end-anchored microsatellites, and the genome sequence of zebrafish, here we conducted extensive comparative genome analysis. We report the identification of conserved syntenies and demonstrate the construction of super scaffolds of contigs by genetic linkage mapping of BAC end-associated microsatellites.

Results and discussions

BAC end sequencing

As shown in Table 1, a total of 42,240 BAC inserts (6.13X clone-coverage of the channel catfish genome) were sequenced from both ends, resulting in 63,387 BES≥ 200 bp in length (75% overall success rate), including 20,366 BES we previously reported (Xu et al. 2006). Mate-paired BESs were produced from 25,676 BAC clones, while only a single BES was obtained from 12,035 clones. The BES were of high quality as the Q20 length ranged from 200 to 810 bp, with an average Q20 read length of 596 bp. All these BES have been deposited into the GenBank GSS database with consecutive accession numbers of [GenBank: FI857756-FI900776]. A total of 37,784,877 bp of genomic sequences was generated from this study, representing approximately 4% of the catfish genome. Analysis using the 37,784,877 bp BES resulted in 11.91% of base pairs masked using the *Danio* repeat database, with the most abundant type of repeat being the DNA transposons. We previously reported the assessment of repetitive elements in the catfish genome and the additional 43,021 BES generated in this study confirmed our

previous findings in general (Xu et al. 2006). These BES [GenBank: DX083364-DX103729] were also used for comparative genome analysis in this study. A BLASTN search against zebrafish genome resulted in a total of 10,943 (17.3%) significant hits at a cutoff value of e^{-5} (Table 2).

***In silico* analysis of the BAC-associated catfish genes on the zebrafish genome**

TBLASTX searches using the 63,387 catfish BES against the ENSEMBL zebrafish cDNA database with chromosome information resulted in 5,066 significant hits (Table 3). Of the 5,066 significant hits, 2,197 unique zebrafish genes were hit by a single BAC end sequences while 1,024 unique zebrafish genes were hit by two or more catfish BAC end sequences, making a total of 3,221 unique zebrafish genes with significant hits from the catfish BAC end sequences. The 3,221 genes cover all 25 zebrafish chromosomes, with the largest number of gene hits being located on chromosome 5 (224 significant hits), followed by chromosome 7 (191 significant hits), chromosome 20 (171 significant hits), chromosome 6 (151 significant hits) and chromosome 19 (134 significant hits); and the smallest number of gene hits on chromosome 24 with 78 hits (Table 3). The number of gene hits on various chromosomes was approximately proportional to the sizes of the zebrafish chromosomes with some exceptions. When the size of chromosomes was taken into consideration, chromosome 25 had the largest number of gene hits with 3.5 hits per Mb or one hit per 286 kb on average, followed by chromosome 5, 4, 20, 19, and 22 with 3.2, 3.1, 3.0, 2.9, and 2.9 hits per Mb, respectively (Table 3).

One particular finding of these BLAST searches is the observation of many highly repetitive genes. Out of 3,221 unique genes, 1,024 genes had hits from two or more BES. A single gene identity had hits from as many as 31 BESs. A total of 14 genes had hits from at least 10 BES each (Table 4); an additional 139 genes had hits from 4-9 BES each; 230 genes had hits from 3 BES each, and 641 genes had hits from 2 BES each (Table 4). Some of the genes with hits from multiple BES may represent a whole array of related genes with similar functional domains. For instance, 18 BES hits NOD3-like gene of channel catfish, which was just recently characterized; NOD3 gene existed as a single copy gene in the catfish genome (Sha et al. 2009), and apparently the multiple BES contained many related genes harbor domains present within the NOD3 gene. Theoretically, a fraction of genes should have hits by more than one BES, simply because of the genome coverage of the BAC clones. We believe that overlapping (including identical) BAC clones do account for some of the observed hits of genes by more than one BES (data not shown), especially for those with 2-3 BES hits. However, the mathematical chances do not support multiple BES hits of a single gene unless the gene itself is repetitive in the catfish genome. Additional research is warranted to fully understand the nature of these genes/sequences in the catfish genome, but clearly many of these represent classes of repetitive gene families such as DNA polymerase gene that had hits from 31 BES.

Establishing microsyntenies

Among the teleost genomes with high sequence coverage, zebrafish is the most

closely related species to catfish (Steinke et al. 2006). Our initial BLAST searches of the catfish BES against the genome of the *Tetraodon nigroviridis* generated many fewer significant hits compared to those against the zebrafish genome. Therefore, we concentrated our comparative analysis efforts with the zebrafish genome in this study.

Conserved syntenies are most often established by comparing genome sequences of related species. However, the whole genome sequence is not yet available from catfish. In the absence of the whole genome sequence, we attempted to establish microsyntenies based on physical linkage of gene sequences. With the genomic resources available in catfish, we have taken three approaches. First, if the genes were identified from both ends of a single BAC clone, they are physically linked with a distance of the BAC clone insert size. If the same two genes are found linked in the zebrafish genome in the same genome neighborhood, a microsynteny can then be established. These genes from mate-paired BES are physically linked with the average distances between them being the average insert size of the catfish BAC library, i.e. 161 kb. From the 63,387 BES, a total of 25,676 mate-paired BES were identified. Of these, 760 mate-paired BES had significant BLASTN hits on BES from both ends against the zebrafish genome sequence. However, only 194 of the 760 significant hit pairs were on the same zebrafish chromosome, allowing syntenic comparison. Further tBLASTX searches against the ENSEMBL zebrafish cDNA database allowed identification of 95 mate-paired BES with genes on both sides. The genomic locations of these 176 mate-paired genes were determined from the zebrafish genome sequence. 52 pairs were found to be present in neighboring genomic locations within one million base pairs, while the other 43 were

present in more distant locations (> 1 Mb) on the same chromosomes. The vast majority of the 50 mate-paired genes were found to be within 500 kb on the zebrafish genome sequence; only 2 of the 50 pairs had a distance of 500-920 kb (Table 5), suggesting conserved syntenies of the involved genes.

We previously reported the relatively high levels of local region conservation. For instance, many genes within the bordering mate-paired genes were well conserved among catfish, zebrafish, and *Tetraodon*, as determined by direct sequencing of the catfish BAC DNA using primers predicted from known genes in zebrafish or Tetraodon (Wang et al. 2007). We did not extend this part of the study, but all known genome information suggested high levels of local genome conservation.

In addition to the 58 microsyntenies, we attempted to determine if significant gene hits in the same catfish BAC contigs also fall on the same chromosome locations comparable to the contig sizes. As shown in Table 3, of the contigs with gene hits, 1,754 contigs had only one gene hit, while 472 contigs had two or more gene hits within each contig. Because the genes in the same contig are physically linked, their linkage in a comparable distance in the zebrafish genome would indicate a conserved synteny. As shown in Figure 1, the vast majority of gene hits within the same contigs were found to be located on the same zebrafish chromosomes with comparable distances as estimated from the catfish BAC contigs. Using such an approach, a total of 336 conserved microsyntenies was identified (Table 3). Presence of multiple gene hits within large BAC contigs would allow identification of extended large conserved syntenic regions. Many of the microsyntenies were conserved with extended genomic distance to span

over several million base pairs (Figure 1). For instance, large conserved syntenies were identified from chromosomes 12, 13, 14, 22, 23, 24, and 25 (Figure 1). In spite of the identification of some relatively large conserved syntenic regions, the vast majority of the identified syntenies were microsyntenies. Such highly segmented microsyntenies are not very useful for genome-wide comparative analysis. However, if scaffolds can be established by determining the relationships among the microsyntenies, large-scale genome comparison should be possible. We, therefore, used two zebrafish chromosomes as the query to demonstrate if super scaffolds can be established. Chromosome 7, one of the chromosomes with highest number of significant gene hits, and chromosome 13, one of the chromosomes with a large number of contigs having two or more hits (indicative of high level of syntenic conservation), were chosen for further analysis using genetic linkage mapping.

Genetic mapping of BAC end-anchored microsatellites

In order to extend the scope of conserved microsyntenies, microsyntenies identified on zebrafish chromosomes 7 and 13 were genetically mapped to determine their chromosomal locations in the catfish genome. There were 373 significant BLASTN hits to zebrafish chromosome 13 involving 178 unique catfish BAC contigs; and 505 significant hits to zebrafish chromosome 7 involving 314 unique catfish BAC contigs. We, therefore, first identified microsatellites from these involved catfish BAC contigs, and then mapped them to the linkage groups when the microsatellites were polymorphic in the resource family. A total of 548 pairs of microsatellite primers were tested, of

which 296 from 188 contigs (the details of the polymorphic markers are shown in the Table 6) were polymorphic in the resource family. Further analysis using JoinMap 4.0 allowed mapping of 290 microsatellite markers, of which 161 microsatellites were from BES with significant similarity to zebrafish chromosome 7, and 129 microsatellites were from BES with significant similarity to zebrafish chromosome 13.

Mapping of microsatellites from contigs with hits to zebrafish chromosome 13 allowed identification of a highly conserved chromosome between catfish and zebrafish. As shown in Figure 2, of the 129 microsatellites from BES with high similarities to the zebrafish chromosome 13, 57 microsatellites from 43 contigs were mapped into a single linkage group, spanning approximately 90 centi-Morgans, suggesting the conservation of a large segment of this chromosome. However, the entire chromosome is not conserved. The 129 microsatellites were mapped into a total of 24 linkage groups, with seven of the 24 linkage groups containing 4-12 markers (Table 6).

Similarly but to a much lesser extent, microsatellites from BES with similarities to the zebrafish chromosome 7 were mapped to three major linkage groups (Figure 3). Once again, many smaller syntenic regions were mapped to various linkage groups, suggesting high levels of local conservation and low levels of chromosomal conservation. Nonetheless, the significant aspect of this is that scaffolds can be established by linking various contigs together through linkage mapping. This will allow integration of genetic linkage and physical maps once microsatellites are identified from most contigs of the physical map. Such scaffolds should guide genome sequence assembly in the future, and should also provide molecular length measurements of

various polymorphic markers along the genome of catfish, providing guidance for the development of the SNP chip technology in catfish. Apparently, SNP chips constructed from evenly distributed SNPs provide the best coverage of the catfish genome when conducting the whole genome association studies.

Genetic linkage mapping of BAC end-anchored microsatellites provided a level of validation of the physical map. Discrepancies were found between the BAC assemblage and the linkage map. Of the 75 contigs with at least two markers, 54 contigs were mapped properly into the same linkage groups. However, 18 contigs were mapped into different linkage groups (Table 6). Of these 18 contigs, 12 are large contigs with at least 40 BACs. Apparently, such discrepancy is indicative of mistakes in the BAC assemblage. Mapping additional BAC end-anchored microsatellites is under way to integrate the genetic linkage and physical maps, and to correct any additional mistakes in the assembly of the physical map (Xu et al. 2007).

Conclusions

Some highly conserved chromosomes or chromosomal regions exist between catfish and zebrafish. High levels of local conservation were found, but a high level of chromosomal shuffling and rearrangements exists between catfish and zebrafish genomes. Comparative genome analysis using zebrafish genome sequence is highly useful for regional comparisons, but not so useful at the chromosomal levels. The significance of comparative genome analysis in catfish is that it will allow more

cost-effective structural genomic analysis, but more importantly, orthologies established through comparative genome analysis should facilitate functional assignment of genes. Given that functional genomics is more difficult with non-model fish species, inference from orthologues should be one of the most efficient and reliable approaches for functional analysis of the catfish genome.

Overall, the evolutionary syntenic conservation appeared to be relatively low between the catfish genome and the genomes of the zebrafish. This indicates many chromosome breakage and rearrangements among the fish genomes occurred during evolution. These findings are consistent with our previous findings that high levels of conservation were found within small genomic regions, whereas high levels of large-scale genome reshuffling were evident when comparing the genomes of catfish and zebrafish (Kucuktas et al. 2009; Wang et al. 2007). These conclusions, however, are based on the assumption that the zebrafish genome assembly is correct. Apparently, due to the assembly mistakes in the zebrafish genome, some of the syntenic breaks may be due to the still poor assembly of the zebrafish genome. We also acknowledge that comparative genome analysis using a partial bank of sequences in catfish and a more complete databank in zebrafish could potentially lead to a bias. Caution should be exercised when establishing concrete syntenic relations. Such limitations themselves justify the need for whole genome sequencing in catfish.

Materials and Methods

BAC culture and BAC-end sequencing

The CHORI-212 Channel Catfish BAC library (http://bacpac.chori.org/library.php?id=103) was used for BAC-end sequencing. BAC culture and sequencing reactions were conducted using standard protocols, and as previously described (Xu et al. 2006; Wang et al. 2007). Briefly, BAC clones were transferred from 384-well plates to 96-well culture blocks containing 1.5 ml of 2X YT medium with 12.5 µg/ml chloramphenicol and grown at 37ºC overnight with shaking at 300 rpm. The blocks were centrifuged at 2000 x g for 10 min in an Eppendorf 5804R bench top centrifuge to collect bacteria. The culture supernatant was decanted and the blocks were inverted and tapped gently on paper towels to remove remaining liquid. BAC DNA was isolated using the Perfectprep™ BAC 96 kit (Eppendorf, Westbury, NY) according to the manufacturer's specifications. BAC DNA was collected in 96-well plates and stored at -20ºC until usage.

Sequencing of channel catfish BAC ends was conducted using the BigDye® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, Foster City, CA), with modifications. Each sequencing reaction mix contained 2 µl of 5X sequencing buffer, 2 µl of primer (3 pmol/µl), 1.5 µl BigDye v3.1 dye terminator, and 4.5 µl of BAC DNA. BAC clones were sequenced from both ends using the primers T7 (5'-TAATACGACTCACTATAGGG-3') and SP6 (5'-ATTTAGGTGACACTATAG-3'). Cycle sequencing was carried out in 96-well plate format using PTC-200 thermal cyclers (MJ Research/Bio-Rad, Hercules, CA) under the following thermal profile: an initial

denaturing at 95ºC for 5 min, followed by 100 cycles of 95ºC for 30 s, 53ºC for 10 s, and 60ºC for 4 min. Products were purified using ethanol/EDTA precipitation according to the BigDye protocol (Applied Biosystems), with the following modifications. After thermal cycling, 1 µl of 125 mM EDTA and 30 µl chilled (-80°C) 100% ethanol were added to each reaction. Plates were gently mixed and incubated at room temperature for 15 min. Plates were then centrifuged at 2,250 x g at 4ºC for 30 min, followed by washing in 30 µl of 70% ethanol at 2,000 x g for 15 min. Ethanol was decanted and 8 µl Hi-Di™ formamide (Applied Biosystems) was added to each well to re-suspend DNA. Products were denatured at 95ºC for 5 min and sequenced on a 3130*xl* genetic analyzer (Applied Biosystems).

Sequence processing and analysis

The raw BES base calling were conducted by using Phred (Ewing et al. 1998a, 1998b) with Q20 as a cut-off. Lucy program (Chou and Holmes 2001) was used to remove the vector sequences and short sequence less than 200 bp. Repeats were masked using REPEATMASKER (http://www.repeatmasker.org) before BLAST analysis. In order to anchor the catfish BES to the zebrafish genome, BLASTN searches of the repeat-masked catfish BES were conducted against the zebrafish genome sequences (Assembly 7). A cut-off value of e^{-5} was used as the significance similarity threshold for the comparison. TBLASTX searches of the repeat-masked BES were conducted against the ENSEMBLE zebrafish cDNA database.

Identification of conserved syntenies between catfish and zebrafish

In the absence of the whole genome sequence, we attempted to establish microsyntenies based on physical linkage of gene sequences. First, if the genes were identified from both sides of a single BAC clone (mate-paired BES), then they are physically linked with a distance of the BAC clone insert size. If the same two genes were found to be linked on the zebrafish genome in the same genome neighborhood, a microsynteny was established.

Initially, BESs were analyzed by BLASTN (*E*-value $\leq$ -5) for the identification of mate-pairs with significant hits on both sides of the BAC insert. Mate-paired BESs were analyzed by tBLASTX (*E*-value $\leq$ -5) for the identification of genes on both sides of the BAC insert. After identification, the two mate-paired genes in each BAC were used as queries to search for their chromosomal locations on the zebrafish genome. Conserved microsyntenies were declared when the mate-paired genes existed within a distance of 1.0 Mb within the zebrafish genome.

Syntenies were also established using genes within contiguous sequences (contigs) based on the catfish physical map (Xu et al. 2007). Genes identified from BES were located along the catfish physical map. Genes identified within the same contig and located on the same zebrafish chromosome with comparable distances as estimated from the catfish BAC contig, an extended synteny was established.

Construction of the catfish syntenic groups using linkage maps

In order to assess the scope of microsyntenies, two zebrafish chromosomes,

chromosome 7 and 13, were chosen for analysis. Chromosome 7 had the largest number of significant hits and chromosome 13 had a large number of contigs having two or more hits (suggestive of high level of syntenic conservation). Syntenies were established using microsatellite-based linkage mapping. A total of 548 microsatellite loci in the contigs which had significant BLASTN hits to the zebrafish chromosome 7 and 13 were tested using a hybrid catfish resource family, F_1-2(female blue-channel catfish hybrid) X Ch-6 (male channel catfish) with 64 progeny.

Microsatellites were identified and analyzed using Msatfinder (http://www.genomics.ceh.ac.uk/msatfinder) and Vector NTI 10.0 (Invitrogen, Carlsbad, CA) as we previously described (Somridhivej et al. 2008). Polymerase chain reaction (PCR) primers were designed using Msatfinder. Mononucleotide repeats were manually excluded. PCR amplification was conducted as previously described (Somridhivej et al. 2008). Briefly, each microsatellite PCR reaction contained 1X PCR buffer, 2 mM $MgCl_2$, 0.2 mM of each dNTP, 4 ng upper primer, 6 ng lower primer, 1 pmol labeled primer, and 0.25 U of JumpStart *Taq* polymerase (Sigma, St. Louis, MO), and 20 ng genomic DNA. PCR amplification was carried out using a touchdown program with the following thermal profile: an initial denaturation at 94ºC for 3.5 min, followed by 94ºC for 30 sec, 57ºC for 30 sec, and 72ºC for 30 sec for 20 cycles as the first step, and at 94ºC for 30 sec, 53ºC for 30 sec, and 72ºC for 30 sec for 15 cycles as the second step. A final extension was performed at 72ºC for 10 minutes. The PCR products were analyzed on 7% sequencing gels by using the 4300 DNA Analyzer (LI-COR® Biosciences, Lincoln, NE). After gel electrophoresis, loci were manually genotyped to determine allele segregation

patterns and polymorphisms in the resource family.

The catfish linkage map was constructed using JoinMap version 4.0 software as we previously described (Kucuktas et al. 2009) using the cross-pollinating (CP) coding scheme, which handles the data containing various genotype configurations with unknown linkage phases (Sekino et al. 2006). Linkage between markers was examined by estimating LOD scores for recombination rate, and map distances were calculated using the Kosambi mapping function. Significance of marker linkage was determined at a final LOD threshold of 3.0.

Table 1: A summary of BAC end sequences.

Category	Numbers
BAC sequence reactions	84,480
Total clean sequences	63,387 (75% success)
T7 sequences	32,074
SP6 sequences	31,313
Pair BAC end sequences	25,676
Total length sequenced	37,784,877 bp
Average length	596 bp

Table 2: Distribution of comparatively anchored BAC clones.

Zebrafish chromosome	Chromosome size (Mb)	No. of BLASTN hits	No. of hits per Mb	Average distance (Kb)
1	56.2	403	7.2	139
2	54.4	368	6.8	148
3	62.9	524	8.3	120
4	42.6	617	14.5	69
5	70.4	499	7.1	141
6	59.2	451	7.6	131
7	70.3	730	10.4	96
8	56.5	449	7.9	126
9	51.5	485	9.4	106
10	42.4	272	6.4	156
11	44.6	543	12.2	82
12	47.5	447	9.4	106
13	53.5	490	9.2	109
14	56.5	414	7.3	136
15	46.6	454	9.7	103
16	53.1	488	9.2	109
17	52.3	393	7.5	133
18	49.3	334	6.8	148
19	46.2	317	6.9	146
20	56.5	356	6.3	159
21	46.1	370	8.0	125
22	39.0	329	8.4	119
23	46.4	374	8.1	124
24	40.3	515	12.8	78
25	32.9	321	9.8	102
Total/Average	**1,277.2**	**10,943**	**8.7**	**120**

Table 3: Distribution of comparatively anchored BAC clones using protein encoding gene sequences only.

Zebrafish chromosome	Chromosome size (Mb)	No. of protein encoding genes*	No. of tBLASTx hits	Hits to unique genes	Unique gene hits per Mb	No. of contigs with single gene hits	No. of contigs with multiple gene hits	No. putative micro-syntenies
1	56.2	818	205	123	1.83	75	17	13
2	54.4	875	194	133	2.13	85	15	13
3	62.9	975	196	127	1.75	72	18	13
4	42.6	743	221	130	2.77	78	16	9
5	70.4	1,173	340	224	2.74	103	35	21
6	59.2	818	232	151	2.31	85	22	18
7	70.3	990	283	191	2.33	80	34	25
8	56.5	864	196	128	1.93	65	22	14
9	51.5	700	212	133	2.17	61	21	16
10	42.4	670	150	87	1.72	51	12	9
11	44.6	627	161	121	2.26	67	16	10
12	47.5	636	177	114	2.21	72	14	7
13	53.5	744	200	113	1.96	67	21	14
14	56.5	701	197	113	1.77	84	11	7
15	46.6	688	177	125	2.25	68	14	8
16	53.1	773	181	124	2.02	77	14	10
17	52.3	715	180	115	1.99	62	19	13
18	49.3	749	193	121	2.23	47	22	21
19	46.2	780	233	134	2.58	86	21	16
20	56.5	1,053	277	171	2.48	76	30	19
21	46.1	721	163	117	2.28	63	14	10
22	39.0	959	178	113	2.59	50	19	16
23	46.4	669	204	121	2.24	68	18	13
24	40.3	513	117	78	1.71	47	8	7
25	32.9	597	199	114	3.04	65	19	14
Total/Average	**1,277.2**	**19,551**	**5,066**	**3,221**	**2.21**	**1,754**	**472**	**336**

* : Annotated genes only from ENSEMBL.

Table 4: Distribution of genes with hits from multiple BAC end sequences, with details provided for genes with 10 or more hits from BAC end sequences.

No. of Genes	Putative Identities	No.of BES hits	Presence in Zebrafish genome	Potential explanation
1	Novel protein similar to DNA polymerases	31	28	Repetitive elements related to retroelements
1	Methionine aminopeptidase 1	22	2	Repetitive elements or multigene family
1	NOD3 protein-like	18	63	Common domains shared by many related proteins
1	Similar to tudor domain containing 7,hypothetical protein LOC393661	17	89	Repetitive elements or repetitive genes
1	Similar to porf2	16	81	Repetitive elements or multigene family
1	Similar to general transcription factor II-I repeat domain-containing protein 2A	16	82	Repetitive elements or multigene family
1	Similar to novel G protein-coupled receptor	13	92	Repetitive elements or multigene family
1	Similar to serine/threonine-protein kinase pim-3;	11	85	Repetitive elements or multigene family
1	Similar to novel protein from Danio rerio;	11	85	Repetitive elements or multigene family
1	Similar to Dynein heavy chain 6	11	20	Repetitive elements or multigene family
1	ORF2 [Mus musculus domesticus]	10	91	Repetitive elements or multigene family
1	PREDICTED: tubulin, alpha, ubiquitous isoform 8 [Macaca mulatta]	10	16	Repetitive elements or multigene family
1	PREDICTED: similar to vacuolar protein sorting 52 [Danio rerio]	10	69	Repetitive elements or multigene family
1	GF20795 [Drosophila ananassae]	10	4	Repetitive elements or multigene family
14		**Subtotal**		
68		5-9		Repetitive elements or multigene family
71		4		Repetitive elements or

		multigene family
139	**Subtotal**	
230	3	Potentially duplicated gene candidates
641	2	Potentially duplicated gene candidates
1024	**Total**	

Table 5: Summary of 50 conserved syntenies identified by comparison of 95 mate-paired genes of channel catfish with genomic locations of those on the zebrafish draft genome sequence.

Catfish BAC ID	GeneBank ID of SP6 hits	GeneBank ID of T7 hits	zebrafish Chr	Chr location (Mb)	Distance (bp)
035I16	NP_001038735	XP_689146	2	38.61	18,048
063L16	NP_001012377	NP_001092217	2	19.25	48,679
026B05	XP_001921249	CAI20867	3	31.82	4,418
098D05	CAI11873	NP_001038462	4	1.82	301,470
007M06	XP_001333162	NP_571105	5	5.46	379,579
028M04	XP_687685	NP_958867	5	22.23	133,063
074L07	XP_001334912	XP_687570	5	15.69	290,598
035N11	NP_001032187	NP_958882	6	18.25	231,524
040J24	XP_686613	NP_991309	6	7.37	443,524
077F04	NP_001034906	NP_001038813	6	11.42	484,877
032F20	XP_691291	NP_001075159	7	28.40	244,425
062B16	NP_001008651	NP_001017550	7	15.05	222,416
075K10	NP_956505	NP_001070187	7	13.34	268,455
103I21	NP_998033	NP_001159825	7	22.34	293,845
047N13	XP_699919	XP_699627	8	23.84	203,260
057N22	XP_001923800	NP_997066	9	36.69	163,677
076H22	NP_001004563	XP_001921910	9	13.84	313,352
105N24	NP_001122018	NP_997992.2	9	27.37	268,271
056A23	NP_998295	NP_001103164	10	7.50	486,745
068C21	NP_001019272	XP_001920077	10	17.72	104,067
093K02	XP_001344325	XP_001919728	10	41.67	152,479
096A14	NP_001120805	NP_997775	11	10.02	980,276
010B15	NP_956715	NP_956756	12	25.17	111,625
041B24	NP_001002607	XP_693784	12	13.59	398,886
109K22	XP_001920588	XP_001920550	12	25.55	449,096
018H11	NP_956915	NP_997743	13	23.23	293,181
026C08	NP_001119869	NP_956611	13	23.50	201,663
027E09	XP_001922173	XP_001921791	14	0.22	574,761
103F19	NP_571477	XP_001919973	14	22.90	256,143
022D09	NP_001096112	XP_684421	15	29.14	418,219
059B20	XP_682817	XP_691794	15	27.80	94,396
077H08	XP_001920240	NP_001070609	15	6.05	339,215
004O03	NP_001020707	NP_001020642	16	10.40	149,618
075M09	NP_001107266	NP_001082899	16	8.30	120,355
104I08	NP_571871	NP_694503	16	27.17	214,517
013P16	NP_001076304	NP_001038173	18	18.65	366,849
042J20	NP_001037796	NP_001038370	18	32.85	223,781
021L13	NP_001038343	XP_688911	19	17.40	286,019

052N18	XP_001921158	NP_956134	19	7.59	345,553
056A09	NP_001018463	NP_001038416	19	13.46	122,217
065M02	XP_001922809	NP_001038686	19	15.37	430,325
080O21	NP_956277	NP_571262	19	17.06	389,643
086C16	XP_001340912	NP_001038562	19	6.00	364,590
010H22	NP_998602	XP_001920851	21	2.24	240,019
034I14	NP_001038838	NP_956334	21	20.16	306,102
081J10	NP_001002411	NP_001073439	21	22.12	277,818
099H22	NP_001076277	XP_699221	22	7.03	115,754
053P11	NP_001002173	NP_001035137	23	30.73	197,703
068O10	XP_001339131	XP_689922	23	18.01	332,973
105H10	NP_001103636	NP_001098596	23	33.36	274,318
Average					**278,648**

Table 6: Microsatellite primers designed from the BES and used in the linkage mapping. This table provides the detailed information of the primers of the polymorphic microsatellite markers, which were designed from the BES and used in the linkage mapping in this study.

Ctg_ID	BAC_ID	Primer Name	UPPER PRIMER SEQUENCE	LOWER PRIMER SEQUENCE	LG
contig0002	035A2C05	AUBES2570_7	CCATGCTGCAACATATCCAG	GCTCATGTTGGAGACGTGAA	1
contig0004	045A1D11	AUBES2397_7	CTGATTCACCCGAGGAAAAA	ATAAAACCACGCAGGAGGAG	21
contig0005	104A1A03	AUBES3068_13	TTAAGTGCATGAGCCCACAC	TGTCCATCATGATTCCCAAA	9
contig0019	063A1D02	AUBES2737_7	AACAAACATCCGACCTCTGC	TGGTACACTGAACTGATGGATGA	20
contig0029	060B2H03	AUBES2606_13	GATCTTTTTCGGTCTCCCATC	GTCTGCCAACAGGAGTGTCA	2
contig0029	088B1C05	AUBES3359_13	AGAAGGGCAATTTGTGCAAT	CAGCAGACCTGTTTGGAGGT	2
contig0029	008A1G04	AUBES3192_13	GTACATCACCTGAGGGCACA	GGGAGCCAAGTGCATAAGAC	20
contig0033	022A1H02	AUBES2731_7	GCCATGCTTTCATCTCTATCC	GGTGTCTGGCTTGCTTATGTC	28
contig0037	011B1H04	AUBES1185_13	TGTTTACCGGCTGTGCGAGC	TCATGCCATTTAGCGGCCTG	2
contig0037	038B2C08	AUBES2623_13	ACCTAGCGTGGATTCAGCAC	CTGCTTCCGTCCACTCCTT	2
contig0037	023B1E10	AUBES1867_13	TGTTCACACACATTGCTGACA	TCTGGCAACTTTTACAGTGCAT	8
contig0040	036A2B09	AUBES1847_13	GCTATTAAAGGCTCGGAAGG	GCCTTCAGCAGGGAAATCTA	10
contig0040	106B2C01	AUBES3091_13	ACTCTGAGCCTGAGGGGAAA	TGGAAGAAATATGAGGATTCTGAC	10
contig0041	056B2G06	AUBES2859_7	TGGCAGTGATCCATTCACAT	AAATCCCGTGGTTTTGAGTG	16
contig0041	106A1A05	AUBES2979_7	TTGTCAGGTGTCCACATGCT	TTGTCAGGTGTCCACATGCT	16
contig0041	105A2C08	AUBES3242_7	GTTCTCCCCGGCGATATTAG	AGTTGCACAACAGGACATGC	18
contig0042	055A2C08	AUBES3166_13	GTCTCGAGCAGGTGATAGCC	AGACGTCCTGTCTCGGAAAA	7
contig0042	064B2B12	AUBES2634_13	TCCTGAGCTGCTGTGAGTTG	TGGTGTCCAGGAAGTGTTCA	12
contig0042	029B1H10	AUBES3108_13	GCATGGGCTGCGTAGTTTA	ACCCGTGTTTTCCGATACAG	12
contig0052	005B1A03	AUBES2708_13	GCTGTGCCGGATATACCATT	AGCTGCCATGACATTTTCCT	0
contig0058	030B1E04	AUBES3213_7	GTCTACTGCCAATCCTTAAGAGC	GCTCCTGCATCGGTAAAGTC	29
contig0059	005A1C12	AUBES2637_7	TGCACAGAGGCAAAATTACG	GACCAAAGGTTCCCACAAAG	3
contig0059	035B1F04	AUBES3678_7	TGAATTTGCACTGATGGTTCA	TCCCCTCATTGTATAACCTCAA	3
contig0059	062B2H04	AUBES3169_7	AACAGGTTAAATGCTGCTTATGA	TCGAATAAGACATGGCAGCTAA	18
contig0060	057B2B03	AUBES2740L_13	GGAGTGTGGGCAGTAGATCG	GAAAAGTGCACCGTCCTGTAA	11
contig0070	110A2B02	AUBES3243_7	CAGAAAGCTGCTGCATCTCA	TTTGTCATGACCACTAAGGTGTG	11
contig0080	010A2D12	AUBES3654_13	CTCGTCTGCACTCGAACCTT	CGACAGCGGAAGAGGATAAA	0
contig0080	028B1H01	AUBES1884_13	CCCAGAAAAGGGATCTTGGT	CCCGTGACCCTGTACGATAA	2
contig0092	027B1G06	AUBES2704_7	TTATGCTGTGTGTTTGCGTGT	CCCAGCATGAAATAAAAGACC	3
contig0092	074A2H07	AUBES2864_7	TTCAATCCCCAGGAAGAAGA	GGACACAACAACTCGGATGA	3
contig0104	027A2E05	AUBES1733_7	CGCATCCCAGCGTACAATTA	AACTGACCGGAACTACCTGTG	1
contig0104	014B1B09	AUBES2622_7	TGGTCGAGGTCATTTCTCATC	CAGTAGTTTTAGGCAGCACGTT	1
contig0104	098B1D06	AUBES3100_7	CCTGAAATCTTTCCTGTTTTTG	GCGGTCATGTTACCTTTGGT	1
contig0104	082A2C10	AUBES3158_7	CGGTGATGGAAATGTACACG	CAGTACGGGGAAGTGTTTTGA	1
contig0113	110A1A08	AUBES3513_7	GCCTATTTAGATTGGCCTTGG	CAGAGGTTCAGCTCAGCAGT	23
contig0114	005B1A11	AUBES2746_13	GGGAGATTGTTGAAAGGGAAT	AGCACATTACTGCCCACCTC	26
contig0116	031B1B09	AUBES3191_7	CCCCATACAAGGTAAAGTGCT	CAGAGGCAGTCAGCTTTTCA	1

contig0116	063B1C11	AUBES3354_7	ATCGCCCCAACTTTCGTTTA	CTCACTCTCGCCACGTGATA	1
contig0132	023B1A06	AUBES2749_7	CGTTCAGCAATAAAACGGAAT	CCAAACTCTGACTGAGAGGACA	21
contig0135	045B2H12	AUBES2593_7	AGGCTAAAATGGCTGGGTTT	CCTGGAAATGAGCTTCATCC	20
contig0135	051B2C10	AUBES3187_7	TACCTTGTTGGCACAGCAGA	TTTGAATCGACTGTTGCTCAG	20
contig0136	045B2G01	AUBES2423_7	TTATCCCTATACACGAGCACATAA	TCCTTGGCATCCATATTTCC	22
contig0139	035B2G10	AUBES2636_7	CAACGCGTGTATGCATTGTT	TGATAAATCCCACACGTTGC	29
contig0139	107B2D09	AUBES3168_7	GACGCAGCGGAATAACTGA	CATTACGCTCCAGCCAGAGT	29
contig0141	083A2B11	AUBES3755_13	ACCTGGCAACATGTGAGCAT	ATCACCAAAACGCAAAAAGC	26
contig0146	086A2A09	AUBES3247_7	TTAAAAAGCCCTCGCAGTTC	ATGGGGTGCCAATAATTCTG	28
contig0155	028A2F04	AUBES1748_7	TGACTGTGAGTAGTTTCCCTGCT	CACCTGCAACTGCACTATGA	9
contig0155	013B2C01	AUBES3094_7	CAGGCGGACCTATTGTTTGT	TGGAAAAGGACATGCATCAG	9
contig0155	048A2G04	AUBES3152_7	CAATGTGAGGAAGCCTGGTC	GTGTTTTTGGTTGCCCAGA	9
contig0166	042A1H02	AUBES2232_13	TGTATCGTAGGACGCCATTT	TGCATTCACAAGGCGATAAA	20
contig0166	044B2B11	AUBES2389_13	ATGGTCAGATTACCCCAAGC	TTCAACGATCAGCGTTTAGG	20
contig0166	009B2F10	AUBES2635_13	TAACGTTTCAATGGGTGCTG	GGTTGTGTGACAAAAACGACAC	20
contig0166	011B1E05	AUBES3109_13	TCACCTGCATCCAATTCAGA	TGGCACCCTTGGTAAATCA	20
contig0166	108A2E02	AUBES3167_13	AACGTTTCAATGGGTGCTGT	GGTTGTGTGACAAAAACGACA	20
contig0166	053B2D03	AUBES3201_13	CCTCCAGGAGCTTGAGATTG	GGCCATAGCGATAAGAGCAC	23
contig0170	026B1C08	AUBES2633M_7	TTATGCTTTGGGGGAAAAAG	ATTGGAAGGTCCGCACAAG	14
contig0170	090A1H09	AUBES3376_7	CATGAGCTGCACACACTCG	TGTCAAACCCCAATACTGAGAG	14
contig0171	078B1A04	AUBES3514_13	TGGCCAGGCATAGTGAAGTA	GCTTCTGCATGCAATCATGT	2
contig0171	071A2C10	AUBES3681_13	TGAAAACACAAATGCATCCAA	AGGCAAACAAGGATGTAGCAC	2
contig0180	051A1A12	AUBES2873_13	ACACAAGCCGTAACATGCAG	TCCCAAACACCATAACAAGGA	29
contig0181	053B1F10	AUBES3092_7	TGCAAGCAACCTTTTAAATCTG	CAGCATTAATGGCGCACTAC	17
contig0181	031A1D03	AUBES1892_7	GCATCACCATTCGTGGTCTA	TGCCATCAAGCGTTAGCATA	26
contig0181	025A2C11	AUBES2612_7	GTGTGAGATGTTGTTTCGTTGAG	GGGACAGCAGTCACACACAC	26
contig0181	090A1H01	AUBES3538_7	CAATTGGGCGTTTTATTTGG	ATGAAGCGTTGGATGGATTC	28
contig0189	057A2F09	AUBES3203_7	GCTGGGTACTCTGGTTACACAA	AACAGCTTTAACCCCGATGA	28
contig0189	092B2B07	AUBES3326_7	GTTAGAGGCCAAGGCGAAAT	GGATTTGCATGTCCAAAATG	28
contig0191	104A1E06	AUBES3329_7	GACTGGGTGTAAACAGAATGACC	GCCATCAGCCATTCATGTAG	25
contig0199	014A1A07	AUBES2609_7	GCACATGACCAGAGCACATT	AGGTTTTTGCCAAACTGCAC	3
contig0199	054A2G09	AUBES3149_7	TGTGAACCTCTAGGATAAGAGTCA	CTTCAGGGGTTTCCTCCAGT	3
contig0199	004B1D10	AUBES3702_7	GCACATGACCAGAGCACATT	GAAGGTTTTTGCCAAACTGC	3
contig0200	057B2A05	AUBES2628U_13	AGAGCCTAGGTGGTGGAATG	CCCGATAGGTCACGGACTAA	9
contig0200	112B1A02	AUBES3161_13	CCTTCATGGCTGAACAGGAC	GCAAAACATGTCCCACCATT	13
contig0200	088A2D01	AUBES3391_7	TGTTTAAAGCCGGATGTCCT	CCTGTCCTGATGACAAACTGG	13
contig0206	021B1C05	AUBES3066_13	CATCAACTGCCTCGGTTTTT	CGGAAGGAGTCTCCAGTGTT	2
contig0206	038B1B08	AUBES2576_13	CATGGTCTGAGCCTGAAGGT	ATATTGCCCAGCCCTAATCC	28
contig0209	035B1G08	AUBES2564_7	CCTGATTCAGAAGTGCTGACC	AGAGACGATGGTGCCACTTT	18
contig0209	075B2G07	AUBES3336_7	GGTTCCAGTTGTTCCACGTT	ACATGTTTGTGAAGTGCTGGA	18
contig0210	038B2D02	AUBES2040_13	GGTCACGTGATGGAGGTACA	CCGAATCTTTTACGGGATCA	0
contig0222	047B1C06	AUBES2608_7	AAGGTGACTGGATCTTCCACA	AGGTTTTGCACTGTGCTTTG	13
contig0222	073A1D03	AUBES3088_13	ACGGAGTTCGCATGTTCTCT	TGCTCACGATGGCAAGTTAG	13

contig0227	058A2C07	AUBES3065_7	TGTTGAGGTGTCTAGGATGCTG	AAAAGGGCCTGGCTAATTGT	22
contig0227	052B1H01	AUBES3536_7	CATTCCCTTGCTCCTTCATC	TTGAAATGGATTTCCCATAGC	22
contig0236	092B1H10	AUBES3080_7	TCCTCTGGAAGGCTCTCAAC	TTCTGCTTGCACCAAATCAG	4
contig0236	032A2G02	AUBES1763_7	CTGCATGTATGTGGCTCCAG	TCATGGCATGCATACTAACACA	28
contig0236	034B1A11	AUBES2597_7	CAGCAGCTCCAGAAAGAGGT	TCGGCATGCTGCTAAAAAC	28
contig0236	034B1C11	AUBES3717_7	AGCAGCAGCTCCAGAAAGAG	CTCACCGTTCTCTGCATCCT	28
contig0236	103A1E11	AUBES3749_7	TGGCAAAAGGTCTTGTTCTTC	TCTGCTGAACTCTCCGATGA	28
contig0241	070A1A08	AUBES2611_7	GATGGGTGATGAGGGTATGC	GGGTTCCTAAAAATGAAGTGGA	21
contig0241	111B2G05	AUBES3151_7	GAGGATGCAGTGAGGACACC	TGTGCACCGAGTGTGTTGTA	21
contig0241	084A2F05	AUBES3362_7	CAACCTCATGTGGGGTCAC	GAGAGCGTGTTGGACTTGGT	21
contig0242	057B1D06	AUBES2876_7	CATCAAGAATCCGCGACATA	TGGCTTTAAGACCTTTTGCAC	18
contig0242	053B1H12	AUBES3001_7	CAGGACCTAAACAGCCCTGA	CAGGACCTAAACAGCCCTGA	18
contig0245	033B1A05	AUBES2877_13	TGAGATGGGTTGACATGTGG	GGAAACAAATGGGGTTATCTCA	21
contig0259	003A1D11	AUBES1305_13	GTAGGCGGTCTGCCTCTCAG	AACACAGCCTGCCTCTCATC	24
contig0259	019B2H06	AUBES2699_13	TGCACCACTGGACTATGGAG	CTGTAAGCTCACTGCCACCA	24
contig0276	004B2B09	AUBES2613_13	ATTTCCCCACAAAGGCAAA	TAAAGGAAGCAGGGGGAAAT	9
contig0276	001A1A10	AUBES3193_13	TTCATTGCAGACAGCTGAAAG	TGGGGAATATCACAAAGTTCCT	9
contig0276	078A2C02	AUBES3364_13	AGGGTATCCCAAAGGTCTCC	CTGGGATAGGCTCCAGGTTC	21
contig0297	046A1F10	AUBES3683_13	TGGATGACCTCCTCCTTCTC	GCGGCATTTTTCAACCTAAA	2
contig0301	031A2G03	AUBES1606_7	GGTTAAACAGCTAGGTGCACTG	ACAGCCAGATGATTTCCAGTT	17
contig0307	027A1E08	AUBES2815_7	GCACGACTCAAAAAGCCACT	GCATCAGCCAAATGAACAAA	22
contig0314	021B1A10	AUBES1610_7	ACTCGCTGAAGAAGGCATTT	CGCTCACTACATAGGGCATGA	28
contig0314	023A1B02	AUBES1635_7	ACTGTGTTTGCCCAGGTGTC	CAGTGGAATGTCCTCACAAGG	28
contig0315	010B2D02	AUBES2716_7	TGGTCATCTGTGGTGTCTGC	TTTTGCAGGTCCTCCTCATC	6
contig0342	034A1G04	AUBES3861_13	CAGATTCAGTGTGATGGTGGA	GTACGGGGACAAATGCAATG	21
contig0371	074A2G06	AUBES2885_13	GCTGTTAAACCATTGCCTGA	ACCCTGCCTGCTTAGTGAAC	4
contig0371	038A1D10	AUBES2025_13	GTCCTCTCGGGTCTCCTCTT	CAGACCCACACACACAGACC	17
contig0374	087B1H04	AUBES3759_13	AACCAGCGGAGTTTGTGTCT	CGATCGTCTGTCATCCATTG	2
contig0375	102A1B06	AUBES3747_7	AGAAGGGAGTGCAGAAGACG	CTCAGACCTGGAGGCCAGTA	28
contig0376	034B2C01	AUBES1786_7	ATCACGGAACACACGTGAAA	TTCAGTCCCATTACACCGATT	28
contig0376	054A1D12	AUBES2886_7	CGCTAAATTGCCCCTAGATG	CTCTTTGTTCCCATCTCAATCA	28
contig0376	083A2D06	AUBES3257_7	CCTCTTGCGTTCTGTGTCCT	TCCTCTTTTTCTAATAACTTCCTTTC	28
contig0378	032A2H07	AUBES1354_7	ACACTGAAGCGGAACGATGG	ACAACGCTCAGTTGCTGGAC	28
contig0385	033B1D05	AUBES1286_7	GGACATTCTTCAGCAGCCAG	GAATGGTATCCTCGCCAAGC	28
contig0388	031B2A04	AUBES3484_7	AACAGCATGGGTGATCATAGG	GCACAGGTGCCTGTCAGTAA	28
contig0388	018A2D09	AUBES3485_7	GCACCTGGAGTGAACTGAATG	TTTCCATTGCATCTGTTCGT	28
contig0399	014B1H11	AUBES1333_7	TTGTGAGGGTCCTCATGCTC	GGATTGTACGTTCTGCTTGACG	10
contig0404	075B1A10	AUBES2591_7	GACGCAGCTAAGTGCCAGAT	TGTGCTGGTTAGCCAAGGAT	20
contig0404	047A2G05	AUBES3216_7	GTGAGGCGGAAATGCTACTT	AATGCCTTGACAGCATGTTC	27
contig0419	068B2D08	AUBES3259_13	CGAATTCTGGATTGTCATGC	CATCGTGGTCGACATACTGG	7
contig0419	006B1E01	AUBES2682_13	TGAAGCAGTATTCCCGACTTG	TTGCCTTATAAGCAGCAGAGG	14
contig0425	030B1G03	AUBES1538_7	TTGCATCGGTGCATCTCTAA	TGAGGGGTGTACTCACTTTTG	16
contig0426	058B2G11	AUBES2948_13	GTGCGGTTTCGGTTAGCTC	GCAGCGTTTCTTACCTTTGG	0

contig0426	061B1G06	AUBES3760_13	CAAGCAGCAAAGTCTGTGGA	CCACATGGACCTTGGAACAC	14
contig0435	049A1H06	AUBES3723_7	TGCATTCATTGTTGTATCATTCTG	AATGTTTTGGGGAAGCACCT	21
contig0436	072A1A03	AUBES2580_7	TGACATCAAAGATGCCCTCA	CAAGGCGTCAAAGAGGATTC	8
contig0436	111A1F02	AUBES3070_7	GCACATCCCAGAACAACCT	ACTGTGCCCTGTAGTTTTGGA	8
contig0442	053B2B04	AUBES2617_7	GGGTCACTAACTCAGGTGTGG	AGGTACAAAATGCCTTGACG	14
contig0442	011A1B06	AUBES3154_7	CGGACGTCACAGAACTCAAG	GCTTAGACGCGCAGAGTGAT	14
contig0442	095A2C07	AUBES3366_7	GTTTCGGCAGACCAATAGGA	CCACCATTGCCGTCTAAAAC	21
contig0442	039B2G10	AUBES2079_7	TCCACAACCCTAAATGTGGAA	CAGCCTGAAGAGACAAAACG	25
contig0443	052B2D06	AUBES3051_7	TTCACTGCTTTCCTCCAACC	GCCACATAACAGGACCAGTG	28
contig0443	059B2F10	AUBES3324_7	TGCAAATTGCCCCAAGTAGT	CCCTGGGATAGGTTCCAGTT	28
contig0454	045A1G03	AUBES3346_7	GACCATGGGATTGTGAAGTTG	GCCCTTCTGCAAGGTTCTTA	29
contig0458	032B2A03	AUBES3631_7	CCATTGGGATAACGTGGTCT	GAGATAGCCGGTGGCACTT	8
contig0458	081A2B09	AUBES3632_7	CGTGGTGGTTGAGCAGATTT	ATCGAGACGATCTTGCCACT	8
contig0463	046A1B03	AUBES3260_7	ACCATCGCACCTAGCAGAAC	AGTTGGAGCCAGGTAAGTGC	5
contig0463	087A1G03	AUBES3261_7	GAACAGGCACCACACACAAC	ACATTTGTGGCCCTGATAGC	5
contig0463	012B1H07	AUBES2694_7	TTGCTGGAGCCTATTTGACA	AGTCATCCCCATGCTTTGAC	6
contig0466	036A1D01	AUBES1935_13	AGCCATGACACCTGCTGATT	TGTTGCCAGCACCTATCATC	2
contig0466	026A1G01	AUBES2734_13	GGCGCGTATAAATCACCAGT	CCACTGCACCCATTAGAGTG	2
contig0466	041A1D03	AUBES3016_13	CCATGGCTCCTGGATTTAAG	CCAGCCAGCAAACATGAATA	2
contig0468	007B1H07	AUBES3420_7	GCGCTTAAAAGCAAATAGATTG	GGGCATGCCTTGGTTATATG	12
contig0468	077B1H08	AUBES3421_7	TTCTTCATGCCTTGCTCTAAG	GGGCATGCCTTGGTTATATG	12
contig0470	081A2C11	AUBES3423_13	TCGCACTCTCATAAGGACCA	CACTCAGCCCCAGTCCTAAC	2
contig0475	065B1F08	AUBES3729_7	GGAACCCCAAATTTCTCTCC	AGAGGGGAGCTGGTGTAGTG	20
contig0480	022A2G04	AUBES2555_7	ACTGTTGTTGCTGTTGTTGTTG	GCACAAACTGACAAACAGATTCA	22
contig0480	070A1D09	AUBES3327_7	GCCCGTAAAGGATTTGTTCA	ACGATCCCATCCACACTGAG	22
contig0480	081A2B06	AUBES3328_7	AAGGAGTCTCCAGTGTCAGCA	TGAGGAACATCCAAGAGTTTCA	28
contig0481	033B1B12	AUBES2889_7	ACAGTTTGTTCATGCGTCCA	GAAACCAAAGCGCAACAGAT	6
contig0481	100A2A04	AUBES2990_7	TTGCGGTTGTGAGTCAGTCT	TTGCGGTTGTGAGTCAGTCT	6
contig0482	036A1G04	AUBES2822_7	CAAAGTCCAGGTCACTGTCAT	GGAACAGGCGAAAGATGTGT	6
contig0483	013A2B01	AUBES3687_13	GCTGCCTGCCCAGTAAAGAT	GGTCAGAGGGTGTGACCTGT	11
contig0486	020A1A02	AUBES1742_7	GCCCAGGGGTCCAATTAC	TGCGTAGGTCAGATCCCTCT	22
contig0486	045B2E12	AUBES2419_7	ACCTTCCTCCATGCTCCTCT	TGAATGGCATGAAATGGAAA	22
contig0488	072B2E03	AUBES2769_7	CCCGTTCAGAGATGTTACGC	GCTGGAAAGATTCAATGTCCA	27
contig0503	011A1C02	AUBES2717_7	TGTGGTGAGAGGCATATGGA	CGCTCTCAAAGCAGAAGACA	16
contig0504	048B1B04	AUBES3688_7	CCCACCATTACAACACAGCA	ATCTGGCTCCCTCCTACCTC	12
contig0530	074A1A10	AUBES3004_13	ATTTTGTTGACACGCCCAGT	ATTTTGTTGACACGCCCAGT	2
contig0534	046A1E08	AUBES2478_7	CAGCCTGAGGAGACTTGTGA	GGGCACAGCAAAGATTGTATT	15
contig0534	023B2H03	AUBES2638_7	CTGCTTGGACTGTGTTGCAT	GCACTCCAGCCAGCTTAGTC	15
contig0534	052B2E02	AUBES3202_7	CCTGAGGCCTGAAACTGAGA	GAAAAGCAGCCCAGTTCATT	15
contig0534	023B2G10	AUBES3237_7	GTGCAGATCCACATAAGTCCA	GGATAAACCCGTGTTGTGGT	15
contig0534	013A2A03	AUBES3111_7	CAGAAATACACCCTGGAGCTG	CCCATGGCTCAGCTTGTAA	23
contig0541	043A1F08	AUBES3764_13	AGCACTCTCAGTCACGCTTTC	GGGATGAACTTCAGGACAGG	10
contig0551	030B1C11	AUBES3765_13	ATGCTGATGTGTTTGCTGCT	CGCGCTGTGATAAGAAAATG	16

contig0552	078A1A02	AUBES3766_13	CGCTTGCCAGCCTAAATTAT	CACTCGAATAGCGAATAGTTGG	2
contig0579	028B1F05	AUBES3767_13	GTTCTCAATCTCCCCTTCCA	CCTGTTAACCTTTGCGGTTC	29
contig0582	064A1A04	AUBES2895_7	AACCGATTGGTCAAGGTTCA	TGGGAGTTCATCCAAATGGT	9
contig0584	069B2A03	AUBES3304_7	CAGTGGAAGCCCCATAATGT	CTTGCCAGAACAGCTGGATA	1
contig0590	026A1A03	AUBES3433_13	GTGGGGGAGAAAATCCATTC	ACAAGCACACATTCCCTCGT	2
contig0594	008A1F01	AUBES2712_13	CCCACGGACACACAAATAGA	GCAATGACTGTTTTGCACCA	16
contig0597	047A2H03	AUBES2358_7	AAGGCTGTGTCACCATCACA	GCATGAGCGACATGAACTTG	12
contig0597	020A2F03	AUBES2772_7	GGAAGGCAGCGTCTAGAGAA	CCTTCTCTGAACCTCTGAACC	12
contig0601	019A1B02	AUBES3708_7	CCCTTAATTCTGGCTGGTTG	CAATTCTCCGAACATGAGCA	23
contig0603	058A2E08	AUBES3263_7	ACCTTCTACCACAGGGAGCA	ATGCGTATGAGGAATGGAATG	13
contig0643	023A1E04	AUBES2701_7	TGTTGGGTAAGGAGCCAGTT	TTGAAAACGCTGTTCCTGTG	6
contig0645	018B1G04	AUBES3789_13	TCCCCTGTGAGACAGAAAGG	GCCCAAAAGTGGATTATCAAA	2
contig0645	070A2B08	AUBES3791_13	TGCATTGATTCACTGTGGTTT	ATCCACACACCGCCTGTAA	2
contig0652	104B1G10	AUBES3440_7	GGAATGAAAACGCTCAAACG	GATGCTCGGCTTGGATAGAG	3
contig0661	077B2C04	AUBES3769L_13	CCTTTATGGATCCGGTAAAACA	TCAGTCTCCCCACTCACCTC	0
contig0661	077B2C04	AUBES3769U_13	CCTTTATGGATCCGGTAAAACA	TCAGTCTCCCCACTCACCTC	2
contig0666	084B1A10	AUBES3309_13	TTCCAAAGGATTGCAAGGAG	CAGGCTAACTGGGGACACTG	2
contig0675	072B2D04	AUBES3731_7	ACGTGTGCTGAGGGAGAGAC	TTCTCACCAAATGCACCTACA	24
contig0690	026B2C07	AUBES2318_7	CTGTCAATGTGTGGCTCACC	GGCAACCATCTGTGTTTGGT	27
contig0690	053B1E10	AUBES2905_7	GTCATGTGTGGCTCCCAAGT	GAGGGGGAAGAGGAAAACAT	27
contig0737	023A2A10	AUBES2152_7	CGGCTGACTGAGGATTTGA	AATCGCAAGCTGGAGAATCA	4
contig0752	002B2D09	AUBES3560_7	TAACGCACAGCTAGGCACAC	GCTGTGGGCTCGTAATTAAA	6
contig0766	019B2H05	AUBES1306_13	CACTCCCGGAATACACGCAG	TCAGCTCCAGCGCAACGAGG	2
contig0766	011B1H02	AUBES2689_13	AGAAATCCAGCCACGGAATA	TGCCAGCTTTAAGCCTCTGT	2
contig0769	040B2G07	AUBES3015_13	TTCTGTTGGTGGAAAGATCTGA	TGACCATAATCCAACTCTTCAAT	2
contig0786	045B2D04	AUBES3792_13	TAACTCCGCGTTCCTAAAGC	GATGCTGGTGGTAGAGAGTGC	2
contig0786	074B1G05	AUBES3794_13	TTCAGGATGGAGGACTTTGC	TGTTGTGGATTCCCCAGAAG	2
contig0790	038B2F10	AUBES2567_7	TGGCTGATCGGTGTAGAGC	TACTGGCAGTGATTGGCTGA	29
contig0790	087B1D01	AUBES3742_7	TGGTTTGTACCCATGCTTCA	GTGCTCCTGTCACTCACCTG	29
contig0798	040A2G03	AUBES3013_13	GCTAACATTCCCCAAATGGA	CATGTTGAGCATATACCTGTATGAA	2
contig0801	032A2E08	AUBES3867_13	CTGTTACAGCTGCGAGATGC	TGCTGCCTACAAGTGACAAAA	16
contig0801	090B2G11	AUBES3872_13	TCAGGCCAAGGTTACCAGTC	TCCAAACAGAGGAGGCAGTT	16
contig0808	073B1E11	AUBES3795_13	TGACCTCAACAATGTATGCTCA	TTCTGACATGCCGATCTTTT	2
contig0818	076B2B08	AUBES3736_7	TTTCACAGCTTTCTCTGCTCA	ACCCCACTGTGAAATCCTGT	1
contig0830	095B2D08	AUBES3692_13	GAGGGCCGTTGGATTACTTT	CTTGCAAACGGTCCATAACA	2
contig0843	019A2C08	AUBES1544_13	TTGCCTATAGTTTACGATCACAAT	AATGCAAGTGGTACAGCCCTA	26
contig0844	033B2G03	AUBES3571_7	GCCTCACTGGGTAACTTGCT	CAGTTGACATTTTGCTGACG	5
contig0844	084B1F08	AUBES3572_7	GCCTCACTGGGTAACTTGCT	CAGTTGACATTTTGCTGACG	5
contig0904	059A2C11	AUBES3317_7	AACCCCTGAGCATTAGTCCA	TCTGGGTATCTGGACAGCAC	8
contig0904	082B1D07	AUBES3739_7	CAGCTCACTCGGTTCTGTCA	CAGCAACGCCTATCACACAG	8
contig0913	062A2C10	AUBES3878_13	CCATGGAGCTCTGGATCTGT	CCGTTTTAACATTGCAAGACA	15
contig0919	045B1E12	AUBES3879_13	GCGACTGAGGGGAATAATGA	ACACACACACACATGCTCGT	22
contig0935	041A1B06	AUBES2084_13	ACCCGGAAGAGAAGTGAACA	GTGGACCCAAAAAGGTTACG	14

contig0939	106B2G01	AUBES3576_7	CCACCATCCCGTACAGTTTC	TGTCGCATTTGGATAAGACG	14
contig0959	039A2F04	AUBES2059_7	CCGCATAACTGGGATCATTT	CGTTTATGTTGGTGCCGAAT	6
contig0974	002B1F04	AUBES3798_13	ACCAAGCCAACAGAGCAAAT	TGATGTCGCGGTGATTATTG	2
contig0974	020A2D03	AUBES3799_13	CAGGAACTTGTATTGCGGAAG	TGTTACCACCCCGGAAATAA	2
contig0994	066A1A07	AUBES3772_13	GGGCACTAATGAGCACACAA	TGTTAAGCCTCAGCCTTTCAT	2
contig1007	041A2F10	AUBES2099_13	CGGCTTCGAGGAAAGTTTTT	CCACGCTTAAAATCGATACTGC	18
contig1017	014A1A01	AUBES3456_13	GGATTCCTCTGCATTTCTGC	GTGAAGTATCAACCTAATCATTGACA	20
contig1017	092B2D03	AUBES3553_13	TGCCTCTATTTGCCTGTTTC	TGTATAAAGTGCCTTGAGAAGCTG	20
contig1029	023A2D08	AUBES3801_13	CATCGCTCAAAAGCAGACAC	AACAGTATGAAACATGCTCCAATTA	6
contig1029	077B2D08	AUBES3802_13	TCTGCTACTTGCTACTGCTGCT	GCGTGACCGGTATGTGAAG	6
contig1031	072A1G08	AUBES3457_13	TGAACACCAGGACAACATGAA	GGTAACCACTACGCCACCAT	2
contig1031	109B1C05	AUBES3458_13	TGGTGTCCCAGATAGGGTGT	CTTAAACCCCTGGACCCACT	2
contig1044	047A2H08	AUBES3695_13	CTGTGCATGAGGCAGACATT	GGCTCCATCCATCCAGTTAT	15
contig1056	049A1E10	AUBES2532_7	TCGGATTCCTCATCATCCAC	ACGGCCAAAGAAACAAACAA	28
contig1058	046B1D06	AUBES3804_13	CCTGGGTTTAGACTGGAGGA	TAAATCTGCGCTCGGCTAAT	2
contig1058	071B2E09	AUBES3805_13	TTTTCTCCCACCTGCTGAAG	AGTGAGTGAGTGGGCTTCGT	2
contig1058	102A2D06	AUBES3806_13	TGTGACCCTTTTAGAACGATG	AAAAACAAACCCAACAAGCA	2
contig1059	020A1H08	AUBES3773_13	AGCCAACAATTTCCATGAGG	TTCATGTCAGCGGTTCACTC	2
contig1085	029B2A11	AUBES2799_7	GCGGTACTGTACTTCCATCCA	GATGACCTTCCATGCCTGTC	27
contig1109	035A1B03	AUBES3774_13	TGATCATGGCAGCTTCTGAG	CTGTGCACAAATCATTCATGG	2
contig1128	045A2F07	AUBES2411_13	ACCCTGACACAAGCTCAAAT	GAGTGACATCTGCTGGTGGA	22
contig1139	073A1D08	AUBES3732_7	TCTCAAGCATGTGAGAAATCTG	AAGTAATCAATGCGGCCAGT	5
contig1142	038A1B06	AUBES3535_7	CTTGAAGTCAGACAGTACTCATGG	CGGCCATGTTTCTGAAGATT	6
contig1148	107A2B10	AUBES3807_13	GCAAATGGGTTTTTATCAGCA	CCCAACCACAACAAGATCAA	20
contig1167	014A1E03	AUBES1277_7	ACGGAGACTCCATGCTGAGC	GAATTTACGGGAGGCCGCTG	6
contig1196	022B1H11	AUBES3462_7	TAGCAGGAAATTAGCGGTCA	CACTTGCACAAATGCTTCCT	5
contig1208	040B2A04	AUBES3321_7	TGCCACCCATGCATTTAATA	CAGTGTGCAGTGTACCTTGTGT	28
contig1211	041A1H09	AUBES2091_7	GCTGTGTGTGCCCTCATCTA	ATGACCACACCCTTCTGACC	22
contig1222	060A1E09	AUBES3885_13	TGGCTCAACACATGCCTTAT	GCCAGTGTACACCTCAGAGC	18
contig1264	038B1E03	AUBES2034_13	CATGTGAGTGAGATGGGCTCT	GCCTGGAGCTTGTGTTTGTA	2
contig1288	046A1E05	AUBES3888_13	ACGCAGACAGAGCTTCCAGT	CCAGCCCTTAGCACATGTTT	9
contig1288	055B1G01	AUBES3889_13	GGTCAAACACTTTGGGATCG	TAATGTGGGAGGGGCCTACT	9
contig1310	109A1E10	AUBES3809_13	CGAATAAGTAACATTTTACACGTATGG	TTGGTTGGTGAGTTTCTACCTG	2
contig1353	022B1B08	AUBES2802_7	TTGGCATCCTTGGTAAATCTG	GCAAAGGGAGGTAGCATTGA	28
contig1353	065B1H12	AUBES3322_7	GCAGGAGAAAGCATCCTGAG	ACCCCATGTTTGGGTAATGA	28
contig1369	023A2H09	AUBES3777_13	CCTGAAATTCCCAGGTAATGA	GCACCAACAACCATGCTACA	2
contig1377	020B2D10	AUBES1951_7	CTGCTGAGATATGGAGGAGGA	CCACCCCTCCCTTTGTTTAT	5
contig1387	047B1G04	AUBES3021_7	TGCAAGTGAGAAAACACCACA	CGCACATACACAGCGTTTCT	5
contig1387	077A2E06	AUBES3737_7	AGCTGCATGGGCTTTCATAA	CTTTTGGCTTCCACTCAAGC	5
contig1396	027B2B05	AUBES1759_7	GCCCTCTGGCAATAACAAGT	TGGAAGATGGTACGGGAGAC	6
contig1401	041B1D07	AUBES2103_7	TCTGAGCTCTGACCCCAACT	TTTTACACGTGACCAACCTCA	7
contig1401	006A1D02	AUBES2709_7	GCCCGACTATGGAAATAACG	TTAACTGCCAAGCCACCACT	7
contig1401	046A2D01	AUBES3720_7	AAGCTCTCCTGGGCGTTTAC	GTGCTGCTGGTTTGTTTGTG	7

contig1455	106A1A08	AUBES3892_13	TGATAGAACTGGAACACACAAATG	CTCAAAAGCAGGATCCGAGA	2
contig1468	027B1A08	AUBES3812_13	GCCTGCTGAGGGATATTCAC	GGCAAATGGGAAGTGAGAAA	2
contig1496	003B2D06	AUBES3813_13	AGTTGTGGGTGGTGGTGAGT	TTGTGTCCCACCATTGTTTG	20
contig1496	048A2C03	AUBES3814_13	TATTCAGCTCCCGAATCACC	TCAGTCTTCTGGCCTTCACA	20
contig1496	094B2A03	AUBES3815_13	AGTTGTGGGTGGTGGTGAGT	TTGTGTCCCACCATTGTTTG	20
contig1502	110A1B06	AUBES2996_7	AAGCCGAGTTAAACCTGCAC	AAGCCGAGTTAAACCTGCAC	6
contig1525	025B2F07	AUBES1377_7	CAGTTACAAGGGGTTTCCCAG	CACATACACACCTCCATGAGC	16
contig1536	027B1G12	AUBES1964_13	TCCTGTTTTCCCTTCACAGC	GCTCCCCATCTCTCTGTTTG	10
contig1584	048B2E07	AUBES3817_13	GAGCAAAGGACAGACACACA	GCTTGAAAGTGTCAATGCAAA	7
contig1604	110B1G07	AUBES3894_13	TAGGAGAACTGGTGGGGACA	AGATTGTGGCTGCGGTAAGT	22
contig1604	038B2A08	AUBES3895_13	TAAATGCAGGTGTGGACTCG	AAGCGACAAACTAGCAAGTGAA	22
contig1653	048A2C04	AUBES2508_7	ATTGCAAGGACTGTTGGACA	CACAGGTTGCAAAGTGACTGA	14
contig1684	105A2E09	AUBES3781_13	CCCACATTATGCCTTCACG	TAAACTCTGCAGCCTTCCAA	7
contig1684	038B2B01	AUBES3818_13	TGTTTAAAGTTGAAGTGAATGGAG	CCCAAAGACTACGTGCAAGA	7
contig1693	019A2D06	AUBES1947_7	CAGATCCTGATCCCTGATGG	TGGACTCTGCCTTTTGATCC	4
contig1702	022B2A04	AUBES2148_7	AAACATTGCGTGAGAGACCA	ACGGATGTGGTGAGCTTCTC	12
contig1702	045A1G01	AUBES2400_7	TGCAGTAATCAGTTGTGGGAAT	GCCAAGACTTTTGCACCCTA	12
contig1736	081A2C09	AUBES3822_13	TTTGTCGAACACGGTTATGG	ATCACACTTGCGTGACATGG	2
contig1776	090A1C05	AUBES3323_13	TGCCTACCTGCAGTATTTTCAA	CAGTTTGCAGGACAAGAACG	2
contig1802	026A2D06	AUBES3823_13	GCTCGATATGACTGACTGGATG	TTTCAGGAACAGCTGAAAAATG	2
contig1813	077B1D03	AUBES3783_7	CTCAGTCTTGCAGCATCAGG	TTTCTCTTCCTCCTACCCCTCT	21
contig1828	065A2G09	AUBES3828_13	TACCTGATGCTCCAGGTTCC	CGACATGAATCCCTGAGACA	2
contig1828	090A2F04	AUBES3829_13	TTCCCTGTGACCCTGAAAAG	AAACAGCAGCCGAACTTCAC	2
contig1894	052B2C07	AUBES3701_13	AGGTGTGAGGTGTGAACGTG	TCTGTCCATTCCCCAGAGAC	2
contig1919	111B2D10	AUBES3835_13	GTTGGCATTGTACACGTTCG	CTTTTCACGGAGCCAGATGT	1
contig1931	043A1F07	AUBES2263_7	TGCAGGAATCATGAAATGGA	CCACATACAAGGCTGCACAC	28
contig2068	026B1B11	AUBES2316_7	CAAATGTCCACAGGGAGCTA	ACACAGGCATGGTCAGTTCA	0
contig2128	021B2H09	AUBES1353_7	GTCCTCCTGAAGTCCAGACG	TCACCAGCTTGCTCTGAGAC	28
contig2128	025B2E12	AUBES1990_7	TGATGTAGAAGACCCCCATGT	TGCAGCTCCTAGTGGTATGC	28
contig2239	092B1A12	AUBES3745_7	ACACCCCTGCTCTGTCACAT	TGAAATTCCACAATCCCACA	12
contig2239	108B1E11	AUBES3752_7	ACACCCCTGCTCTGTCACAT	TGAAATTCCACAATCCCACA	12
contig2257	035B2D06	AUBES3839_13	CGGTTCTGTGGCTTCATTTT	CAACGAACGGCTCCATATCT	6
contig2348	031A1H08	AUBES3841_13	ATCACCTTCCCATTCCCTTC	AAGCGGATAATGTATACACATAGGG	15
contig2495	032B1H08	AUBES1764_7	CAATGTCGAGGCCAATCTG	AGCACGGAATGACTGCTTTT	23
contig2533	024A1A04	AUBES3844_13	GGATAAGGACGTCTGCCAAC	GGTGCAAGGAAAATGACAGG	2
contig2533	111B2D04	AUBES3845_13	CATTTCGATCCAGGGTGTTT	CAAATCTGTCGGGCTTGACT	2
contig2630	068A2D11	AUBES3899_13	CTGTGCAGCCCAACATGA	GGGACAGAGAGGCAGAGAGA	2
contig2630	100B1B06	AUBES3900_13	AGCAGCAGCACAATGCATAA	TCCATAATCCCCCATTTCTTC	2
contig2654	042B2D03	AUBES3846_13	CATTGCCAGAACCTCACAGA	GCAATTGTGCATTGATCAGC	2
contig2724	019A2C04	AUBES1281_7	TGGAGAGAGTCAACACGCAG	ACCATAATCTTGCCACAAGCAG	5
contig2833	002A1H01	AUBES1282_7	TCATTCTGCCACCTGCTGAC	CTGTGACTCCCTCAGACTGC	2
contig3026	085B1F06	AUBES3852_13	GCGAGAGGGAATAAATGCAG	GTTTGAGCAAAGCCTGATGG	2
contig3420	050A2H04	AUBES3854_13	TACAACTCGCTCACGCACTC	AGTAAACGTGTCCCGGTTTG	24

Figure 1: Microsyntenies established through sequence homology comparisons using tBLASTX searches of the catfish BAC contig-associated BAC end sequences against the zebrafish genome sequence. The putative conserved microsyntenies are presented along the 25 zebrafish chromosomes (chr 1 through chr. 25 labeled at the top of each chromosome). The position of the zebrafish sequence along each of its chromosome is shown on the left of the chromosome bars in million base pairs, e.g. Chromosome 2 is 55 million base pairs long. The conserved microsyntenies are indicated on the right side of the chromosome bars, with the numbers representing the contig numbers of the BAC assembly of the catfish physical map. Circles represent short syntenic regions and short vertical lines represent relatively longer conserved syntenic regions proportional to the length of the bar with a number in parenthesis representing the number of conserved sequences within the microsyntenies. The microsyntenies designated with asterisks (*) are those with duplicated conservation of the microsyntenies that are color-coded to facilitate the visualization of the duplicated syntenic regions along the chromosome. Duplicated syntenic regions refer to a conserved genomic segment between the catfish genome and the zebrafish genome that is duplicated in the zebrafish genome such that identical or nearly identical significant hits are generated from two chromosomal regions of the zebrafish genome using a single catfish genome segment (say it is a contig or a scaffold) as the query. In just a few cases, this term is used in an extended fashion to include those that are tripled in the zebrafish genome.

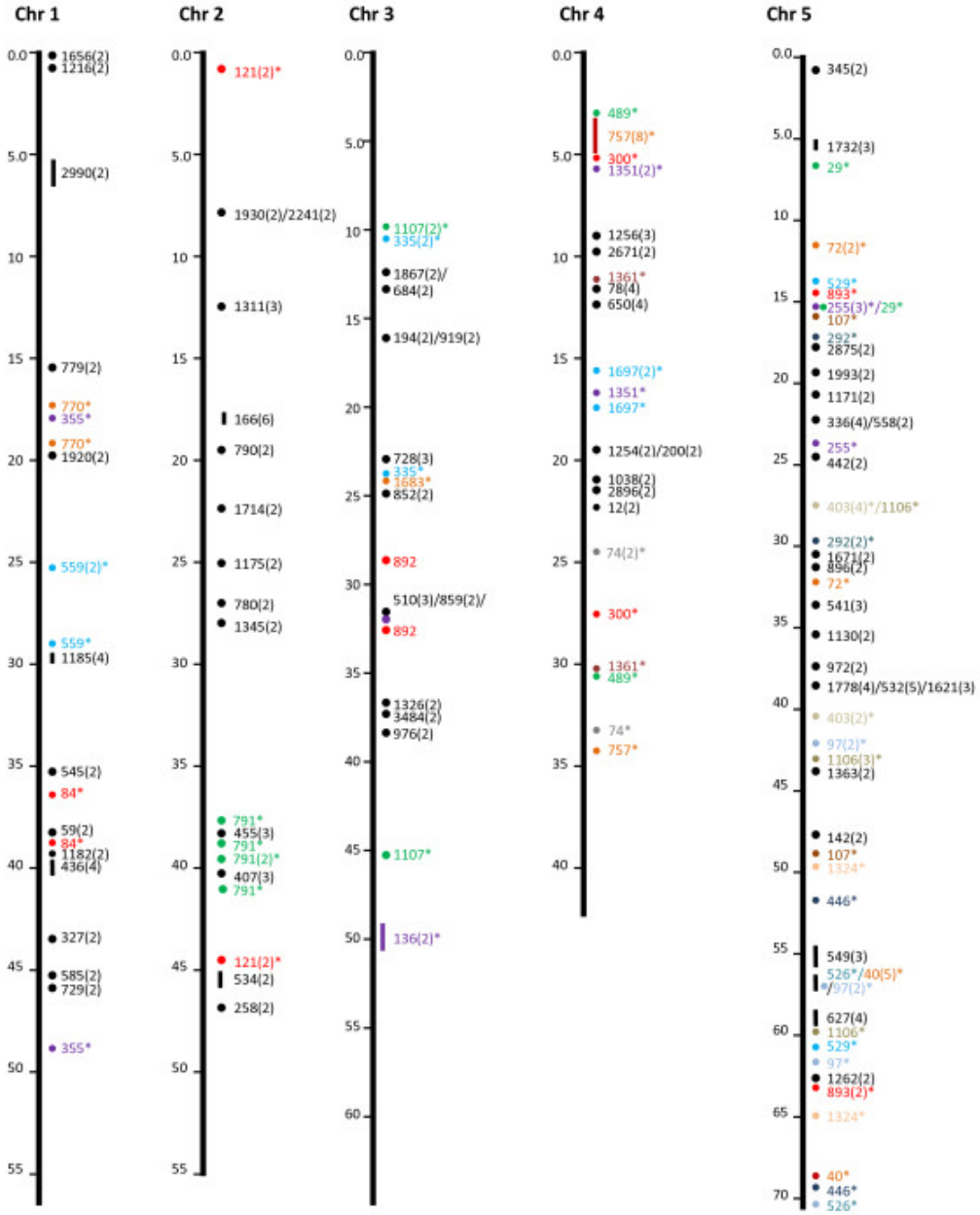

Chr 1
0.0
1656(2)
1216(2)
5.0
2990(2)
10
15
779(2)
770*
355*
20
770*
1920(2)
25
559(2)*
30
559*
1185(4)
35
545(2)
84*
40
59(2)
84*
1182(2)
436(4)
45
327(2)
585(2)
729(2)
50
355*
55
Chr 2
0.0
121(2)*
5.0
1930(2)/2241(2)
10
1311(3)
15
166(6)
20
790(2)
1714(2)
25
1175(2)
30
780(2)
1345(2)
35
40
791*
455(3)
791*
791(2)*
407(3)
791*
45
121(2)*
534(2)
50
258(2)
55
Chr 3
0.0
5.0
1107(2)*
335(2)*
10
1867(2)/
684(2)
15
194(2)/919(2)
20
728(3)
335*
1683*
25
852(2)
892
30
510(3)/859(2)/
892
35
1326(2)
3484(2)
976(2)
40
45
1107*
50
136(2)*
55
60
Chr 4
0.0
489*
757(8)*
5.0
300*
1351(2)*
10
1256(3)
2671(2)
1361*
15
78(4)
650(4)
1697(2)*
1351*
1697*
20
1254(2)/200(2)
1038(2)
2896(2)
12(2)
25
74(2)*
300*
30
1361*
489*
35
74*
757*
40
Chr 5
0.0
345(2)
5.0
1732(3)
29*
10
72(2)*
529*
15
893*
255(3)*/29*
107*
292*
2875(2)
20
1993(2)
1171(2)
336(4)/558(2)
255*
25
442(2)
403(4)*/1106*
30
292(2)*
1671(2)
896(2)
72*
35
541(3)
1130(2)
972(2)
1778(4)/532(5)/1621(3)
40
403(2)*
97(2)*
1106(3)*
45
1363(2)
142(2)
50
107*
1324*
446*
55
549(3)
526*/40(5)*
/97(2)
627(4)
60
1106*
529*
97*
1262(2)
893(2)*
65
1324*
40*
70
446*
526*

Figure 1 continued

Figure 1 continued

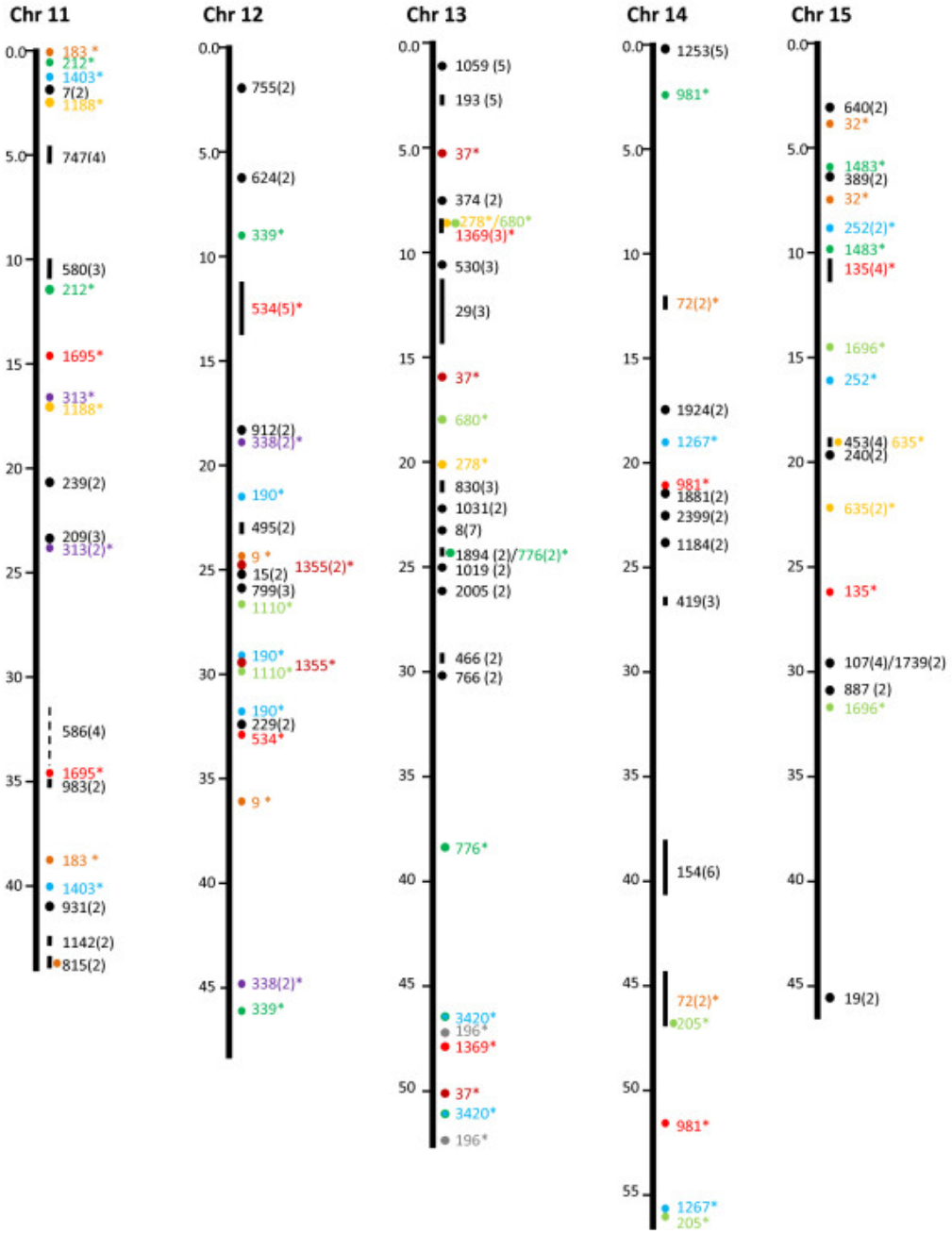


Figure 1 continued

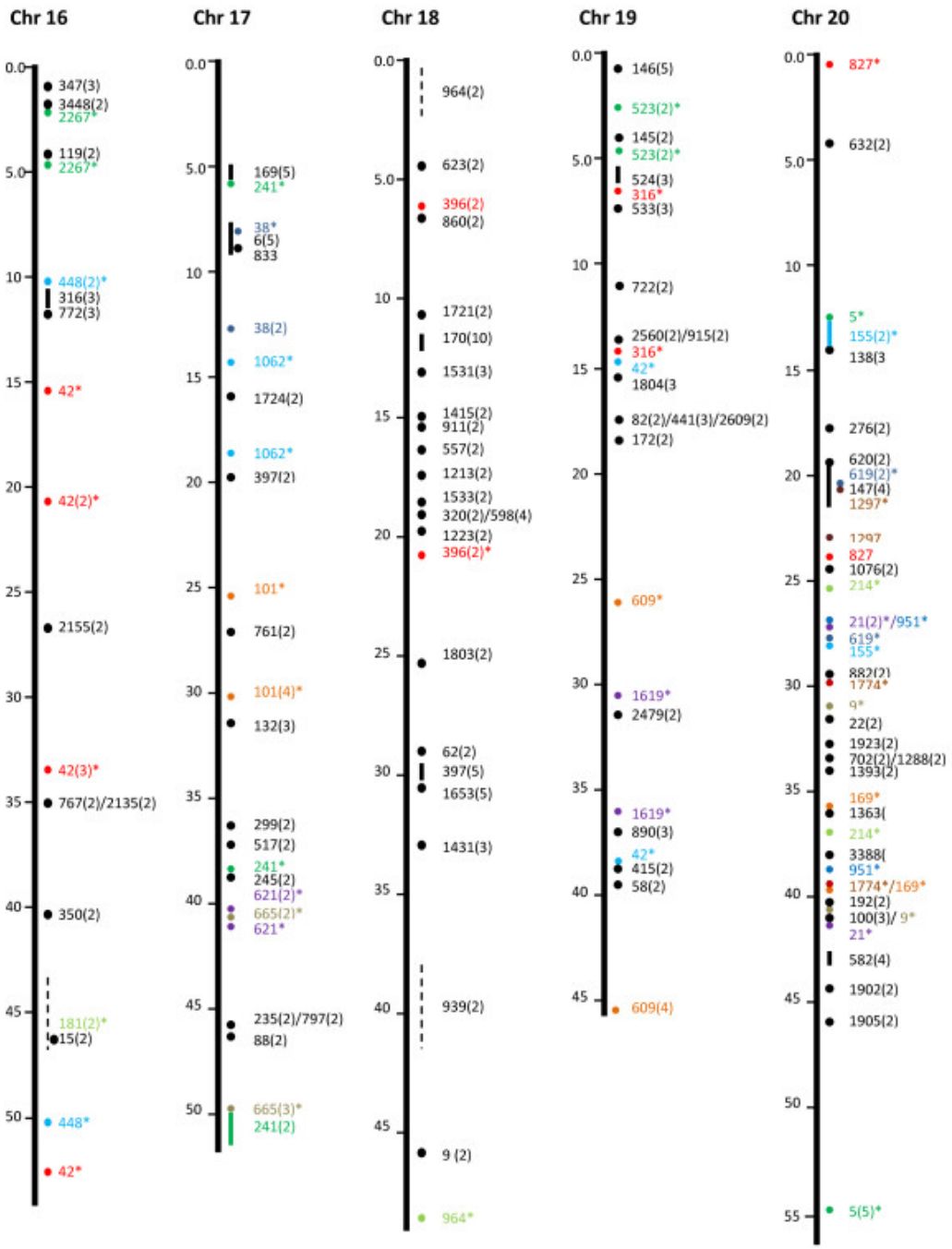


Figure 1 continued

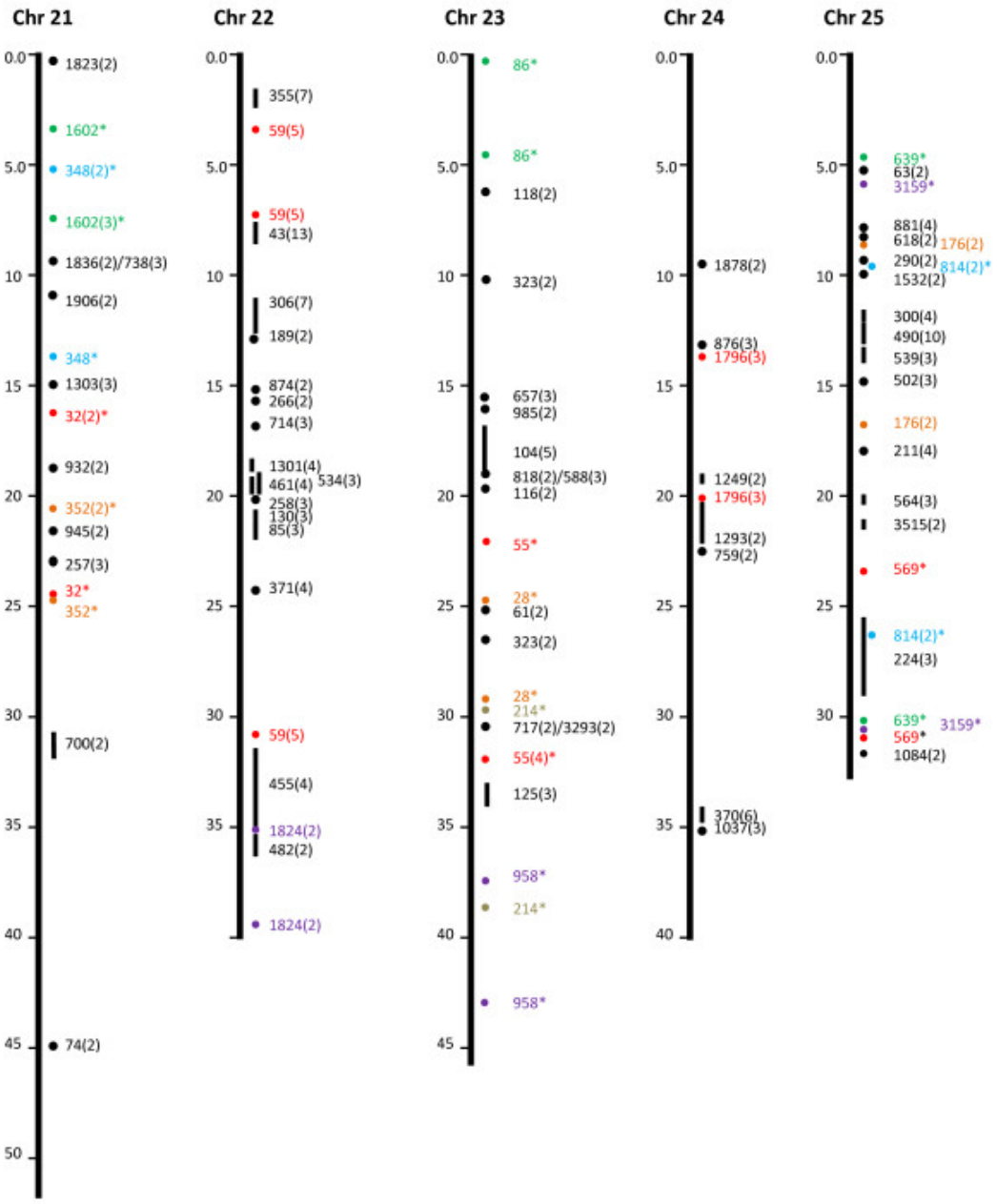


Figure 2: Scaffolds of conserved syntenic regions between the catfish and zebrafish genomes as established by genetic linkage mapping of BAC contig-associated microsatellites. The zebrafish chromosome 13 is shown left (chr. 13) with its base positions indicated on the far left in million base pairs; The conserved syntenic regions between catfish and zebrafish is shown immediately left and immediately right of the chromosome, with numbers on the left representing the contig numbers of the catfish BAC contig assembly of the catfish physical map. The numbers in the parenthesis are the number of conserved sequences; the circles and bars represent relatively short and long conserved syntenic regions; the asterisks represent duplicated syntenic regions with color coding to facilitate the visualization of duplicated regions, the same way as described under Figure 1 legend, except that the open circles represent conserved sequences coming from non-gene sequences while the solid circles represent conserved gene sequences. Microsatellites from the BAC contigs were genetically mapped to linkage groups as shown on the right, with the names of microsatellites being labeled on the second most right, e.g. AUBES1884. The positional relationship of the conserved syntenies on the zebrafish genome sequence and within the catfish linkage group is indicated by thin lines linking the zebrafish chromosome and the catfish linkage group positions. The positions of markers within the linkage group are shown on the furthest right in centi-Morgans.

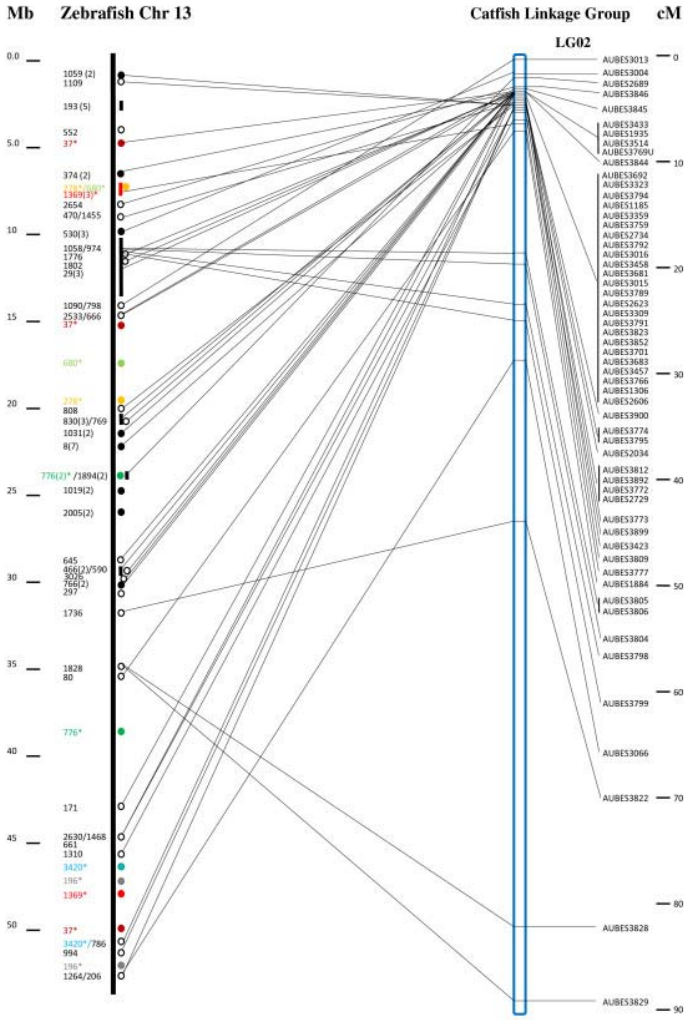
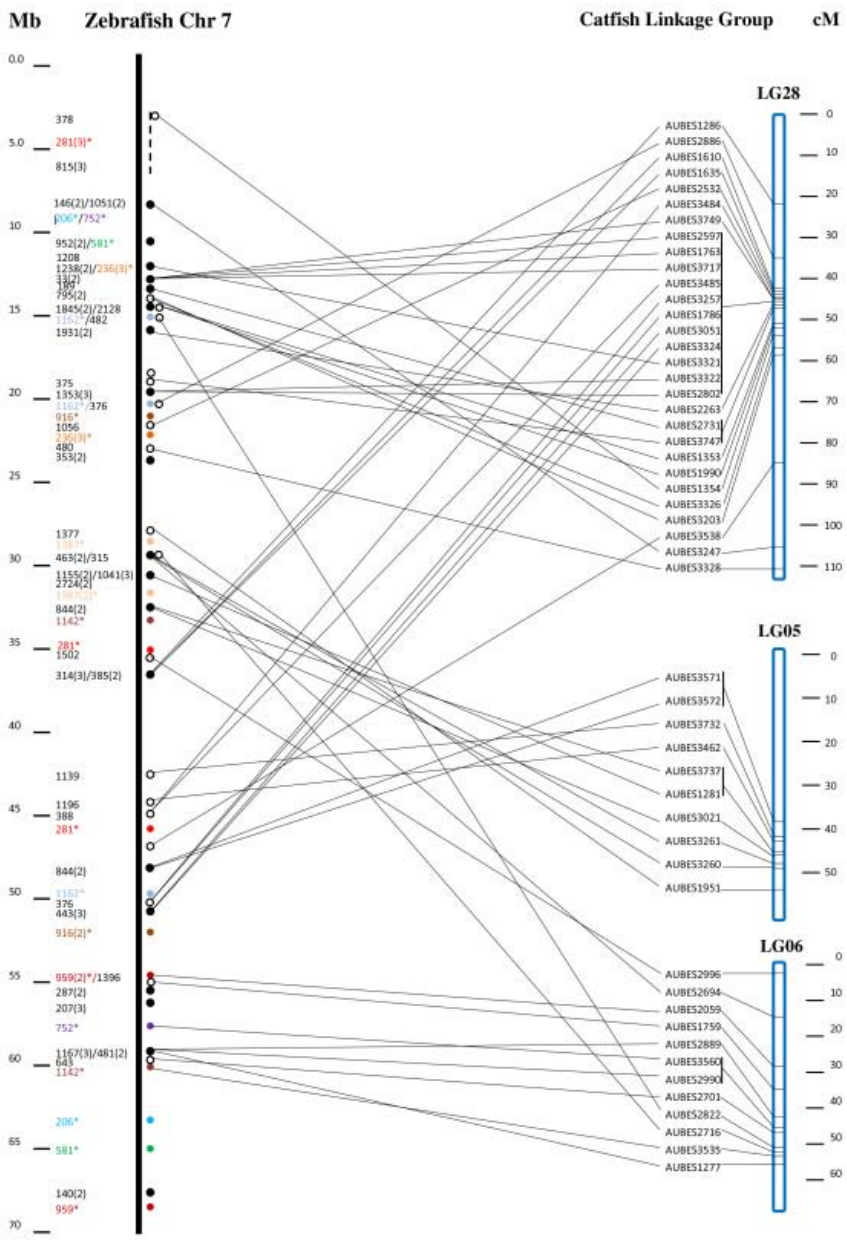


Figure 3: Scaffolds of conserved syntenic regions between the catfish and zebrafish genomes as established by genetic linkage mapping of BAC contig-associated microsatellites, as shown in Figure 2 except that the conserved syntenic regions on the zebrafish chromosome 7 was mapped to various linkage groups, with three major linkage groups shown here.



V. GENERATION OF PHYSICAL MAP CONTIG-SPECIFIC SEQUENCES

Abstract

Along with the rapid advance of nextgen sequencing technologies, more and more species are being added to the list of organisms whose whole genomes are sequenced. However, the assembled draft genome of many organism consists of numerous small contigs, due to the short length of the reads generated by nextgen sequencing platforms. Great efforts have been made on the genomics studies of channel catfish, the primary aquaculture species in the United States. The ongoing assembly of catfish whole genome sequences has already yielded over 200,000 contigs. In order to improve the assembly and bring the genome contigs together, more genomic resources are needed. In this study, we added a valuable genome resource to the list. We generated physical map contig-specific sequences, which are randomly distributed genome sequences in a physical contig, with a minimal cost. The physical map contig-specific sequences can serve as the anchor points to link the draft catfish genome contig to the physical contig, thus improve the whole genome sequences assembly and scaffolding. We developed a strategy to generate the physical map contig-specific sequences. A two-dimensional tagging method was used to create specific tags for 1,824 physical contigs, in which the cost was dramatically reduced. A total of 94,111,841 100-bp reads and 315,277 assembled contigs are identified containing physical map contig-specific tags. The physical map contig-specific sequences along with the currently available BAC end sequences were then used to anchor the catfish draft genome contigs. A total of 156,457

genome contigs (~608 Mb) were anchored and grouped into 1,824 pools, in which 16,680 unique genes were annotated. This study describes a strategy to generate the physical map contig-specific sequences at a low cost. The physical map contig-specific sequences are valuable resources to link the physical map, the linkage map and draft whole genome sequences, which have the capability to improve the whole genome sequence assembly and scaffolding, and improve the genome-wide comparative analysis as well. The strategy developed in this study could also be adopted in other species whose whole genome assembly is still facing a challenge.

Background

With the advances of sequencing technologies, genomes of many species with biological or economic importance are currently under sequencing. With the exception of the PacBio sequencing platform, several nextgen sequencing technologies such as 454 sequencing, Illumina sequencing, and SOLiD sequencing produce relatively short sequencing reads, making subsequent sequence assembly a great challenge.

With most eukaryotic genomes, several factors further complicate whole genome sequence assembly: 1) the genome size is most often large with a billion base pairs; 2) most eukaryotic genomes contain repetitive elements that are either in tandem repeats or dispersed in the genome; 3) eukaryotic genomes are loaded with long tracts of simple sequence repeats (microsatellites) that most often pose sequencing challenges because of frequent terminations at such sites (Jiang et al. 2011). All these result in segmented genome assembly. Typically, repetitive DNA sequences such as transposons and short tandem repeats that are interspersed in the genome shatter the *de novo* assembly because the sequencing reads are not long enough to include the entire repetitive sequence plus unique flanking sequences. As a consequence, assembly algorithms cannot uniquely assign sequences that arise from within the repeat, resulting in shortened sequence contigs; similarly, sequencing reactions terminate when encountering microsatellite repeats causing assembly breakage at the microsatellite sites even though microsatellite loci themselves are relatively short, most often within the 100 bp in size. Of course, with a large genome and short contigs, the consequence is a large numbers of contigs.

Such challenges become even more significant when dealing with teleost fish genomes because they went through an additional round of whole genome duplication (the 3R hypothesis) (Meyer and Van de Peer 2005; Steinke et al. 2006). With some teleost fish such as common carp (*Cyprinus carpio*), Atlantic salmon and other salmonid fishes, the situation is even more complex because their genomes went through yet another round of whole genome duplication (the 4R hypothesis) (Moghadam et al. 2009; Moghadam et al. 2011). The duplicated genes cause confusion in assembly because the assembly software cannot uniquely place the highly conserved gene sequences into its own contigs, but rather in many cases, sequence stacking has resulted, leading to mistakes in sequence assembly.

Several pilot studies on genome sequence assembly of fish species have been conducted (Quinn et al. 2008; Kuhl et al. 2011; Jiang et al. 2011) utilizing the high throughput sequencing data generated by next-generation sequencing platforms. In all cases, however, it was concluded that it is difficult to achieve a good level of genome sequence assembly when employing high throughput sequencing data alone, primarily due to the repetitive sequences within the genomes. Apparently, longer reads or paired end reads with larger insert size may help one go through the repetitive region and improve the assembly (Jiang et al. 2011). However, generation of long and accurate sequences has always been a challenge because of the high cost. Therefore, other genomic resources such as a genetic linkage map, BAC-based physical map, and BAC end sequences are needed to improve the whole genome sequence assembly, especially for scaffolding. Even with such genomic resources, assembly of complex genomes,

particularly for scaffolding, requires additional genomic resources. For instance, the best genome sequence assembly historically relied upon the availability of sequences generated by using the clone-by-clone sequencing strategy (Waterstonet al. 2002). Such successes come from: 1) The clone-by-clone approach allows generation and assembly of sequences to be divided into local assemblies within a clone, thereby reducing the complexity of the sequences and their assembly. This approach minimizes the presence of more than one interspersed repeat in each clone, thus a sequence that is repetitive from a genome-wide perspective may now be unique in a clone; 2) By using the minimal tiling path, the locally assembled sequences can be assembled into scaffolds based on physical maps. However, the historical clone-by-clone strategy is too expensive and too laborious with traditional sequencing. The aim of this study was to determine if physical map contig-specific sequences can be generated using the nextgen sequencing, and if such sequences can bring existing small genome sequence contigs into scaffolds corresponding to the physical map contigs.

Channel catfish is the predominant aquaculture species in the United States. The channel catfish genome is estimated to be 0.95Gb in size (Tiersch et al. 1990; www.genomesize.com). It is highly AT-rich, with 60.7% A+T (Xu et al. 2006). The catfish genome contains one main type of tandem repeats named Xba elements (Liu et al. 1998) and several types of dispersed repetitive elements, with the predominant dispersed repetitive elements being Tc1/mariner DNA transposons (Nandi et al. 2007). In addition, retrotransposons, LINE and SINE elements also exist in the catfish genome, with several SINE elements being well characterized such as the *Mermaid* and *Merman*

SINE elements (Kim et al. 2000; Xu et al. 2006; Liu et al. 2009; Liu et al. 2011). Short tandem repeats (microsatellites) are also highly abundant in the catfish genome with AC and AG being the most abundant types of microsatellites (Liu et al. 1999; Tan et al. 1999; Serapion et al. 2004). All these repeats within the catfish genome added more complexities to whole genome sequence assembly.

A pilot study for the catfish genome assembly was conducted (Jiang et al. 2011). In that study, we found that assembly of Illumina sequences was not very effective even with a reasonable fraction of sequences generated from 454 sequencing. Using only Illumina sequences, an initial draft genome assembly confirmed our conclusion, i.e., sequence contigs break at the repetitive sequences, most often at the Tc1 transposons or microsatellite sequences. As a result, use of only Illumina sequences resulted in a relatively large number of sequences of over 200,000 contigs (unpublished data).

A number of catfish genomic sequences are currently available to assist whole genome sequence assembly. These resources include genetic linkage maps (Waldbieser et al. 2001; Kucuktas et al. 2009; Ninwichian et al. 2012), physical maps (Quiniou et al. 2007; Xu et al. 2007), BAC end sequences (BES)(Xu et al. 2006; Liu et al. 2009), and integrated linkage and physical maps using BES-derived markers (Ninwichian et al. 2012). Use of these resources will definitely enhance whole genome sequence assembly. For instance, BES are useful to bring contigs into scaffolds because BAC end sequences from both ends are available (Xu et al. 2006; Liu et al. 2009) and they span an average distance of 161 kb (Wang et al. 2007). The BAC-end sequences associated BAC clones can relate them to the physical map; the integrated physical and linkage map can bring

the physical map contigs to linkage groups (chromosomes). However, as the anchor point to link the genome contigs, the BAC end sequences are not long because they were generated by single pass sequencing, with an average length of ~580 bp (Xu et al. 2006; Liu et al. 2009). In addition, the number of BAC end sequences is still limited, approximately 60,000 representing 25,677 paired reads. It is apparent that additional sequences specific for the physical map contigs will greatly enhance the anchorage ability of such sequences. The objective of this study was to generate more sequence tags from distinct physical contigs, namely physical map contig-specific sequences, to allow vast majority of genome sequence contigs to be anchored to physical map contigs, at a reasonable cost. Here we report a simple strategy for the production of physical map contig-specific sequences, their assembly, and their anchorage capability of random genome sequence contigs to the physical map contigs.

Results

Strategy for generating physical map contig-specific sequences

The strategy for generating physical map contig-specific sequences is illustrated in Figure 4. The strategy includes the following steps: 1) selection of clones representing a minimal tiling path (MTP) of each physical map contig; 2) Pooling of BAC DNA representing MTP of each contig; 3) Restriction digestion using two 4-bp cutter restriction endonucleases, separately. The two restriction enzymes, *Mse* I and *Bfa* I, were used for the digestion. The basic requirements of the two restriction enzymes were: a) they have different recognition sequences such that their restriction fragments are overlapping, allowing sequences to be assembled; and b) their restriction fragments

harbor compatible 5’-TA overhangs such that the restriction fragments can be ligated to the adaptor sequences with a 5’-TA overhang; 4) Ligation of specific bar-coded adaptors to fragments generated from each physical map contig, separately; 5) PCR amplification of the specific bar-coded fragments using barcoded PCR primers for fragments generated from each physical map contig; and 6) Illumina sequencing of the PCR-amplified fragments, followed by decoding of the sequences to specific physical map contigs using specific bar coding of the adaptors and the PCR primers (Figure 4).

Barcoding Tags design

In this strategy, the physical map contig-specific sequences were designated by a two-dimensional tagging strategy using both the adaptor and the PCR primer to reduce the total number of bar codes required for the 1,824 physical map contigs. For instance, if only a single bar-code is used, the labeling of the 1824 physical map contigs would require 1,824 different bar codes. When the two-dimensional tagging strategy was used, 38 different adaptors and 48 different PCR primers were used (total of only 86 sequence tags), which allowed a combination of 38 x 48 bar codes, i.e., 1,824.

Each adaptor was designed to include a 5 base pair specific bar code immediately adjacent to the 5’-TA-overhang, proximal to the fragments for ligation, and a 13 base pair common sequence for PCR. Each PCR primer was designed to include 5 base of common sequence as the shield sequence, then 5 base of specific bar-coded sequence, followed by a 13 base sequence complementary to the 13 base common sequences of the adaptors. Thus, the adaptors were 18 bp long, and the PCR primers were 23 bases long

(Table 7).

All 1,824 pooled BAC DNAs were arrayed into two dimension 38(row) x48(column) using 20 96-well plates, in which the row represented one sets of tags, A_i ,where i = 1, 2, 3, …38; and the column represented another set of tags, P_j, where j = 1, 2, 3, …48. As such, each pool of PCR products represents fragments derived from a single physical map contig with A_i and P_j at the ends. The adaptor and primer sequences are presented in Table 7.

The selection of the minimal tiling path

Two physical maps were published (Quiniou et al. 2007; Xu et al. 2007). One was constructed using the BAC library CHORI 212 (Wang et al. 2007) with 3,307 contigs, and the other was constructed using BAC library from a gynogenetic channel catfish (Quiniou et al. 2003) with 1,782 contigs (Quiniou et al. 2007). We have merged the two physical maps together (unpublished) with 1,824 physical contigs that include the BAC clones from the CHORI 212 library. A total of 6,701 BAC clones representing the minimal tilling path from 1,824physical map contigs were selected and cultured for BAC DNA preparation. On average, ~ 4 BAC clones were selected in each physical map contig.

Generation of physical map contig-specific sequences

Following the strategy described above, all PCR products were pooled together and sequenced using Illumina HiSeq 2000. Figure 5 shows the overall processing with the

raw data. A total of 334,381,996 100-bp paired-end reads were generated (Table 8). Low quality reads (Q < 20) were filtered and reads shorter than 20-bp were discarded, with CLC Genomics Workbench 5.5. After trimming, a total of 328,229,917 high quality reads were used for *de novo* assembly (Table 8), by using ABySS 1.3.0 with a k-value of 55. Of the 328,229,917 reads, there are 94,111,841 reads attached with the physical map contig-specific tags. After assembly, a total of 315,277 assembled contigs and 57,545,833 singletons were identified as the physical map contig-specific sequences, and assigned to each physical contig based on its attached specific tags.

Anchoring the draft genome sequence contigs to the physical map contigs

After the identification of physical map contig-specific sequences in each physical map contig, the barcoding tags were first trimmed as they were not part of the genome sequences from catfish. After trimming, the clean physical map contig-specific sequences represent the randomly distributed sequences in each physical contig, and serve as the anchor point to link the catfish draft genome sequence contigs (unpublished data) to the physical map contigs.

In order to determine how many genome sequence contigs can be anchored to the physical map contigs by the newly generated physical map contig-specific sequences (PMCSS) (contigs plus singletons), BAC end sequences, with or without PMCSS were used to BLAST the catfish whole genome sequence contigs. As shown in Table 9, with just the existing BAC end sequences, 27,770 whole genome sequence contigs had significant hits to the BES. However, this number was drastically enhanced with the

PMCSS, bringing the number of whole genome sequence contigs with significant hits to 156,457, i.e. 61% of whole genome sequence contigs had significant hits to the BES plus PMCSS, as compared to only 11% when only BES were used. In terms of genome sequence contig length, an even greater % of the whole genome sequence contigs were anchored by using the newly generated PMCSS. Only 26% of the whole genome sequence length was anchored by using BES only, and when the PMCSS were also used, over 79% of the whole genome sequence assembly were anchored (Table 9).

We also assessed the impact of PMCSS on scaffolding by examination of the genes that can be anchored to the whole genome sequence contigs. When only BES was used to anchor the whole genome sequence contigs, a total of 6,732 genes were identified within the whole genome sequence contigs with hits to BES. This number was drastically enhanced, to 16,680 genes (Table 9), when the PMCSS were also used, confirming strong anchorage capability of the PMCSS.

We previously reported the integration of the catfish physical map with its linkage map by using BAC end sequences-derived microsatellites (Ninwichian et al. 2012). The total physical lengths of all the physical map contigs that were genetically mapped to the linkage groups (chromosomes) were approximately 162 Mb. By using also the PMCSS, a total of 421 Mb of genome sequences can now be anchored to linkage groups (chromosomes). With the total length of the draft genome assembled whole genome sequences being 766 Mb, the PMCSS allowed 55% of the physical genome to be anchored to chromosome-scale scaffolds. The total number of genes that could be anchored to linkage groups were increased from 5,489 genes to 12,423 genes when the

PMCSS were used (Table 9).

Discussion

Assembly of whole genome sequences using next generation sequencing short reads is a great challenge. Such a challenge becomes even a greater challenge for teleost fish genomes because the presence of highly abundant interspersed repetitive elements such as Tc1 transposons, and long tracts of simple sequence repeats. As a result, whole genome sequence assemblies reflect the status of repetitive elements with contigs broken at the repetitive sequences, resulting in two unwanted consequences: 1) the segmented genome assembly with a large number of contigs; and 2) the repetitive elements stay unassembled such that the completeness of the genome assembly is reduced. However, for many fish species, the most significant issue for practical applications is the ability to scaffold the relatively short genome sequence contigs to large, chromosome-scale scaffolds. We have previously generated genomic resources for scaffolding such as BAC end sequences (Xu et al. 2006; Liu et al. 2009). However, these BAC end sequences were able to only scaffold approximately 21% of the genome to chromosome-scale scaffolds. Additional resources for scaffolding are greatly needed. This project intended to generate sequences specific to physical map contigs to provide greater scaffolding capacity.

Here we report a simple strategy for the generation of such physical map contig-specific sequences using a two-dimensional tagging strategy. This strategy is very simple. It is based on a two-dimensional specific sequence bar coding, one bar code on

the adaptors linked to each physical map contig fragments, and the other linked to PCR primers used to amplify fragments from each physical map contigs. This two dimensional design allowed the reduction of total number of bar codes. For instance, to differentiate the 1,824 physical map contigs, a total of at least 1,824 bar codes are needed. Synthesis of 1,824 primers would cost much more. By using the two dimensional design, a total of 86 primers were used, significantly reducing the cost without any loss of the differentiating power.

Rather than using a random shearing approach for breaking DNA into small segments, we used two 4-bp cutters that had the same 5'-overhang but different recognition sequence. This would allow overlapping fragments to be generated such that the sequences can be assembled. The restriction digestion eases the complexities for adaptor ligation.

The most tedious step is the DNA preparation, adaptor ligation and PCR amplification step for each of the physical map contigs because each physical map contig is treated as a separate sample. In our case, this involves 1,824 physical map contigs, thus 1,824 separate samples for DNA isolation, adaptor ligation, and PCR amplification. Considering how much information this project provides, such a tedious step is still worthwhile. A total of over 330 million reads were obtained, with over 28.7% harboring the sequence tags specific to each of the physical map contigs, thus allowing generation of 94,111,841 tagged sequences. One nice thing is the ability to assemble these sequences, allowing longer and greater percentage of sequences to be assigned to physical map contigs (Table 8).

We made the assessment of how these physical map contig-specific sequences affected the scaffolding capabilities. Several parameters were used including the number of whole genome sequence contigs that can be anchored to the physical map, the physical lengths of the whole genome sequence contigs that can be anchored to the physical map, and the number of genes that can be anchored to the physical map. With all these indicators, the physical map contig-specific sequences provided a high ability of enhancing the scaffolding capability (Table 9).

However, we realize that with just the physical map contig-specific sequences generated here, the sequences are anchored into 1,824 sets of sequences belonging to the physical map contigs. We cannot yet resolve the sequence stacking without additional information such as scaffolding using mate-paired reads of various sizes. Apparently, the ability to generate larger sequence contigs and the ability to place sequence contigs into linear scaffolds using mate-paired reads with various insert lengths are all very important for the whole genome sequence assembly, and such genomic resources are being generated. Nonetheless, these physical map contig-specific sequences will provide another level of scaffolding capability for the whole genome assembly and annotation of the catfish genome.

In the absence of a well-assembled whole genome sequence assembly, the physical map contig-specific sequences can greatly enhance the comparative genomics studies. For instance, most genes located within a physical map contig can now be readily identified. All anchored genome sequence contigs within a physical map contig can be used to search against Uniprot database, generating the genes that are located within the

physical map contig (Zhang et al. In press). Using all the annotated genes in all the physical map contigs, along with map integration information (Ninwichian et al. 2012), will greatly benefit the genome-wide comparative analysis of catfish. As shown in Table 9, if using BES in each physical contig alone as the anchor point to link genome contigs and physical contigs, there are 27,770 genome contigs anchored, resulting in 6,732 unique genes annotated, of which 5,489 genes can be anchored to the chromosome-scale scaffolds in linkage map. Addition of the physical map contig-specific sequences as anchor points, the number of anchored genome contigs increased to 156,457, resulting in 16,680 unique genes annotated, of which 12,423 genes can be anchored to the chromosome-scale scaffolds. Consequently, this would greatly improve the genome-wide comparative analysis of catfish. However, as discussed above, sequence stacking, which would also leads to gene stacking, has yet to be resolved with additional genomic resources before a thorough whole genome comparative map can be constructed for catfish, but a forest view of the whole genome, even with sequence stacking, still provide useful genome information.

Conclusions

In this study, we developed a strategy to generate physical map contig-specific sequences in an economical way, by using a two-dimensional tagging strategy. Analysis of the physical map contig-specific sequences indicated that such sequences can add significant scaffolding capabilities, serving as a useful resource for whole genome sequence assembly and comparative genomic studies. However, such physical map

contig-specific sequences cannot yet resolve sequence stackings, but rather group random genome sequence contigs into physical map contigs.

Materials and Methods

BAC clone culture and BAC DNA isolation

The catfish BAC clones from a minimum tiling path covering each physical contig from the CHORI-212 BAC library (Wang et al. 2007) were selected. The BAC DNA isolation was conducted as previously described (Xu et al. 2006), with modifications. Briefly, BAC clones were transferred from 384-well plates to a 96-well culture block, which contained 1.5 ml of 2X YT medium with 12.5 μg/ml chloramphenicol and grown at 37ºC overnight at 300 rpm. The block was centrifuged at 2000 x*g* for 10 min in an Eppendorf 5804R bench top centrifuge to collect bacteria. The culture supernatant was decanted and the block was inverted and tapped gently on paper towels to remove remaining liquid. BAC DNA was isolated using the Perfectprep™ BAC 96 kit (Eppendorf North America, Westbury, NY) according to the manufacturer's specifications.

Physical map contig-specific tag design

All 1,824 physical contigs were arranged into two-dimension 38(row) x 48(column) table. Instead of designing 1,824 tags for each physical contig, 38 adaptors, $A_{i\ (i=1,2,\ldots..38),}$ and 48 primers, $P_{j\ (j=1,2,\ldots48)}$ with distinct barcodes, were designed. A 5-bp distinct barcode was designed for each adaptor and primer. The combination of A_i and P_j served as a distinct tag for each physical contig. The detailed sequences of primers and adaptors

are shown in Table 7.

Sample preparation and sequencing

An amount of 100 ng BAC DNA from each physical contig was digested with 4-bp restriction enzymes *Mse* I and *Bfa* I, respectively. The two enzymes recognize different 4-bp sites but produce same 5'-TA overhangs. The digestion was conducted following the manufacturer's specifications, with modifications. Briefly, 100 ng BAC DNA was digested using 1 U of *Mse*I / *Bfa* I in a final volume of 5 µl reaction. The digestion mixture was incubated at 37 ℃ for 3 hours, and then inactivated at 80 ℃ for 20 minutes. After digestion, 0.5 µl [5 µM] in-house-designed adaptor was added for ligation using T4 ligase at 4 ℃ overnight . 1 U T4 ligase was used for each reaction in a final volume of 10 µl.

One microliter ligation product was used as template for amplification with the following reaction cocktail: 1 µl 10X buffer, 1 U polymerase, 0.4 µl $MgCl_2$ [50 mM], 0.8 µl dNTPs [2.5 mM], 2µl [5 µM] in-house-designed primers. Reactions were conducted in a thermocycler with the following thermal profile: 72 ℃ for 2 minutes, denaturing at 94 ℃ for 3 minutes, followed by 30 cycles of 94 ℃ for 30 seconds, 58 ℃ for 1 minutes and 72 ℃ for 3 minutes. A final extension was performed at 72 ℃ for 15 minutes to complete the PCR. Every single reaction was checked by electrophoresis on a 1% agarose gel and documented with a gel documentation system (Bio-Rad, Hercules, CA). All PCR products were pooled together and purified using the Qiaquick PCR Purification kit (Qiagen). A total of 1 µg high quality PCR products were sent to

Genomic Services Lab at HudsonAlpha Institute for Biotechnology (Huntsville, AL) for sequencing using Illumina HiSeq 2000.

Identification of physical map contig-specific tags

CLC Genomics Workbench 5.5 (CLC Bio, Cambridge, MA) was used to remove the BAC vector and low quality reads, with quality score limit of 0.01 (Q20). Illumina 100-bp PE reads shorter than 20 bp were discarded. *De novo* assembly was conducted by using ABySS 1.3.0, with a k-mer value of 55. A script was used to search the physical map contig-specific tags in all Illumina reads as well as the assembled contigs. After physical map contig-specific sequence identification and assignment to each physical contig, the specific tags were then trimmed using CLC Genomics Workbench 5.5.

Identification of anchored genome contigs

The clean physical map contig-specific sequences were used as queries to search against the draft catfish genome sequences by BLASTN, with an E-value cutoff of 1e-20. The query sequences with multiple hits of genome contigs were considered non-specific and discarded. Only the query sequences with a single hits and identity value greater than 98% were considered as specific sequences. The corresponding genome contig hits were retrieved to use as queries to BLASTX search against the Uniprot database for gene annotation, with an E-value cutoff of 1e-10.

Table 7. The sequences of adaptors and primers.

	Forward	**Reverse**
A_1	GACGATGAGTCCGCTCCA	TATGGAGCGGACTCAT
A_2	GACGATGAGTCCGTTCGA	TATCGAACGGACTCAT
A_3	GACGATGAGTCCGAATGA	TATCATTCGGACTCAT
A_4	GACGATGAGTCCGGTTGA	TATCAACCGGACTCAT
A_5	GACGATGAGTCCGCAGTA	TATACTGCGGACTCAT
A_6	GACGATGAGTCCGCTCAC	TAGTGAGCGGACTCAT
A_7	GACGATGAGTCCGAATAC	TAGTATTCGGACTCAT
A_8	GACGATGAGTCCGTCTAC	TAGTAGACGGACTCAT
A_9	GACGATGAGTCCGACACC	TAGGTGTCGGACTCAT
A_{10}	GACGATGAGTCCGCTACC	TAGGTAGCGGACTCAT
A_{11}	GACGATGAGTCCGATCCC	TAGGGATCGGACTCAT
A_{12}	GACGATGAGTCCGAGGCC	TAGGCCTCGGACTCAT
A_{13}	GACGATGAGTCCGTTAGC	TAGCTAACGGACTCAT
A_{14}	GACGATGAGTCCGCAGGC	TAGCCTGCGGACTCAT
A_{15}	GACGATGAGTCCGGTGGC	TAGCCACCGGACTCAT
A_{16}	GACGATGAGTCCGCCATC	TAGATGGCGGACTCAT
A_{17}	GACGATGAGTCCGTACTC	TAGAGTACGGACTCAT
A_{18}	GACGATGAGTCCGCAAAG	TACTTTGCGGACTCAT
A_{19}	GACGATGAGTCCGTCAAG	TACTTGACGGACTCAT
A_{20}	GACGATGAGTCCGACGAG	TACTCGTCGGACTCAT
A_{21}	GACGATGAGTCCGGTGAG	TACTCACCGGACTCAT
A_{22}	GACGATGAGTCCGGTACG	TACGTACCGGACTCAT
A_{23}	GACGATGAGTCCGCACCG	TACGGTGCGGACTCAT
A_{24}	GACGATGAGTCCGTTCCG	TACGGAACGGACTCAT
A_{25}	GACGATGAGTCCGACTCG	TACGAGTCGGACTCAT
A_{26}	GACGATGAGTCCGCATGG	TACCATGCGGACTCAT
A_{27}	GACGATGAGTCCGATCTG	TACAGATCGGACTCAT
A_{28}	GACGATGAGTCCGTAGTG	TACACTACGGACTCAT
A_{29}	GACGATGAGTCCGGGTTG	TACAACCCGGACTCAT
A_{30}	GACGATGAGTCCGTTTTG	TACAAAACGGACTCAT
A_{31}	GACGATGAGTCCGGGAAT	TAATTCCCGGACTCAT
A_{32}	GACGATGAGTCCGGTCAT	TAATGACCGGACTCAT
A_{33}	GACGATGAGTCCGTGACT	TAAGTCACGGACTCAT
A_{34}	GACGATGAGTCCGCTGCT	TAAGCAGCGGACTCAT
A_{35}	GACGATGAGTCCGCTAGT	TAACTAGCGGACTCAT
A_{36}	GACGATGAGTCCGACTGT	TAACAGTCGGACTCAT
A_{37}	GACGATGAGTCCGCACTT	TAAAGTGCGGACTCAT
A_{38}	GACGATGAGTCCGAGCTT	TAAAGCTCGGACTCAT
P_1	GAGTTTGAACGACGATGAGTCCG	GAGTTTGAACGACGATGAGTCCG
P_2	GAGTTACAAGGACGATGAGTCCG	GAGTTACAAGGACGATGAGTCCG
P_3	GAGTTGGAATGACGATGAGTCCG	GAGTTGGAATGACGATGAGTCCG
P_4	GAGTTCTACCGACGATGAGTCCG	GAGTTCTACCGACGATGAGTCCG

$\mathbf{P_5}$	GAGTTGAACTGACGATGAGTCCG	GAGTTGAACTGACGATGAGTCCG
$\mathbf{P_6}$	GAGTTTGACTGACGATGAGTCCG	GAGTTTGACTGACGATGAGTCCG
$\mathbf{P_7}$	GAGTTCAAGTGACGATGAGTCCG	GAGTTCAAGTGACGATGAGTCCG
$\mathbf{P_8}$	GAGTTTCAGTGACGATGAGTCCG	GAGTTTCAGTGACGATGAGTCCG
$\mathbf{P_9}$	GAGTTTTATAGACGATGAGTCCG	GAGTTTTATAGACGATGAGTCCG
$\mathbf{P_{10}}$	GAGTTTGATCGACGATGAGTCCG	GAGTTTGATCGACGATGAGTCCG
$\mathbf{P_{11}}$	GAGTTTAATGGACGATGAGTCCG	GAGTTTAATGGACGATGAGTCCG
$\mathbf{P_{12}}$	GAGTTATATTGACGATGAGTCCG	GAGTTATATTGACGATGAGTCCG
$\mathbf{P_{13}}$	GAGTTGCCACGACGATGAGTCCG	GAGTTGCCACGACGATGAGTCCG
$\mathbf{P_{14}}$	GAGTTTTCACGACGATGAGTCCG	GAGTTTTCACGACGATGAGTCCG
$\mathbf{P_{15}}$	GAGTTTACAGGACGATGAGTCCG	GAGTTTACAGGACGATGAGTCCG
$\mathbf{P_{16}}$	GAGTTGGCAGGACGATGAGTCCG	GAGTTGGCAGGACGATGAGTCCG
$\mathbf{P_{17}}$	GAGTTGTCCTGACGATGAGTCCG	GAGTTGTCCTGACGATGAGTCCG
$\mathbf{P_{18}}$	GAGTTACCGAGACGATGAGTCCG	GAGTTACCGAGACGATGAGTCCG
$\mathbf{P_{19}}$	GAGTTTTCGAGACGATGAGTCCG	GAGTTTTCGAGACGATGAGTCCG
$\mathbf{P_{20}}$	GAGTTGACGGGACGATGAGTCCG	GAGTTGACGGGACGATGAGTCCG
$\mathbf{P_{21}}$	GAGTTCACTCGACGATGAGTCCG	GAGTTCACTCGACGATGAGTCCG
$\mathbf{P_{22}}$	GAGTTTCCTGGACGATGAGTCCG	GAGTTTCCTGGACGATGAGTCCG
$\mathbf{P_{23}}$	GAGTTATCTGGACGATGAGTCCG	GAGTTATCTGGACGATGAGTCCG
$\mathbf{P_{24}}$	GAGTTGGCTTGACGATGAGTCCG	GAGTTGGCTTGACGATGAGTCCG
$\mathbf{P_{25}}$	GAGTTCTGAGGACGATGAGTCCG	GAGTTCTGAGGACGATGAGTCCG
$\mathbf{P_{26}}$	GAGTTCCGATGACGATGAGTCCG	GAGTTCCGATGACGATGAGTCCG
$\mathbf{P_{27}}$	GAGTTGTGATGACGATGAGTCCG	GAGTTGTGATGACGATGAGTCCG
$\mathbf{P_{28}}$	GAGTTTAGCCGACGATGAGTCCG	GAGTTTAGCCGACGATGAGTCCG
$\mathbf{P_{29}}$	GAGTTCTGCTGACGATGAGTCCG	GAGTTCTGCTGACGATGAGTCCG
$\mathbf{P_{30}}$	GAGTTCGGGAGACGATGAGTCCG	GAGTTCGGGAGACGATGAGTCCG
$\mathbf{P_{31}}$	GAGTTCAGGCGACGATGAGTCCG	GAGTTCAGGCGACGATGAGTCCG
$\mathbf{P_{32}}$	GAGTTAGGGCGACGATGAGTCCG	GAGTTAGGGCGACGATGAGTCCG
$\mathbf{P_{33}}$	GAGTTTTGGCGACGATGAGTCCG	GAGTTTTGGCGACGATGAGTCCG
$\mathbf{P_{34}}$	GAGTTCTGTAGACGATGAGTCCG	GAGTTCTGTAGACGATGAGTCCG
$\mathbf{P_{35}}$	GAGTTCGGTGGACGATGAGTCCG	GAGTTCGGTGGACGATGAGTCCG
$\mathbf{P_{36}}$	GAGTTCCTAAGACGATGAGTCCG	GAGTTCCTAAGACGATGAGTCCG
$\mathbf{P_{37}}$	GAGTTAGTAAGACGATGAGTCCG	GAGTTAGTAAGACGATGAGTCCG
$\mathbf{P_{38}}$	GAGTTTCTACGACGATGAGTCCG	GAGTTTCTACGACGATGAGTCCG
$\mathbf{P_{39}}$	GAGTTGCTAGGACGATGAGTCCG	GAGTTGCTAGGACGATGAGTCCG
$\mathbf{P_{40}}$	GAGTTATTAGGACGATGAGTCCG	GAGTTATTAGGACGATGAGTCCG
$\mathbf{P_{41}}$	GAGTTGATATGACGATGAGTCCG	GAGTTGATATGACGATGAGTCCG
$\mathbf{P_{42}}$	GAGTTGGTCAGACGATGAGTCCG	GAGTTGGTCAGACGATGAGTCCG
$\mathbf{P_{43}}$	GAGTTAATCCGACGATGAGTCCG	GAGTTAATCCGACGATGAGTCCG
$\mathbf{P_{44}}$	GAGTTAATGAGACGATGAGTCCG	GAGTTAATGAGACGATGAGTCCG
$\mathbf{P_{45}}$	GAGTTGTTGGGACGATGAGTCCG	GAGTTGTTGGGACGATGAGTCCG
$\mathbf{P_{46}}$	GAGTTACTGTGACGATGAGTCCG	GAGTTACTGTGACGATGAGTCCG
$\mathbf{P_{47}}$	GAGTTGATTAGACGATGAGTCCG	GAGTTGATTAGACGATGAGTCCG
$\mathbf{P_{48}}$	GAGTTGTTTCGACGATGAGTCCG	GAGTTGTTTCGACGATGAGTCCG

Table 8. Summary of the physical map contig-specific sequences.

Total number of raw reads	334,381,996
numberof trimmed reads	328,229,917
Number of reads with tags	94,111,841
Number of assembled contigs	315,917
Numberof assembled contigs with tags	315,277
Average number of assembled sequence contigs with tags per physical map contig	173
Number of singleton reads with tags	57,545,833
Total number of tagged sequence contigs &singletons	57,861,110
Total number of BAC end sequences available for the 1,824 physical map contigs	42,616
Number of BAC end sequences from the 1,824 physical map contigs with hits from tagged sequences	31,809

Table 9. Comparison of effect on comparative genomics study with and without physical map contig-specific sequences (PMCSS).

	With BES only	**With BES &PMCSS**
Number of unique genome sequence contig hits	27,770	156,457
% of genome contig hits	11%	61%
Total length of genome contig hits	202 Mb	608 Mb
% of total length genome contig hits	26%	79%
Number of unique gene hits	6,732	16,680
Total length of genome sequences anchored into chromosome-scale scaffolds	162 Mb	421 Mb
% of total length genome sequences anchored into chromosome-scale scaffolds	21%	55%
Number of unique gene anchored into chromosome-scale scaffolds	5,489	12,423

Figure 4. Flow chart illustrating the physical map contig-specific fragment preparation. The minimal tilling path BAC clones from each physical contig were selected (highlighted). The pooled BAC DNA from each physical map contig was digested with two 4-bp restriction enzymes, *Mse* I and *Bfa* I, respectively. The digestion product was then ligated with in-house designed adaptors, followed by PCR using in-house designed primers. The combination of adaptor and primer formed a specific tag representing each physical contig ID. All PCR products with a physical map contig-specific tag then were pooled together, and sequenced using Illumina HiSeq 2000 platform.

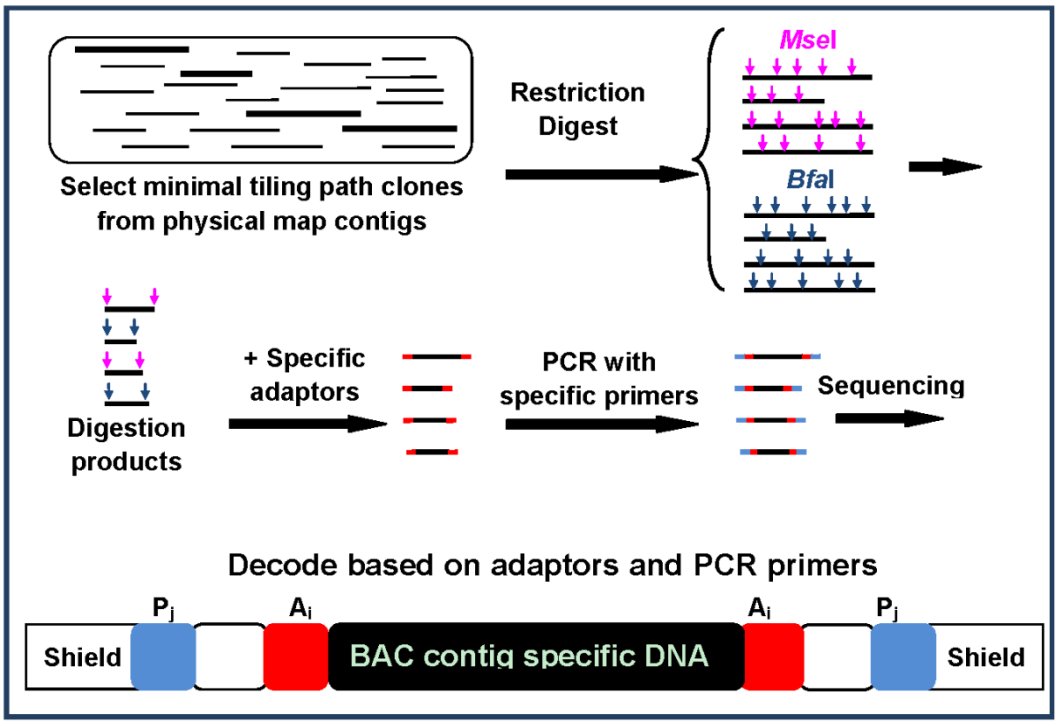
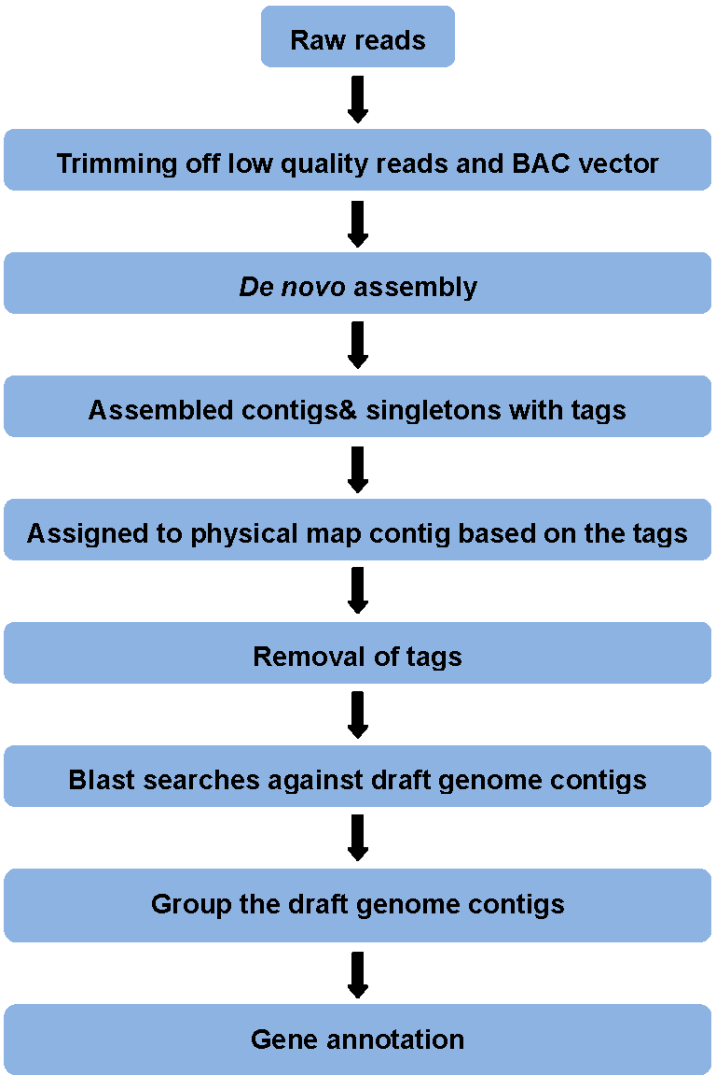


Figure 5. The workflow of data processing. The raw reads were first trimmed off the low quality reads (Q20) and BAC vectors. *De novo* assembly was then conducted with the filtered high quality reads. The assembled contigs with tag on it, plus singletons which have tag on it were then assigned to each physical contig, based on the specific tag. The tags were removed. The clean sequences were then used as queries to BLAST search against the draft catfish whole genome contigs. The targeted genome contigs were then retrieved and annotated.



V. GENOME-WIDE COMPARATIVE ANALYSIS OF CHANNEL CATFISH

Abstract

With the limitation of completely assembled whole genome sequences in most organisms, comparative mapping serves as a powerful tool to advance the genetic and genomic studies on non-model species, by transferring the genome sequences from the well-studied model species. Utilizing all existed genomic resources including genetic linkage map, physical map, BAC end sequences, physical map contig-specific sequences, and the draft genome sequences, comparative analysis of channel catfish with zebrafish were conducted in this study. Conserved microsyntenies were identified and the comparative map was established. The genome-wide comparative analysis of channel catfish was first reported here, which will greatly benefit understanding the genetic basis as well as the evolutionary role of catfish.

Introduction

As evolutionarily and ecologically important vertebrates, fishes not only provide a main source of animal protein to world populations, but also offer key answers to questions on human evolution (Boffelli et al. 2004). Although the importance of aquaculture is increasing, the genetic knowledge as well as genomic resources are still a limitation for most aquaculture species, even with the dramatic advanced next generation sequencing technologies in recent years.

Comparative map is a powerful tool to transfer the genomic information from one genome-sequence-available species to the species whose genome sequences is not available yet, allowing identification of potential syntenies, and therefore to understand the genome organization and rearrangement event and how the genome is remodeled during evolution. Moreover, the conserved synteny can facilitate the identification of heritable traits of interest by comparative QTL analysis directed in a targeted way by synteny conservation and associated gene content information (Chistiakov et al. 2008). Such methods were initially demonstrated by Fujiyama et al. (2002) for constructing a human-chimpanzee comparative map using chimpanzee BESs to hit against human genome sequences. Putative orthologs were identified between these two closely related species. Later on, the comparative mapping approach was extensively used in mammals, for instance, the construction of human-cattle, the human-horse, and the human-porcine comparative maps (Larkin et al. 2003; Leeb et al. 2006; Meyers et al. 2005). The success of the comparative mapping mainly depends on the high percentage of BLAST hits and/or high level of genome colinearity.

Several model fish genomes have been fully sequenced and assembled including zebrafish (*Danio rerio*), medaka (*Oryzias latipes*), fugu(*Tetraodon nigroviridis*) and three-spined stickleback (*Gasterosteus aculeatus*). The availability of these whole genome sequences results in extensive usage of comparative mapping in other fish species. For instance, a study on comparative genomics between six different teleost species, including *Tetraodon nigroviridis, Oryzias latipes, Gasterosteus aculateus, Sparus aurata, Dicentrarchus labrax,* and *Oreochromis spp*., have shown the syntenic

relationship between those species (Sarropoulou et al. 2008). Another comparative genomic analysis study on European sea bass has been reported, with five model teleosts, by using a combined AFLP and microsatellite linkage map (Chistiakov et al. 2008), which established a high level of conservation of synteny in comparison with stickleback. Soler et al. (2010) conducted the comparative analysis using tilapia BESs against the genome assemblies of stickleback, medaka and pufferfish, which allowed identification of homologies for each species, as well as the rearrangement breakpoint between those species.

Channel catfish, *Ictalurus punctatus*, is the predominant aquaculture species in the United States, as well as an important model species for the study of comparative immunology, reproductive physiology and toxicology (Wang et al. 2010). Considerable efforts have been made toward the genetic improvement of catfish, and a number of genomic resources have been developed, including BAC libraries (Quinious et al. 2003; Wang et al. 2007), BAC-based physical maps (Quinious et al. 2007; Xu et al. 2007), genetic linkage maps (Waldbieser et al. 2001; Liu et al. 2003; Kucuktas et al. 2009), a large number of ESTs (Li et al. 2007; Wang et al. 2010), over 1,700 unique full length cDNAs (Chen et al. 2010), over 60,000 BESs (Xu et al. 2006; Liu et al. 2009), and a large number of identified molecular markers such as microsatellites and single nucleotide polymorphism (SNP) (Wang et al. 2010; Liu et al. 2011). Without the completely-assembled whole genome sequences, several comparative analysis have been conducted on catfish with other fish species. Kucuktas et al. (2009) utilized a genetic linkage map using EST-based microsatellite and SNP markers to identify the

conserved syntenies between catfish and other model fish species. BESs serve as a valuable source as well for comparative analyses (Xu et al. 2006; Liu et al. 2009).

Catfish whole genome sequencing and assembly is ongoing. A draft catfish genome will greatly improve comparative analyses. In this study, a first genome-wide comparative map between catfish and zebrafish is reported, utilizing the catfish genetic linkage map, BAC-based physical map, BESs, physical mp contig-specific sequences, and the draft catfish genome sequences. The comparative analysis of the commercial fish to model fish species will gain benefits to improve the structural and functional genomics.

Materials and methods

Establishing chromosome-scale scaffolds

Similar to a previous pilot study described (Zhang et al. 2012 in press), this study started with the catfish genetic linkage map based on microsatellite markers that were derived from BAC end sequences (Ninwichian et at. 2012), the BAC-based catfish physical map (Xu et al. 2007), BAC end sequences (Xu et al. 2006; Liu et al. 2009), and physical-contig specific sequences generated in this dissertation study (Chapter IV), and the catfish draft genome sequences (Unpublished data). The basic concept for this study is utilizing BAC end sequences and physical map contig-specific sequences to anchor the draft catfish genome sequences for further analysis. When a BAC end sequence is mapped to a linkage group of the genetic linkage map, the entire BAC contig is then

mapped to this linkage group. BAC clones from the same BAC contig as the mapped BAC clone were identified through the examination of the catfish physical map. All available BAC end sequences and physical map contig-specific sequences within the physical contigs which were mapped to linkage group were collected, and used to BLAST against the catfish draft genome sequences, with an E-value cutoff of 1e-10. The top hits of the genome contigs were then filtered with a bit score value cutoff of 400 to ensure the identification of true homologous sequences (Zhang et al. 2012 in press).

Identification of gene in each linkage group and the homologous chromosome in zebrafish

Within each linkage group, all indentified genome contig sequences anchored by BAC end sequences and physical-contig specific sequences were used to remove repeat sequences using RepeatMasker (http://www.repeatmasker.org/). The repeat-masked genome contig sequences were then searched against ENSEMBL zebrafish protein database by conducting BLASTX searches.

The homologous chromosome and gene location in zebrafish for each catfish linkage group were obtained based on the results of above BLASTX searches, retrieved by using BioMart in ENSEMBL. The zebrafish chromosome with a high number of gene hits for each catfish linkage map as identified as the homologous chromosome.

Identification of putative conserved syntenies

As Zhang et al. (2012 in press) described in the pilot study, conserved syntenies

were identified based on the genetic positions of microsatellite markers derived from the BAC end sequences as well as the associated genes, with the location of homologous genes in zebrafish. The putative conserved syntenies in this study were identified only if a set of genes in a certain order located in the same linkage group in catfish and in the same chromosome in zebrafish. Microsyntenic blocks were identified based on genes included within BAC contigs of the catfish physical map and their locations on the zebrafish chromosome. The putative conserved microsyntenies were identified as segments of zebrafish chromosomes with a set of adjacent genes that are homologous to a set of adjacent genes in catfish that are reflected by their location within a single BAC contig.

MapChart (Voorrips, 2002) were used to construct the comparative map of catfish and zebrafish. The catfish genome contigs were anchored to the linkage map based on the BAC end sequences and physical map contig-specific sequences. The comparative map was constructed based on the position of genome contigs on each linkage group and the location of homologous gene in zebrafish homologous chromosome.

Results and discussions

Establishing chromosome-scale scaffolds

Without well-assembled whole genome sequences, this study started with microsatellite markers derived from BAC end sequences in the catfish genetic linkage map. As shown in Table 10, there are 2,099 microsatellite markers from 931 physical

contigs located in 29 linkage groups. A total of 32,500 BAC end sequences (BESs) were identified in all 931 physical contigs. BLASTN researches were then conducted by using all 32,500 BESs plus 57,861,110 physcial-contig specific reads to search against the catfish draft genome sequences, resulting 435 Mb (55%) genome contigs can be mapped into catfish linkage map.

To determine the gene content associated with the anchored genome contigs, BLASTX searches were then conducted, using the genome contig sequences as query to search against the ENSEMBL zebrafish protein database. The BLASTX searches resulted in 14,032 unique genes. Since the locations of physical contigs were known in the linkage group based on the position of mapped BESs-associated marker, the BLASTX analysis allowed anchor the 14,032 genes to each linkage group, and consequently established the chromosome-scale scaffolds for comparative analysis.

Identification of homologous chromosomes for each linkage group

All genes identified in all catfish linkage groups were searched against zebrafish protein database. The number of homologous genes in each chromosome/linkage group between two species is shown in Figure 6. The zebrafish chromosome with the largest gene hits was considered to be the homologous chromosome. A roughly one-to-one linear relationship was observed, without major chromosome rearrangements. However, some catfish linkage groups might be created by split or fusion of zebrafish chromosomes because the unequal chromosome number between catfish and zebrafish, e.g. linkage group 6, 10 and 29 in catfish have been shown (Figure 6) to merge into

zebrafish chromosome 1, linkage group 14 and 26 merged into chromosome 5, linkage 8, 20 and 27 merged into chromosome 7, etc.

Constructing the comparative map

To gain detailed insight of evolutionary relationship between two species, a comparative map was constructed between 29 catfish linkage groups with their homologous chromosomes in zebrafish. Only gene sequences were used for this comparative analysis because gene sequences are highly conserved in the teleost genome while sequences from intergenic regions are more divergent. The position of physical contigs can be mapped in the linkage group based on the location of BES-associated marker, however, the positions and order of genes within each physical contig is not clear yet because the catfish whole genome sequence is not well assembled, and the resolution of the genetic linkage map that positioned the physical contigs was not high enough to put the catfish genes in a linear order. Many genes are “stacked”, but the gene positions and orders can be compared at the chromosomal level, ignore the stacked genes.

Figure 7 shows the comparative map of catfish linkage groups to the corresponding zebrafish chromosomes. The results indicated a high level of genome conservation existing between catfish and zebrafish, however, numerous chromosome breaks, shuffling and rearrangement are highly likely during evolution while intensive conservation existed on a small segmental scale, e.g. the identified conserved microsyntenies. Similar results were observed in the previous studies (Wang et al. 2007;

Kucuktas et al. 2009). Catfish has 29 chromosomes (linkage groups) while zebrafish has only 25 chromosomes, thus, in some cases, one zebrafish chromosome is equivalent to more than one catfish linkage groups, e. g. linkage groups6, 10 and 29, linkage groups 14 and 26, linkage groups 8 and 27, share the same homologous chromosome 1, 5, and 7, respectively (Figure 7). It is interesting that the homologous chromosome segments in one catfish linkage group are distributed into two or three zebrafish chromosomes, such as linkage group 1, 8, 9, 12, 20 and 24 (Figure 7), which suggests that these chromosomes might be large in order to contain genes from several catfish chromosome equivalents or the significant chromosomal rearrangements occurred during evolution.

Conclusion

In this study, integrated genomic resources including a genetic linkage map, BAC-based physical map, BAC end sequences, physical map contig-specific sequences and catfish draft genome sequences were used to conduct the genome-wide comparative analysis with zebrafish. Strong evidence supported a high level of conservation between catfish and zebrafish, however, inter-chromosomal and intra-chromosomal rearrangement, shuffling and fusion occurred in catfish during evolution.

Table 10. Summary information of identification of genes in all catfish linkage groups by using BAC end sequences and physical map contig-specific reads.

Number of BAC-associated markers in linkage map	2,099
Number of BAC contigs containing the BAC-associated markers	931
Number of all BAC-end sequences (BES) from mapped BAC contigs	32,500
Total number of BESs with significant hits to draft genome	32,365
Total number of physical map contig-specific reads	57,861,110
Total length of draft genome contigs anchored on linkage map	435 Mb (55%)
Number of unique genes with mapped genome contig hits	14,032

Figure 6. Syntenic relationship between catfish and zebrafish genome. X-axis, chromosomes of zebrafish; Y-axis, linkage group of catfish. Number in the cell is the number of homologous genes between two species. Numbers larger than 60 is highlighted.

	Zebrafish chromosome (C)																								
Catfish Linkage group (L)		C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	C21	C22	C23	C24	C25
	L10	290	0	8	1	2	2	0	0	17	1	1	0	1	3	0	1	1	1	4	2	0	1	0	1	7
	L29	71	1	4	1	13	0	33	1	0	1	1	1	33	7	0	1	0	2	1	2	2	3	3	0	0
	L2	1	205	1	2	1	22	4	7	0	4	2	0	0	1	28	13	0	2	3	0	1	8	0	4	0
	L3	8	2	347	16	15	1	1	3	3	3	3	16	3	41	18	27	1	3	11	38	2	5	7	8	3
	L23	17	0	12	221	1	6	5	1	1	0	0	3	4	6	1	3	0	1	9	0	2	11	0	0	8
	L14	2	7	10	6	256	2	3	2	3	1	2	4	0	1	4	1	1	0	2	0	2	2	2	1	2
	L26	0	2	1	1	299	1	1	1	1	7	0	0	0	1	0	1	0	0	2	0	7	3	0	1	1
	L7	4	14	11	2	2	304	6	5	9	3	1	0	15	4	2	3	7	5	2	2	0	2	7	3	2
	L12	2	1	27	4	19	202	12	1	10	2	0	13	2	0	1	13	1	7	197	1	2	1	1	15	0
	L8	3	206	8	2	0	19	269	3	3	8	4	1	7	2	3	2	2	0	3	1	26	3	1	8	0
	L27	1	5	3	2	0	2	170	1	12	1	24	0	4	0	7	0	0	1	1	6	0	4	1	3	4
	L11	3	8	6	11	7	6	2	453	1	4	6	4	1	4	4	1	13	0	4	7	3	2	18	11	4
	L17	36	6	25	22	4	19	31	2	376	21	0	13	10	1	6	4	5	6	4	12	5	7	4	1	1
	L25	8	7	7	7	11	2	7	5	2	386	17	12	19	11	8	1	2	7	2	6	0	1	1	0	5
	L21	0	3	0	2	2	1	4	2	0	3	275	0	0	0	2	1	0	0	2	0	0	51	0	0	2
	L5	6	2	33	10	17	3	6	3	15	1	4	454	5	0	1	2	9	4	4	1	8	38	2	4	2
	L6	71	9	7	22	24	4	4	4	1	21	1	15	426	30	1	4	4	11	3	3	14	4	3	3	2
	L9	1	0	1	7	3	0	4	1	0	4	10	1	4	290	0	2	16	9	2	0	3	2	2	1	145
	L22	22	45	5	6	9	0	7	6	4	10	2	23	5	2	373	3	2	0	3	2	3	2	1	2	2
	L1	1	18	8	7	1	3	1	9	2	1	3	5	12	11	3	402	1	4	14	8	2	3	7	197	1
	L16	1	4	6	1	11	0	20	6	3	8	1	7	12	3	8	6	454	13	8	37	1	2	2	14	18
	L4	3	9	1	7	0	2	12	3	1	8	7	2	11	12	3	4	2	338	15	5	0	7	5	7	6
	L24	3	0	131	4	4	1	1	3	3	0	35	3	1	0	4	7	2	0	232	1	2	0	3	1	0
	L28	1	1	1	8	2	1	7	3	1	0	2	0	8	0	0	10	4	3	0	415	0	5	0	2	1
	L13	1	14	6	3	8	2	5	12	1	2	6	1	2	2	14	10	4	17	7	2	447	1	5	1	0
	L20	2	33	6	3	5	6	131	2	1	5	109	2	3	1	7	3	1	2	2	7	2	214	4	0	4
	L15	4	17	14	6	2	5	1	1	1	1	10	2	0	1	3	3	1	0	6	13	0	0	457	5	2
	L19	0	7	4	3	4	0	6	0	1	1	1	1	0	1	10	106	1	6	17	1	1	2	0	150	0
	L18	6	4	1	11	1	18	13	1	0	3	0	0	1	59	2	1	2	0	8	1	3	1	2	0	281

Figure 7. The comparative map of catfish with zebrafish.

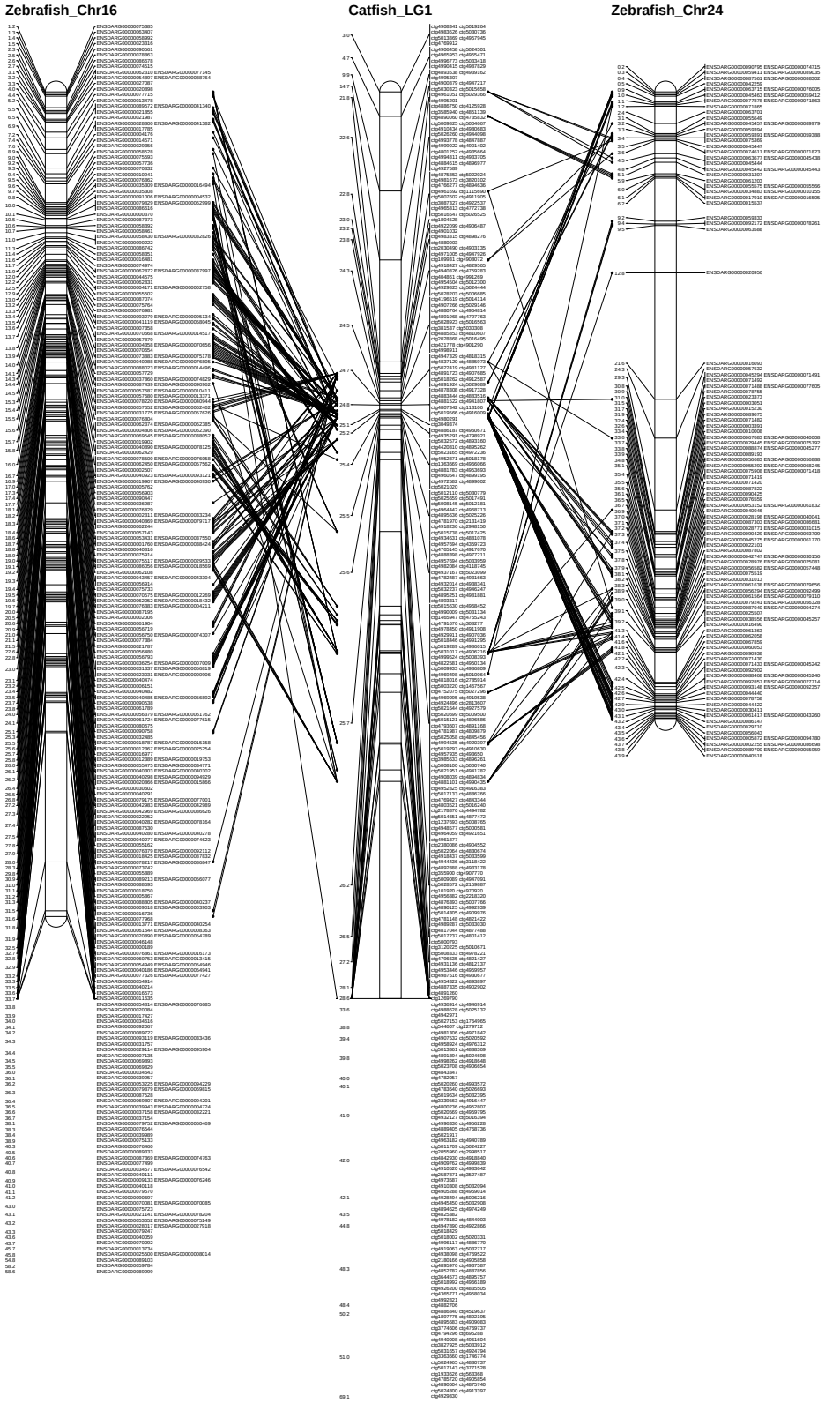


Figure 7 continued.

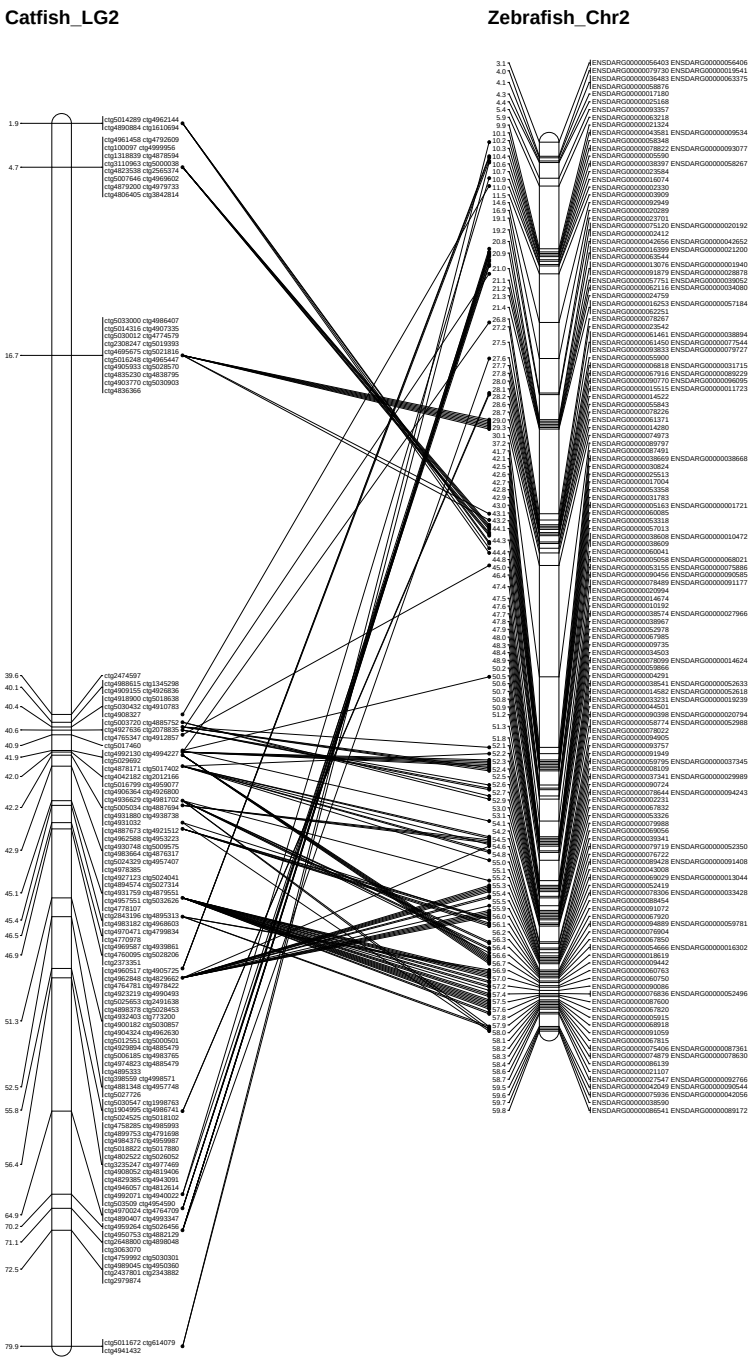


Figure 7 continued.

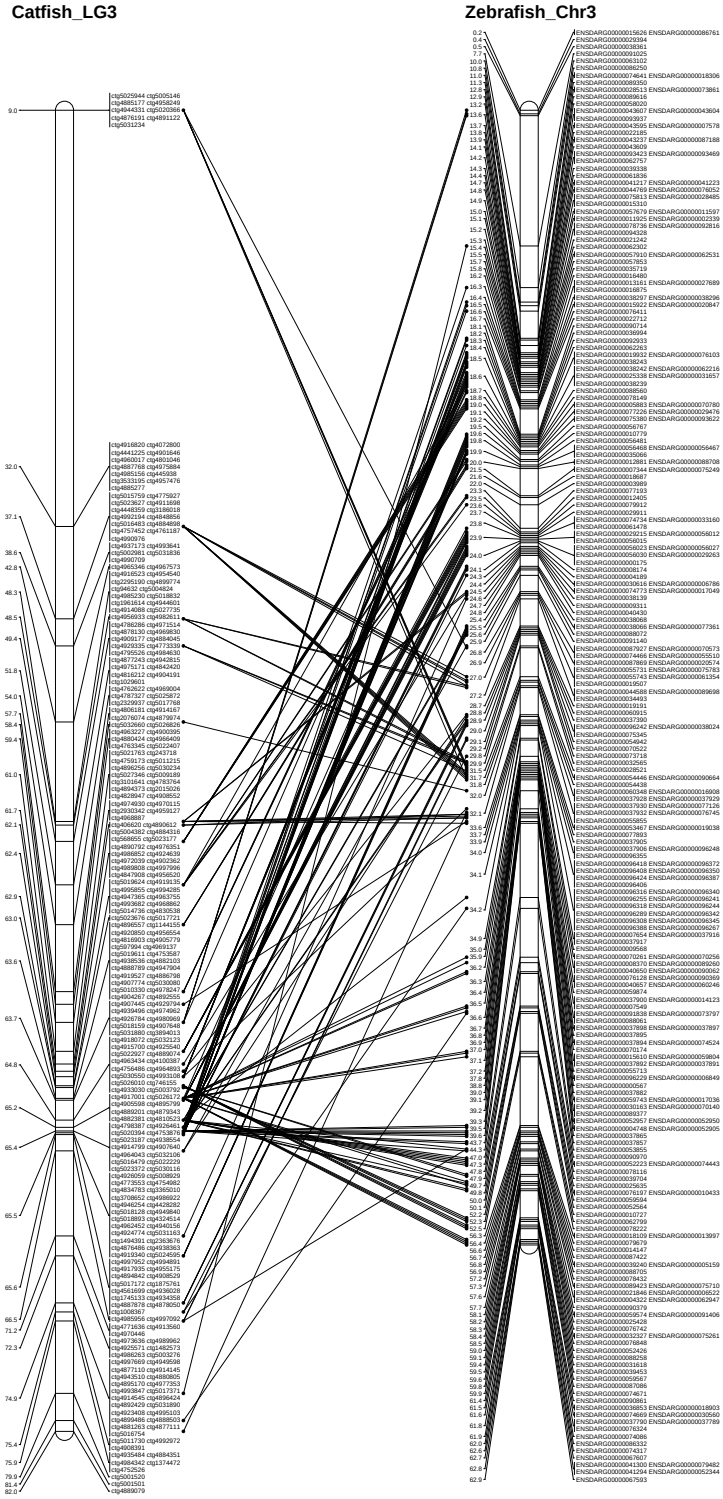


Figure 7 continued.

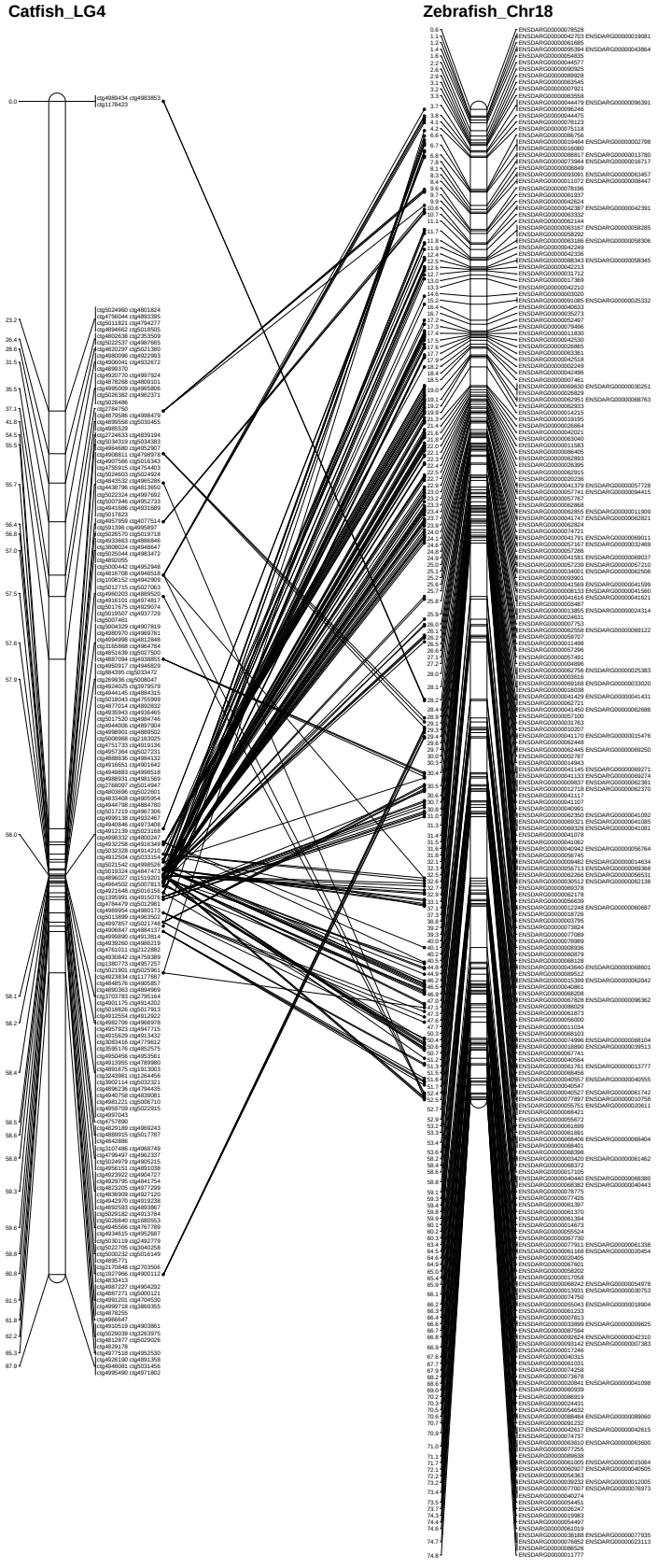


Figure 7 continued.

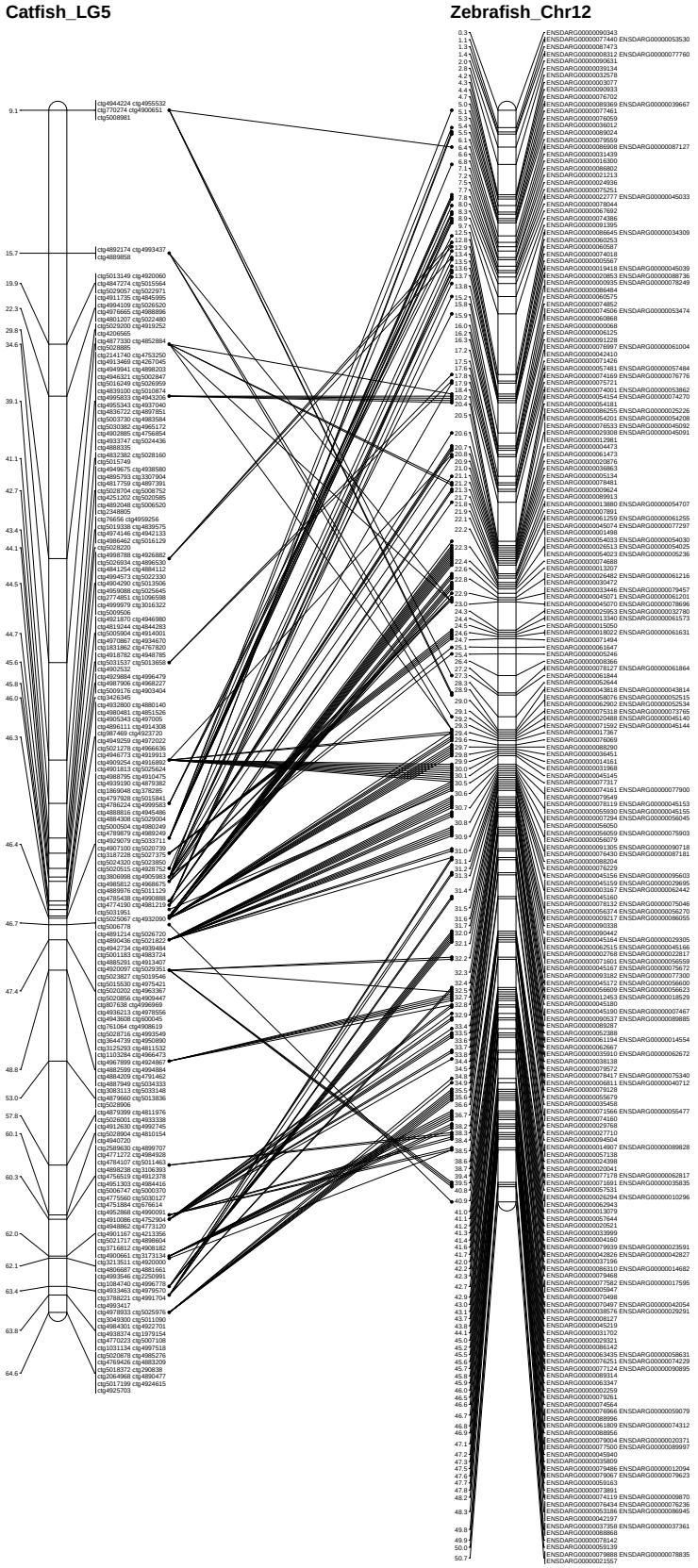


Figure 7 continued.

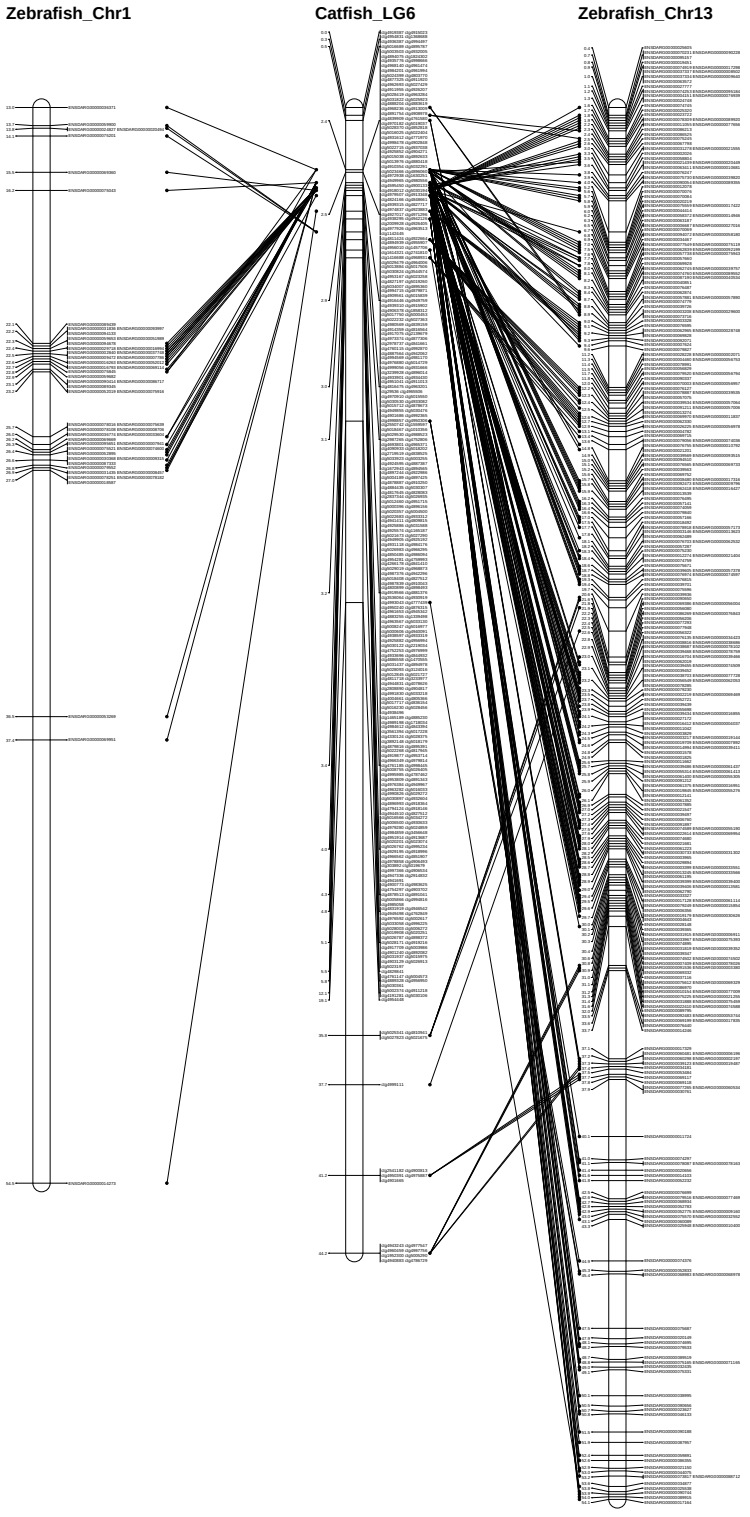


Figure 7 continued.

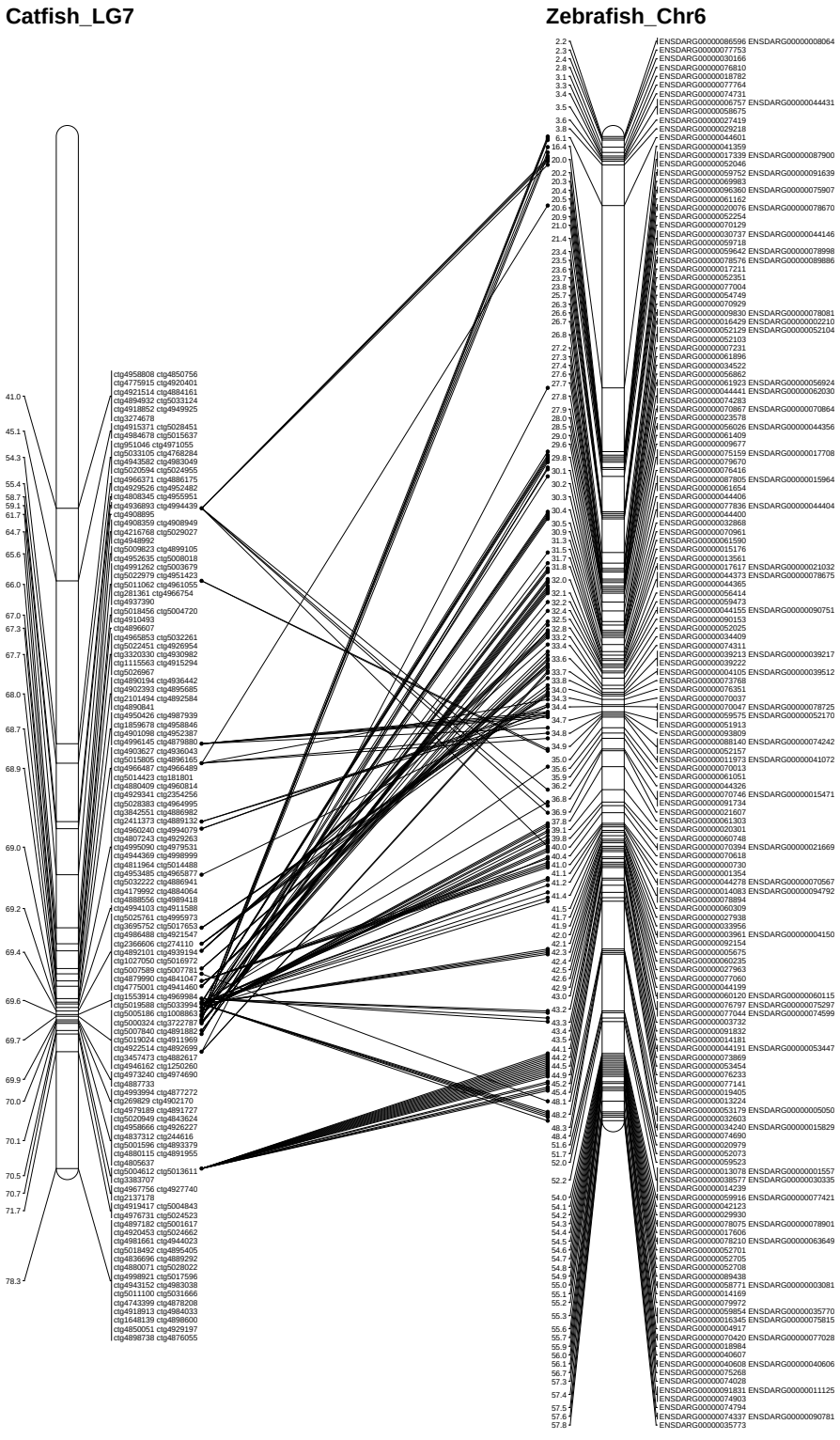


Figure 7 continued.

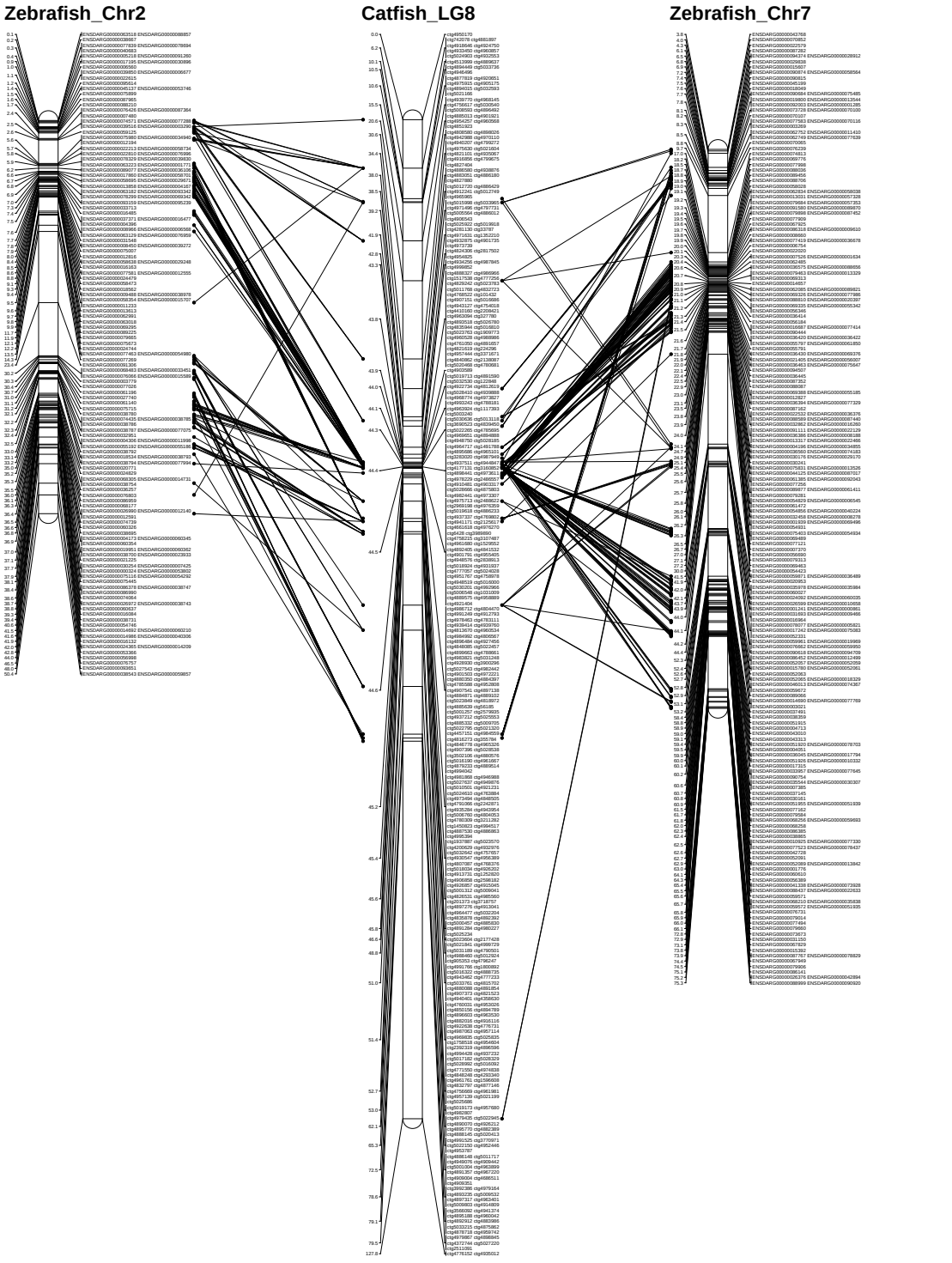


Figure 7 continued.

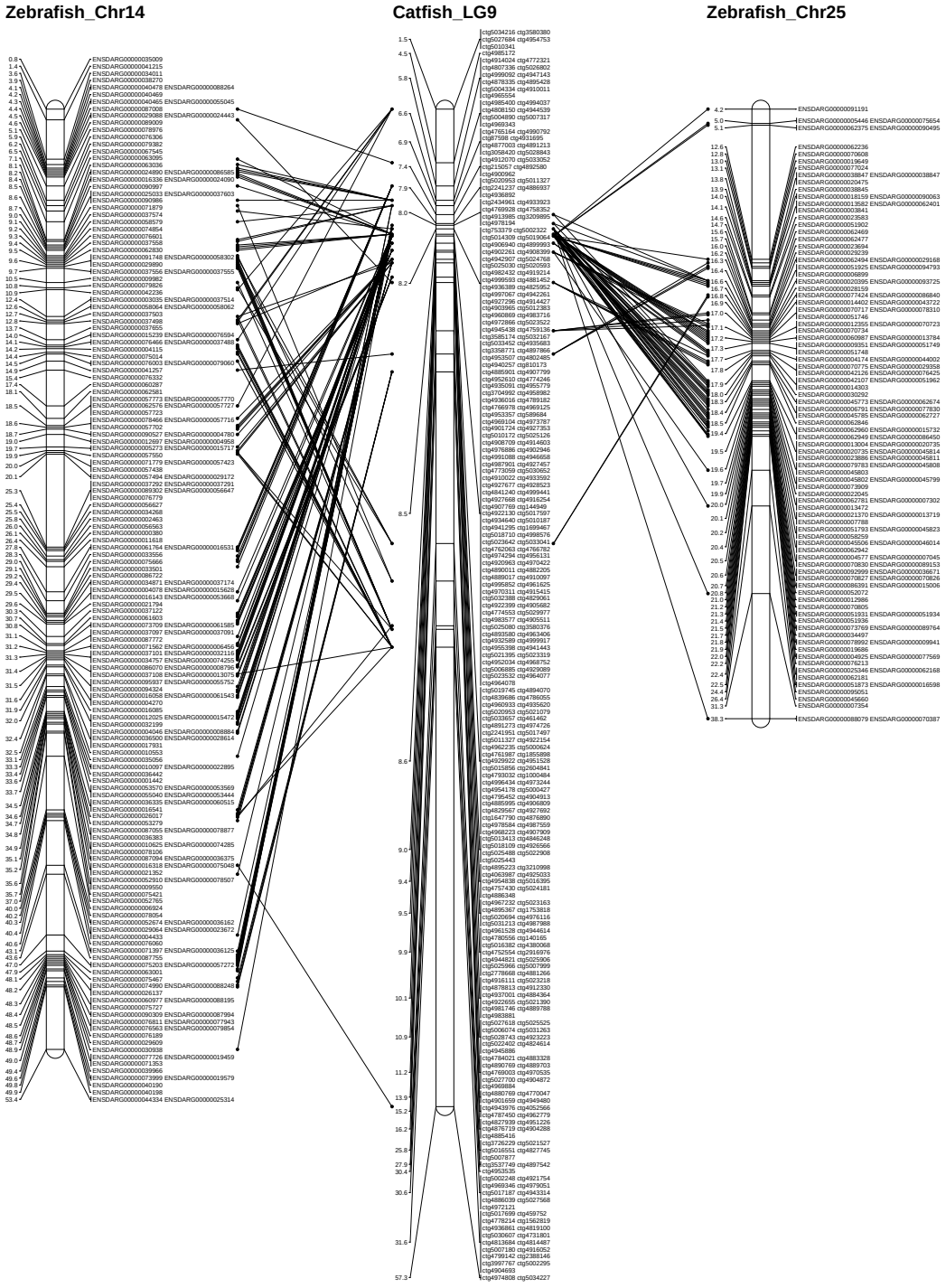


Figure 7 continued.

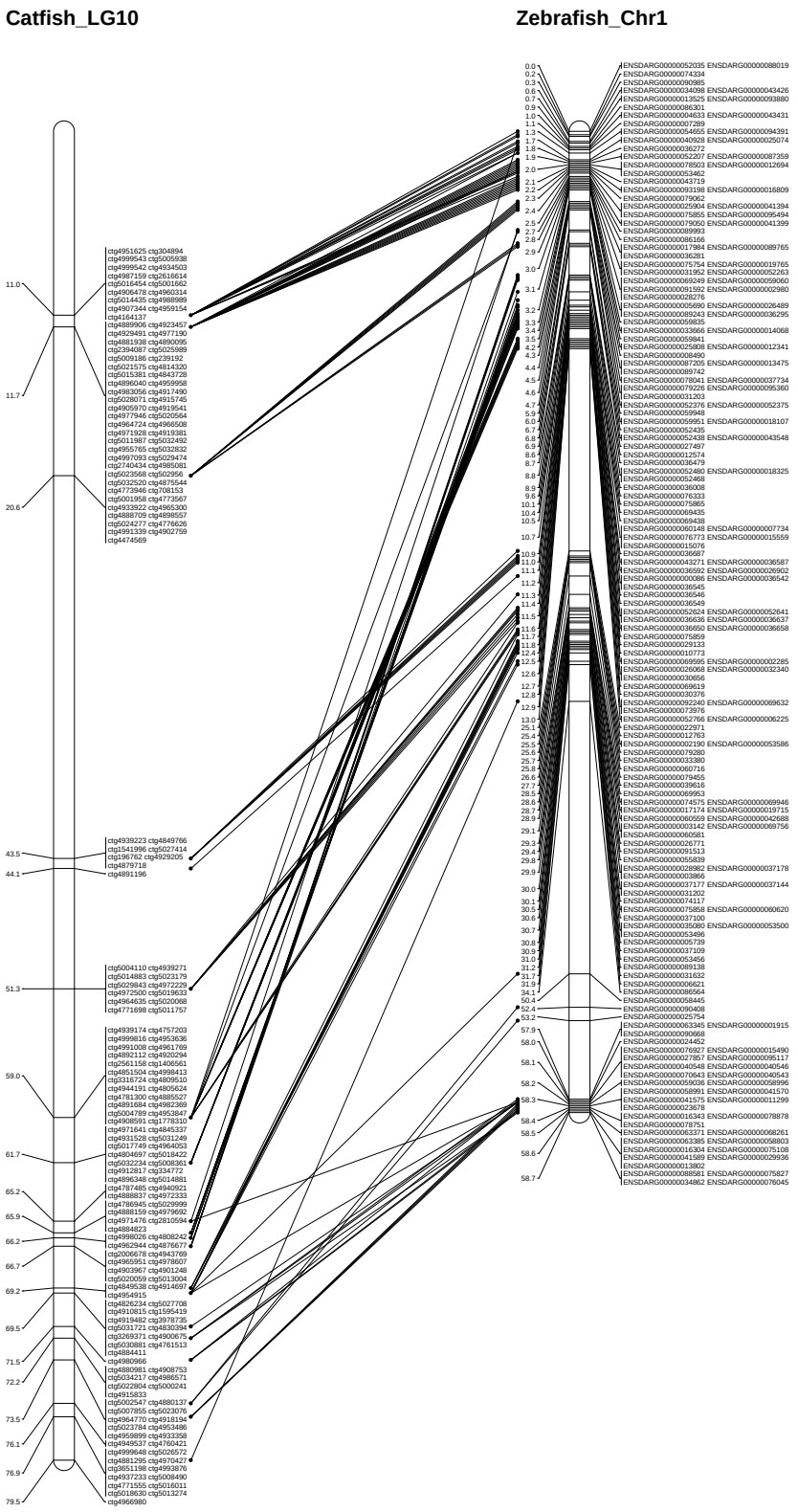


Figure 7 continued.

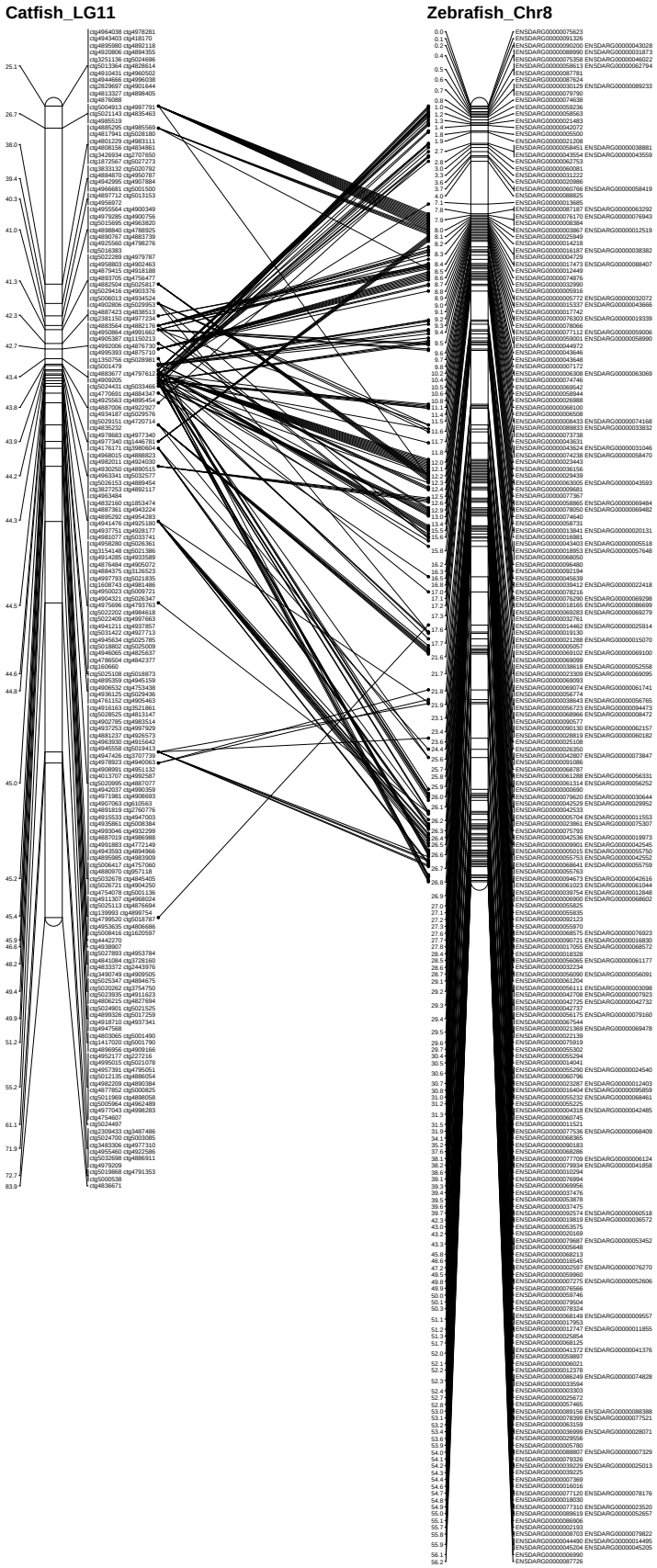


Figure 7 continued.

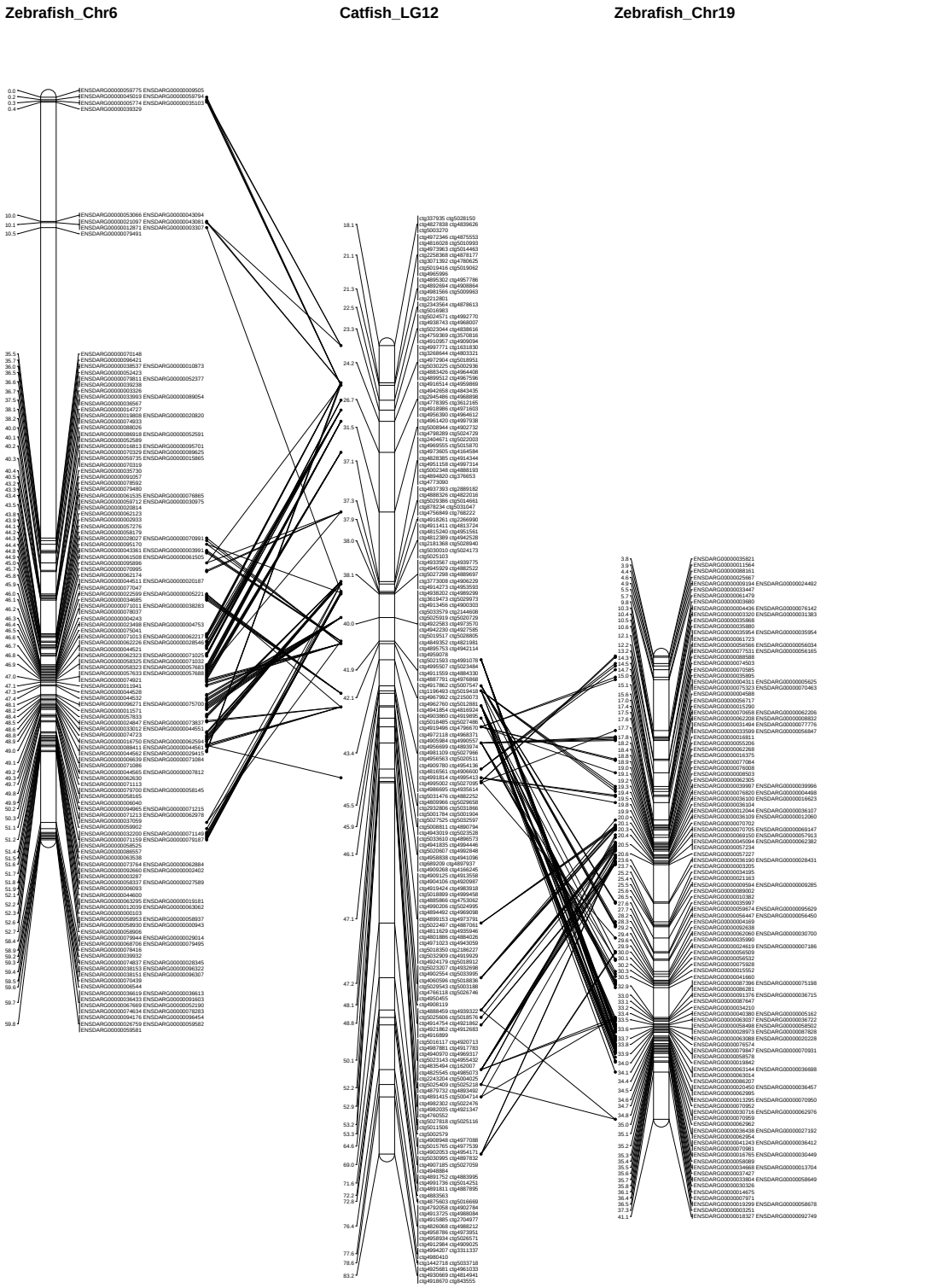


Figure 7 continued.

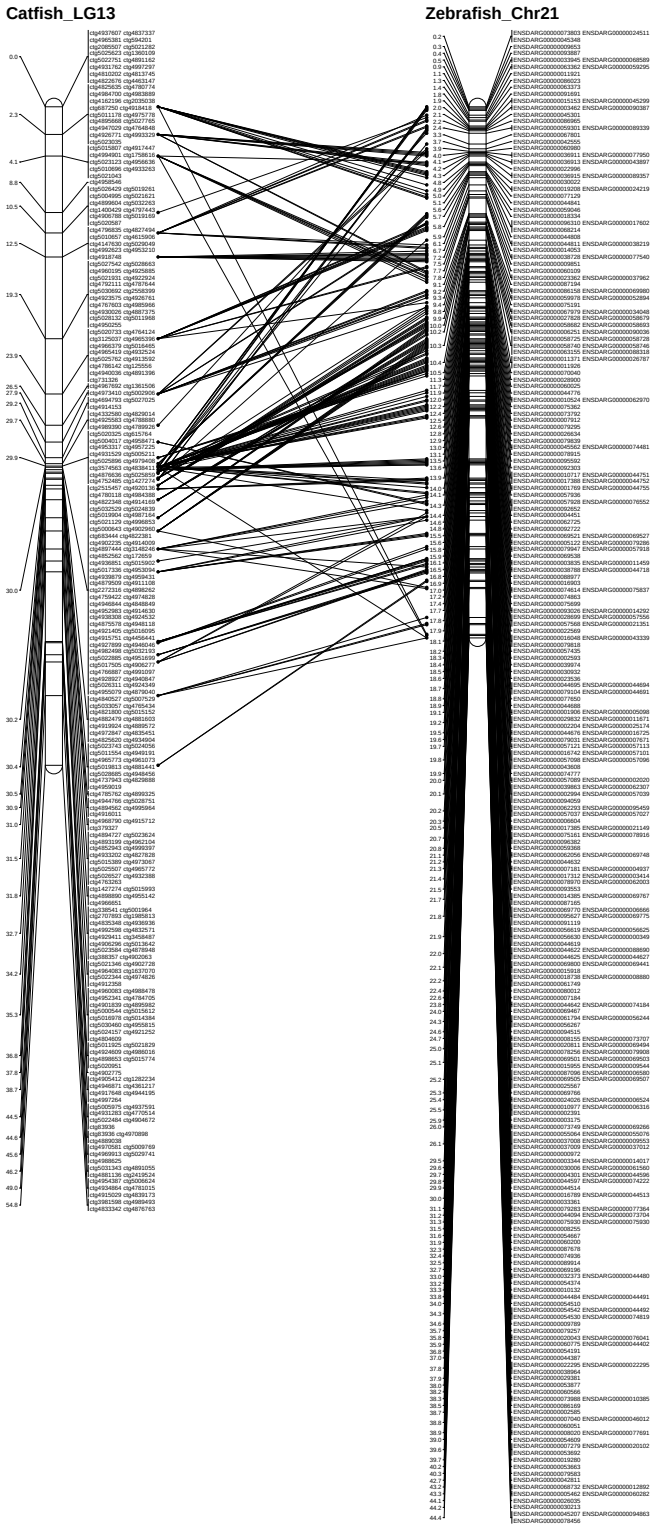


Figure 7 continued.

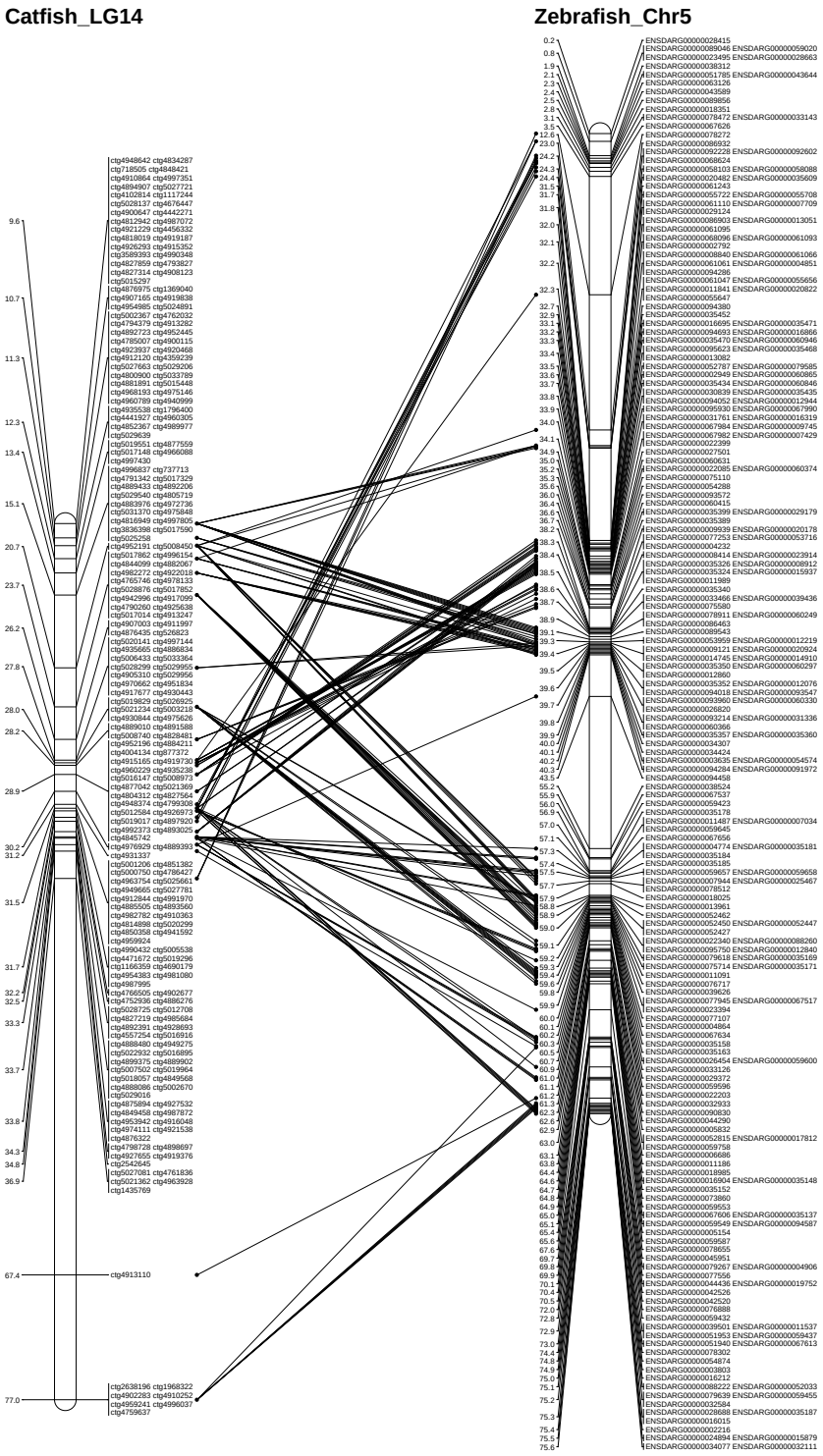


Figure 7 continued.

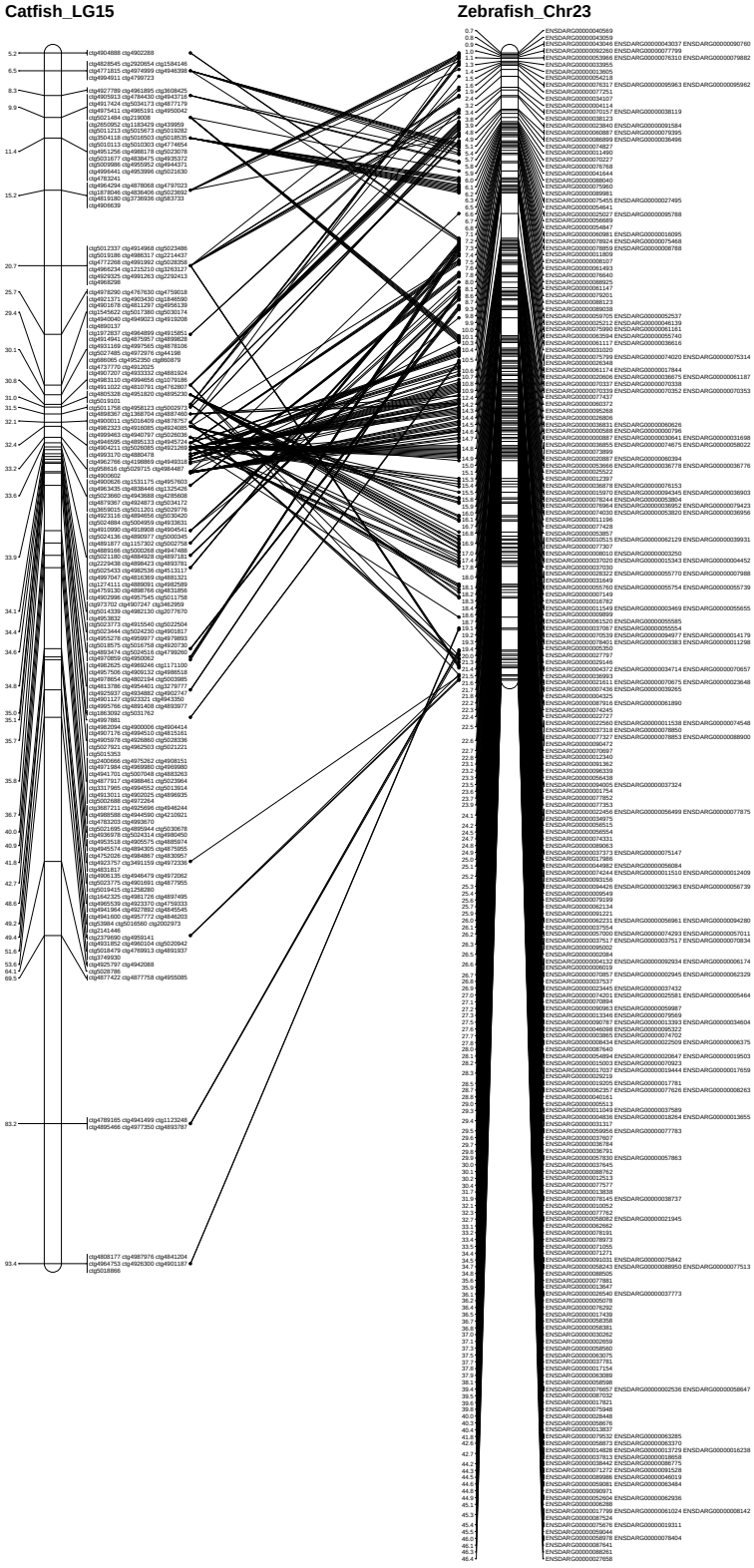


Figure 7 continued.

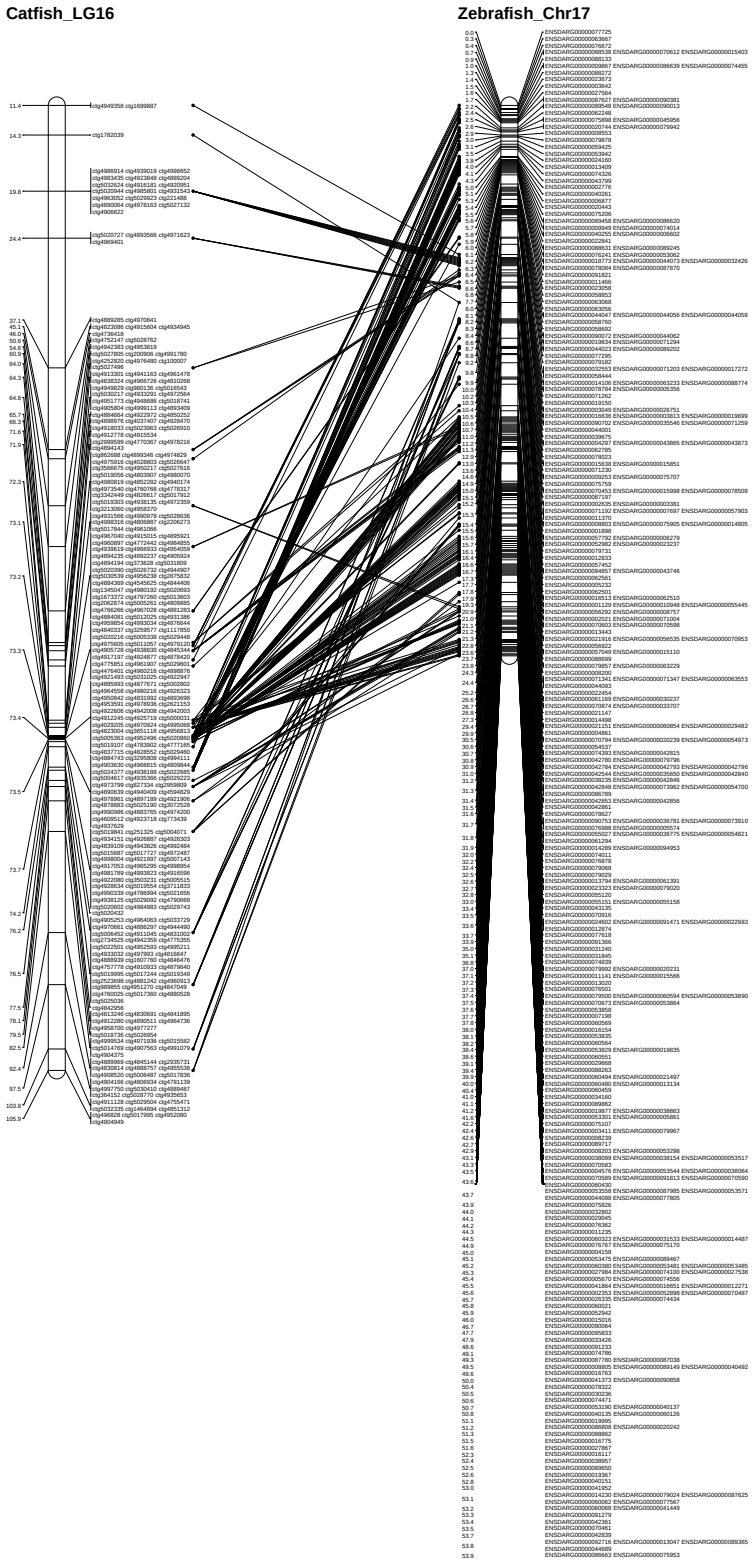


Figure 7 continued.

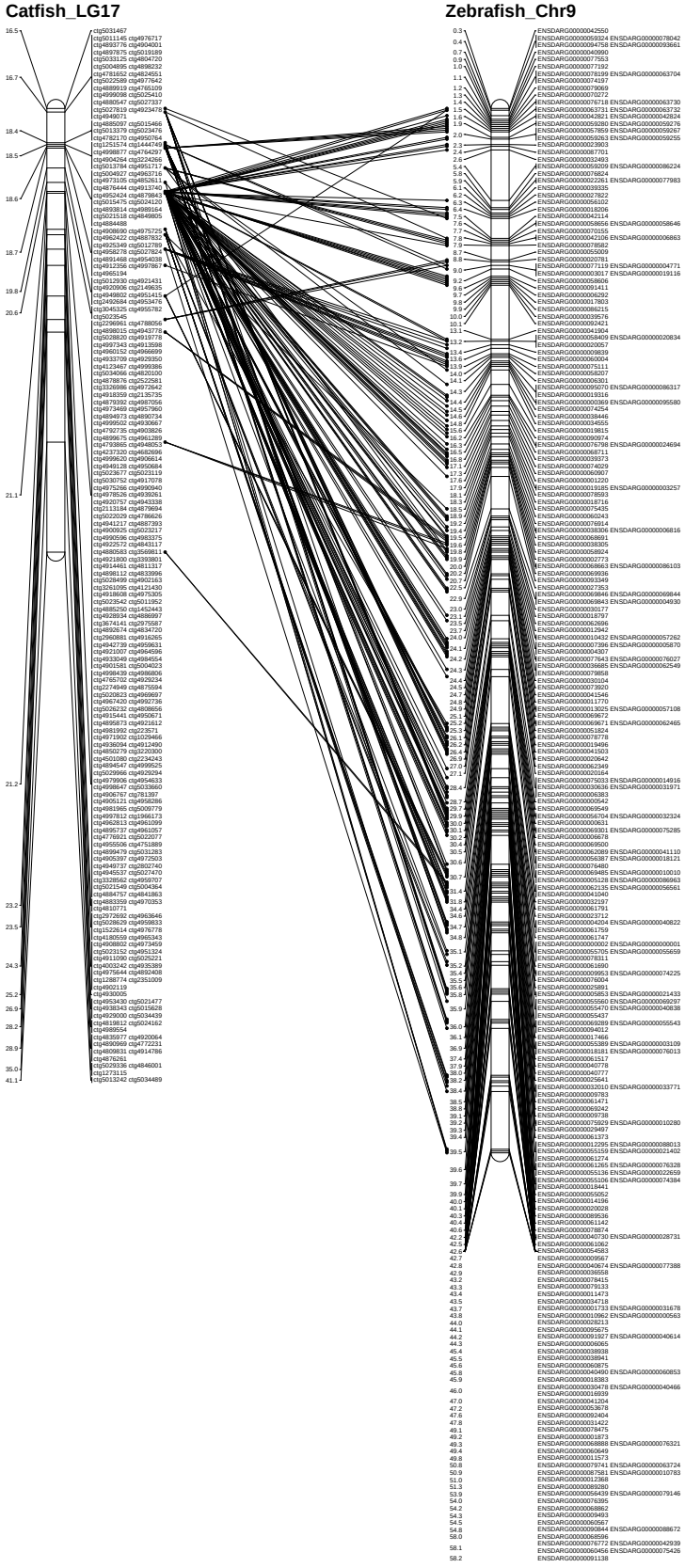


Figure 7 continued.

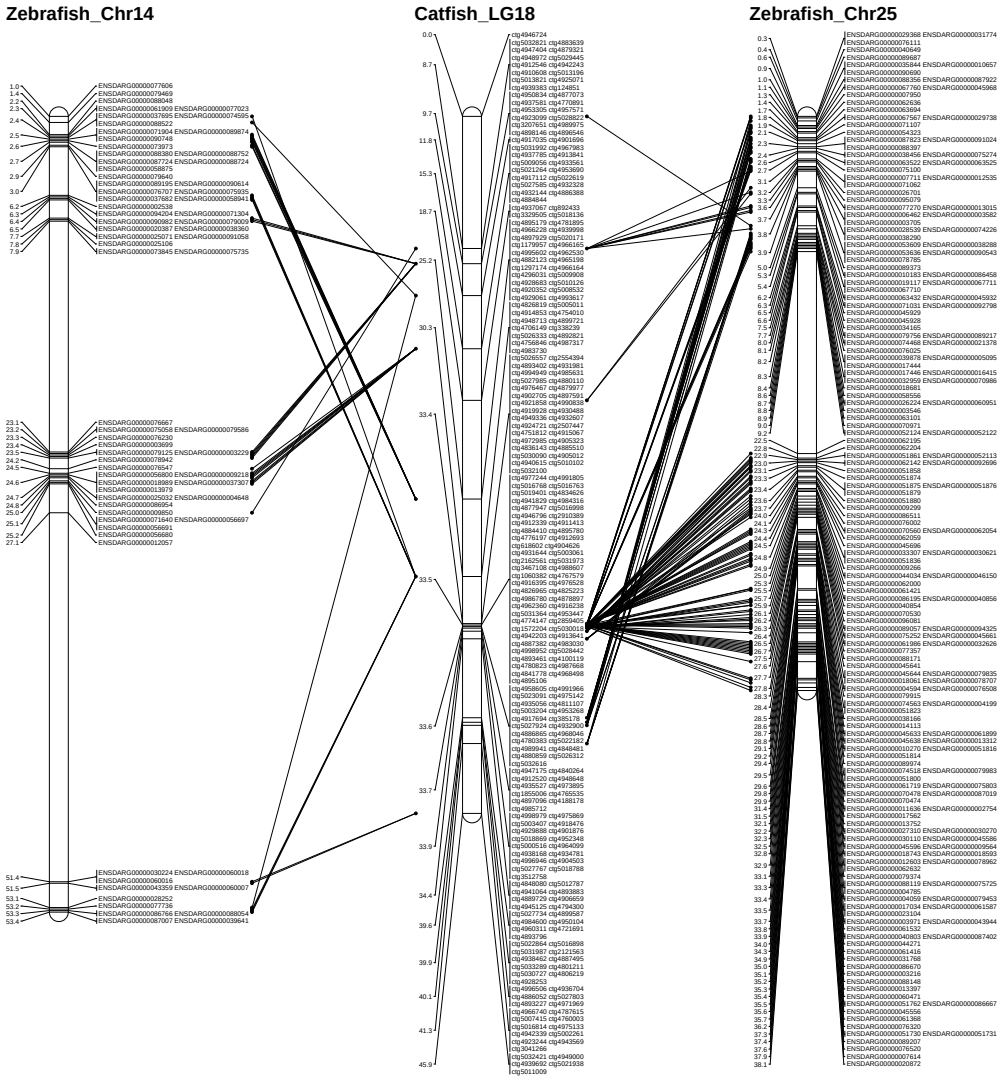


Figure 7 continued.

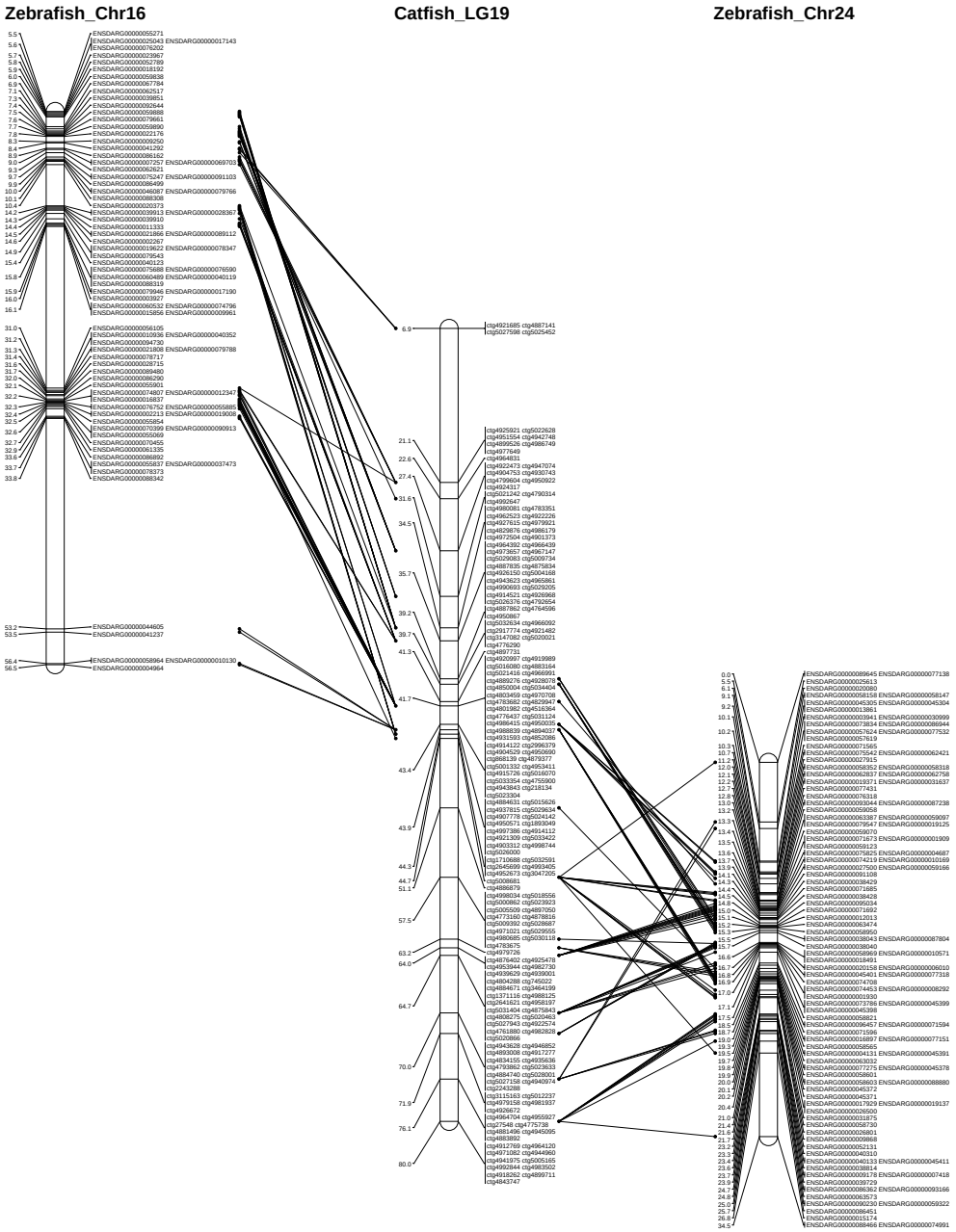


Figure 7 continued.

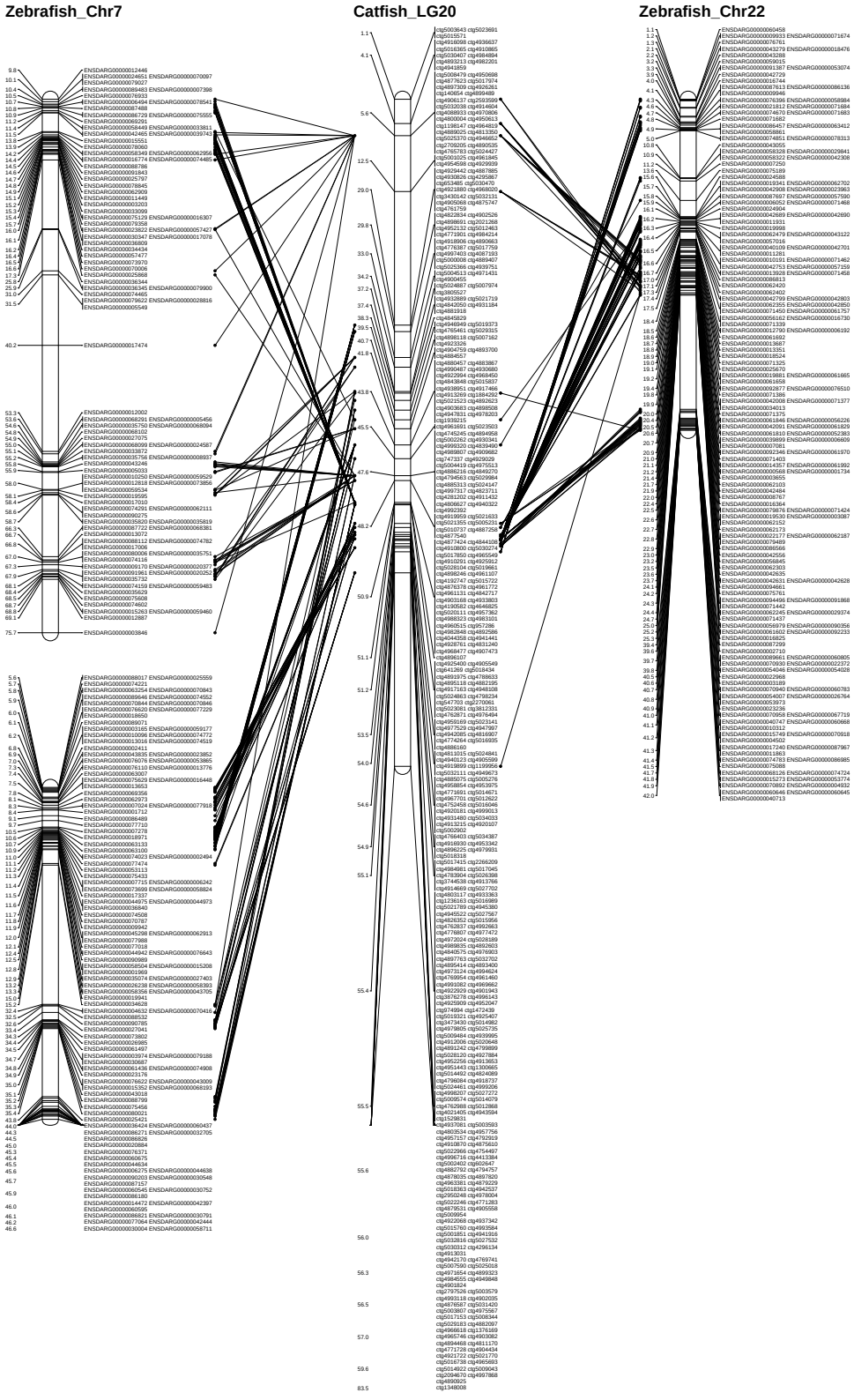


Figure 7 continued.

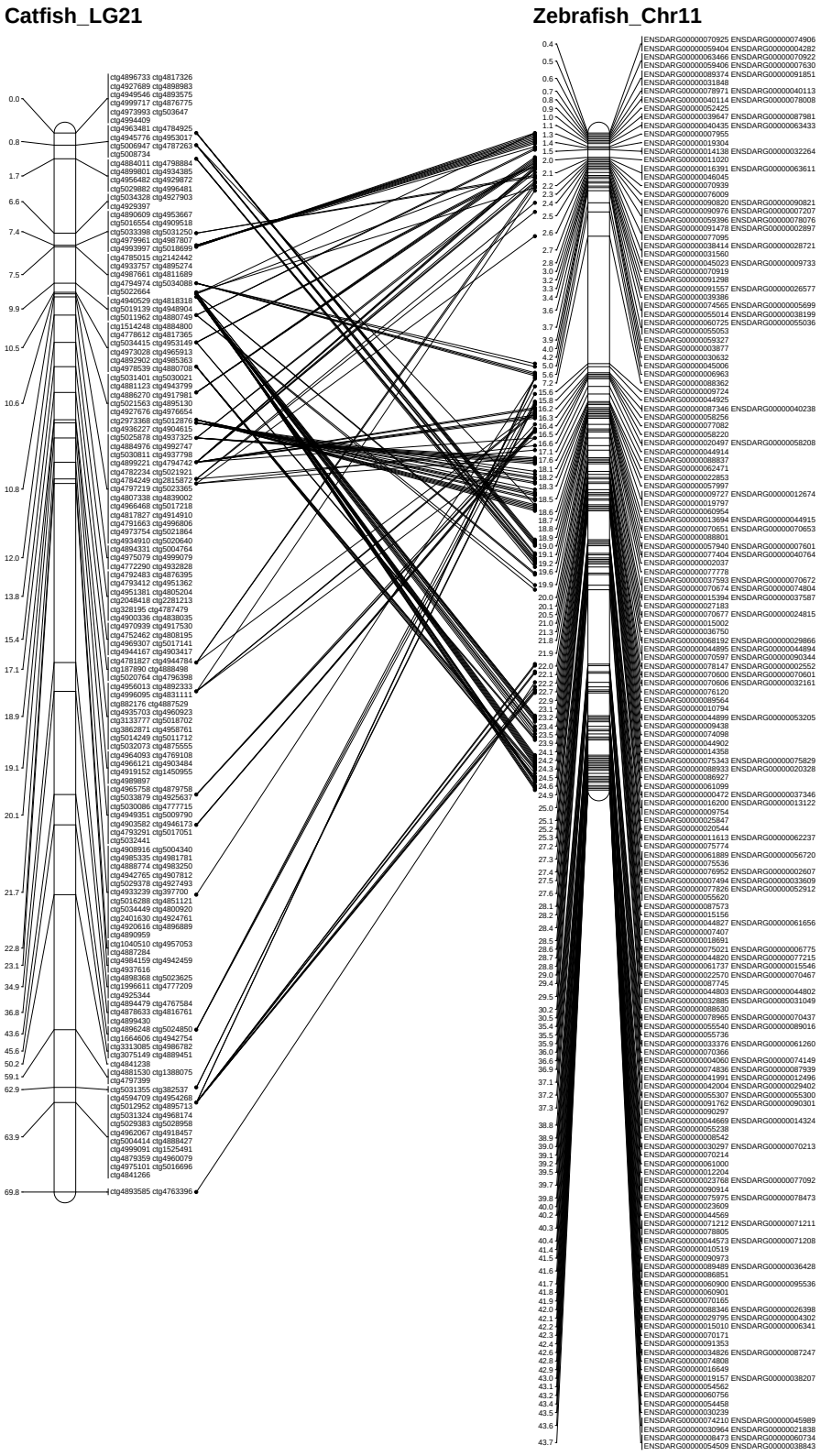


Figure 7 continued.

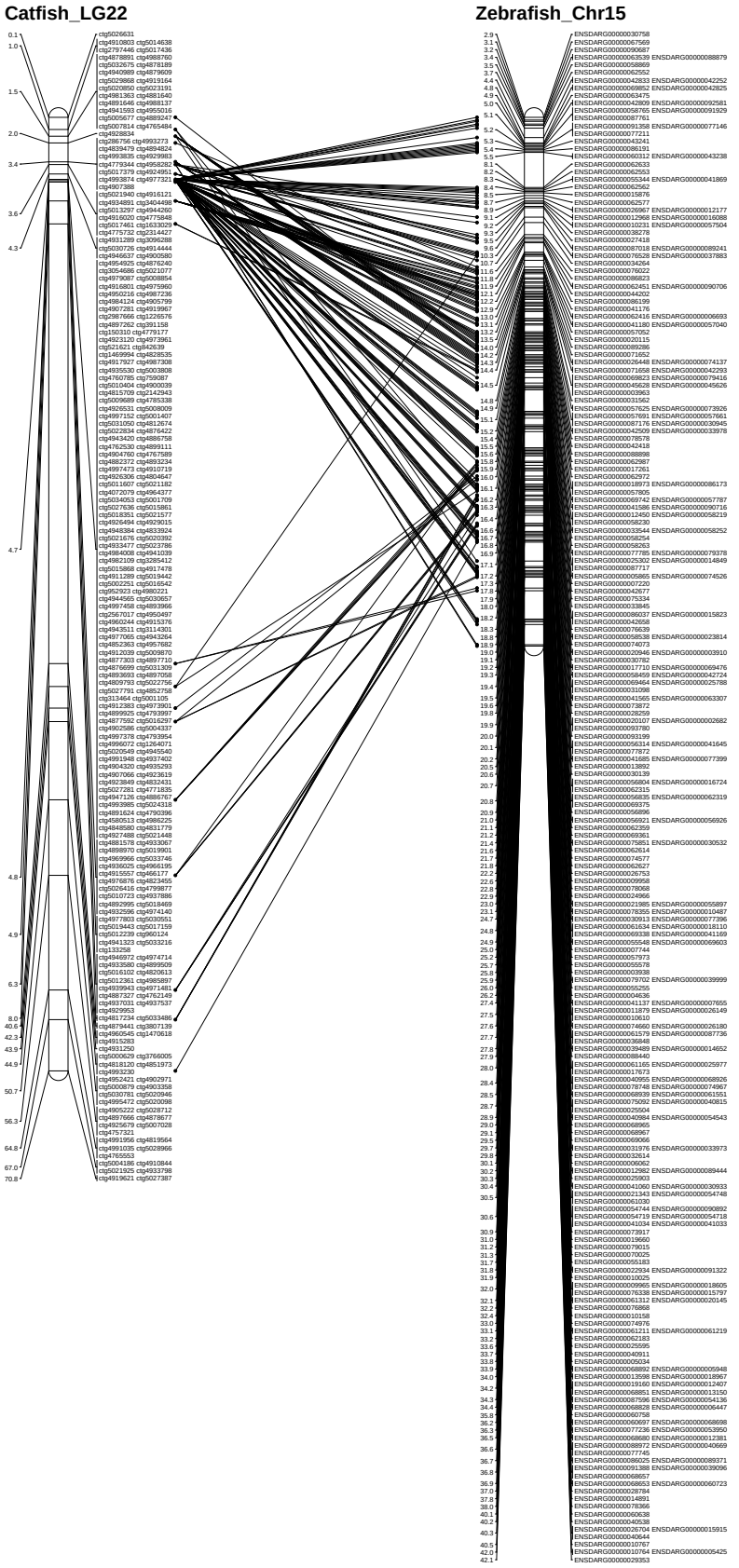


Figure 7 continued.

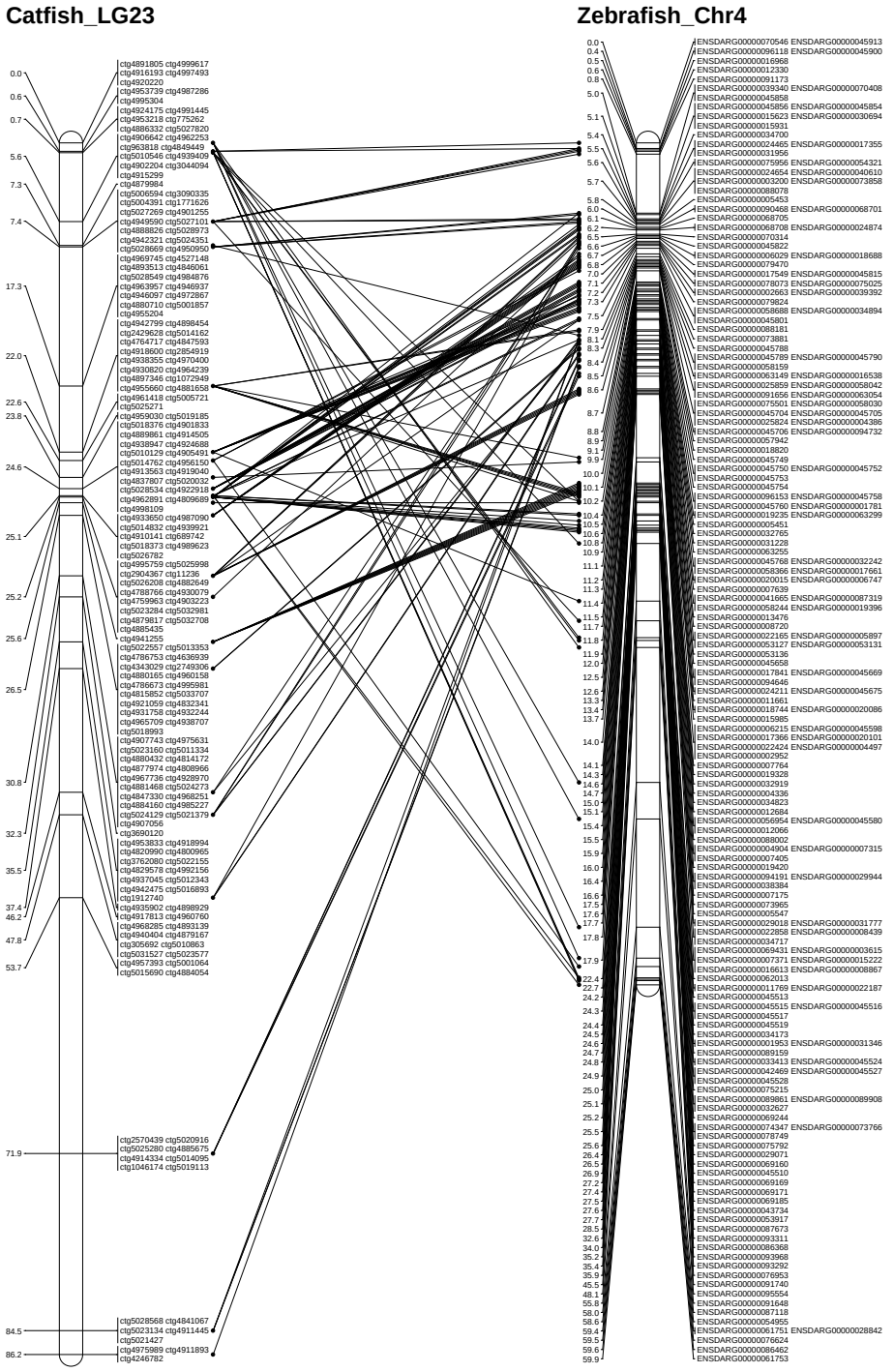


Figure 7 continued.

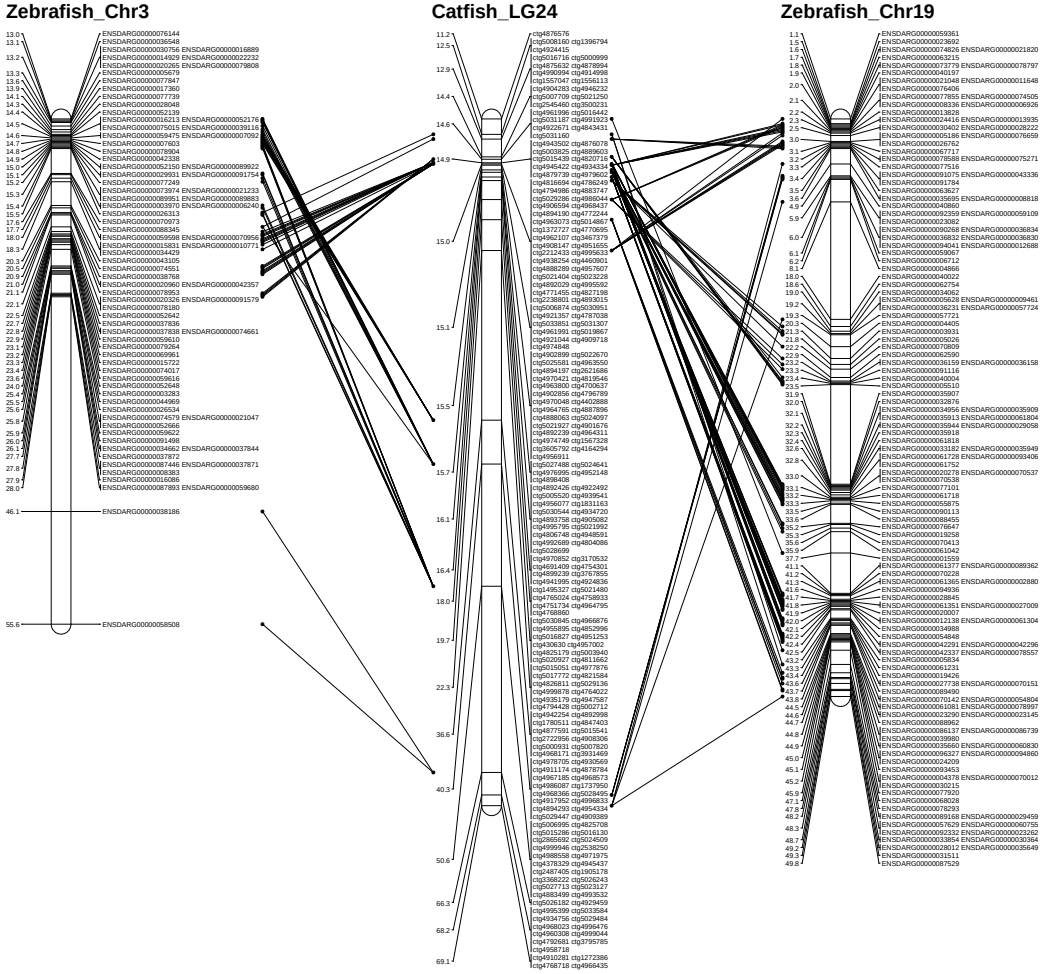


Figure 7 continued.

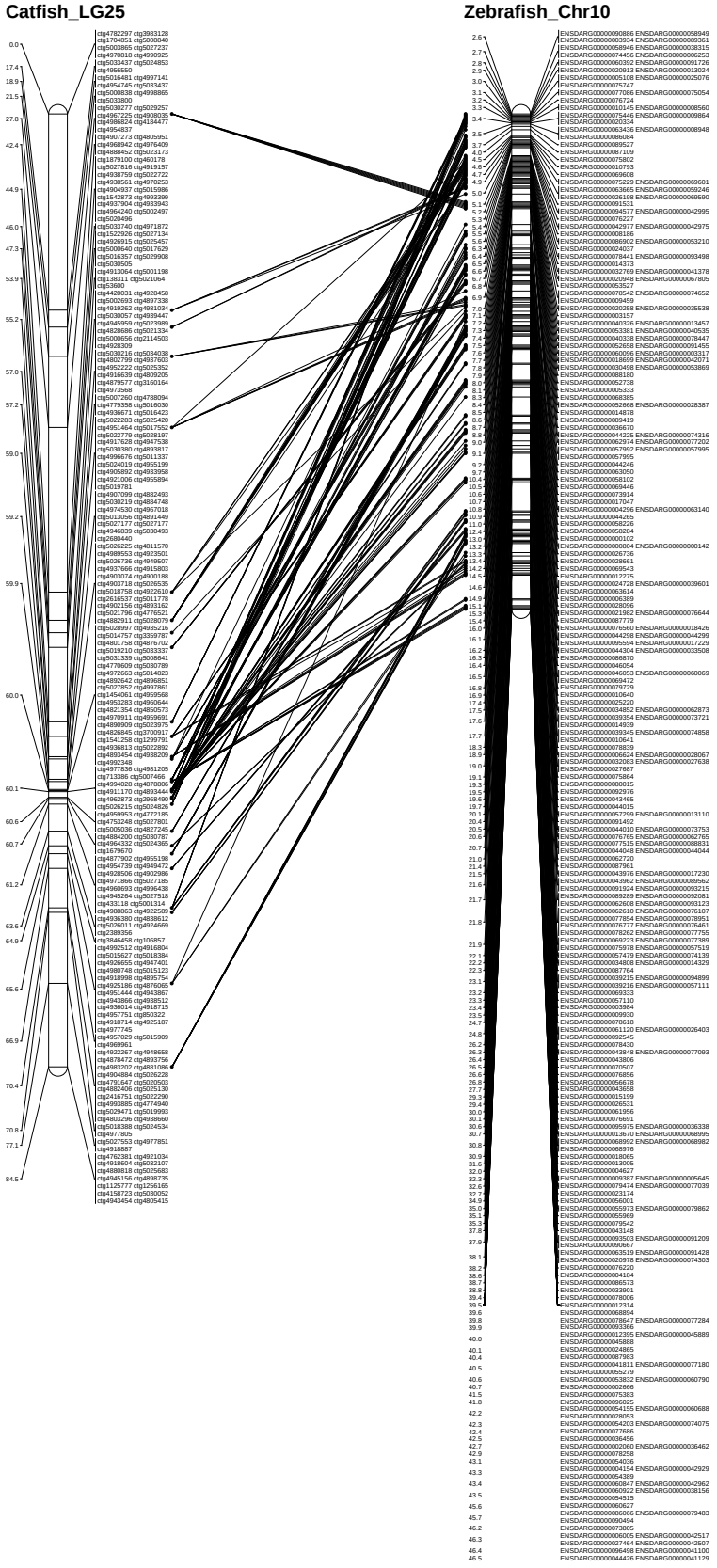


Figure 7 continued.

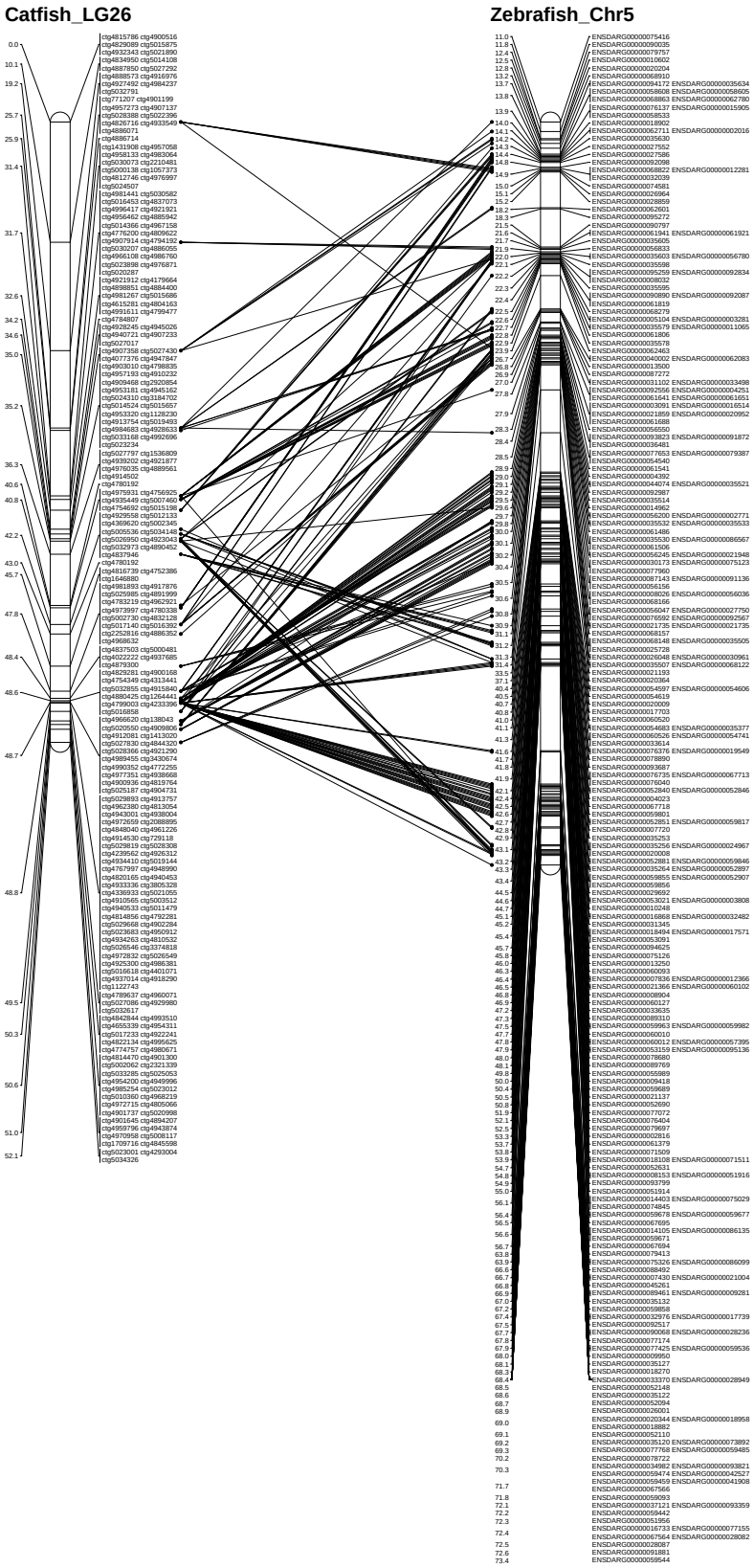


Figure 7 continued.

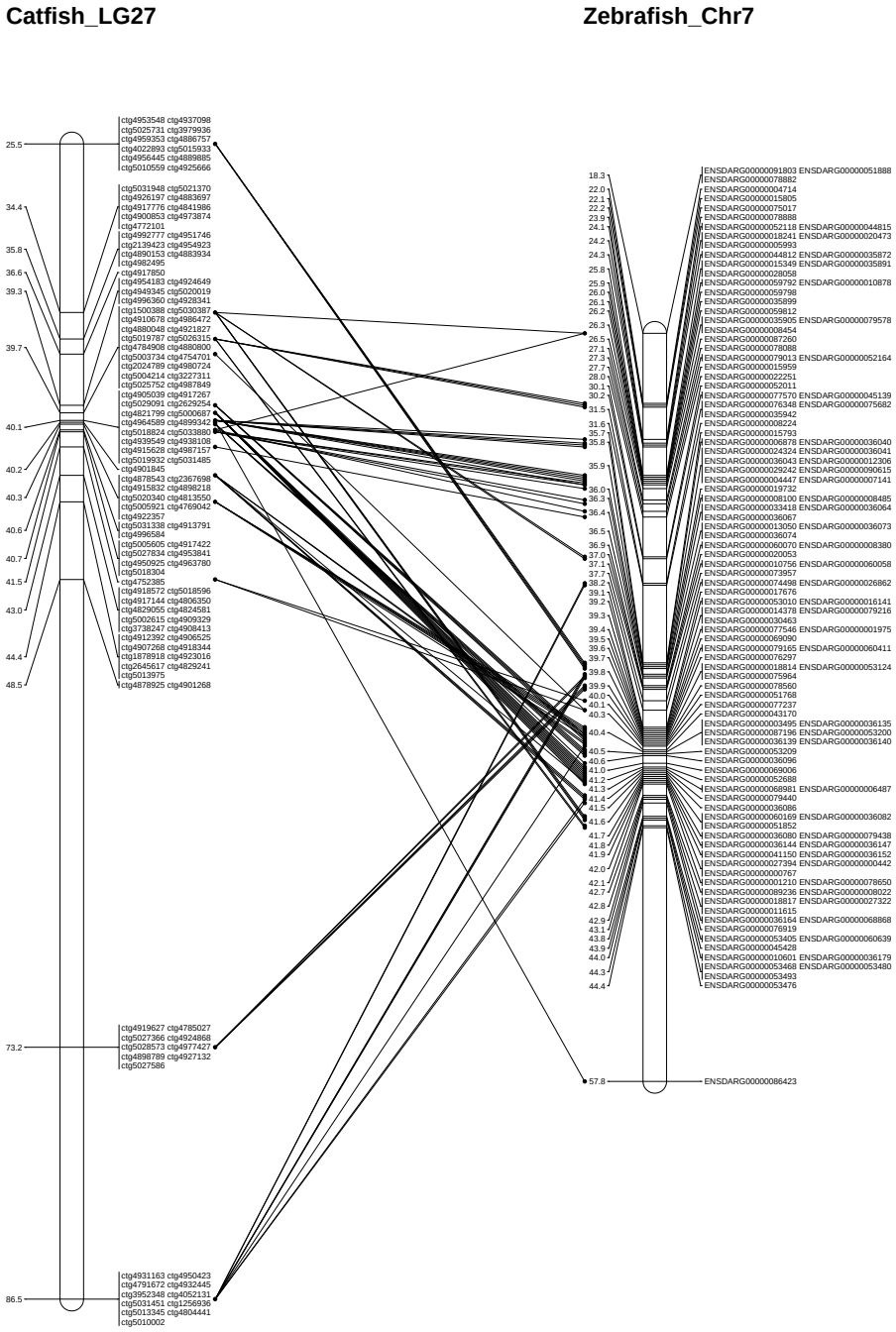


Figure 7 continued.

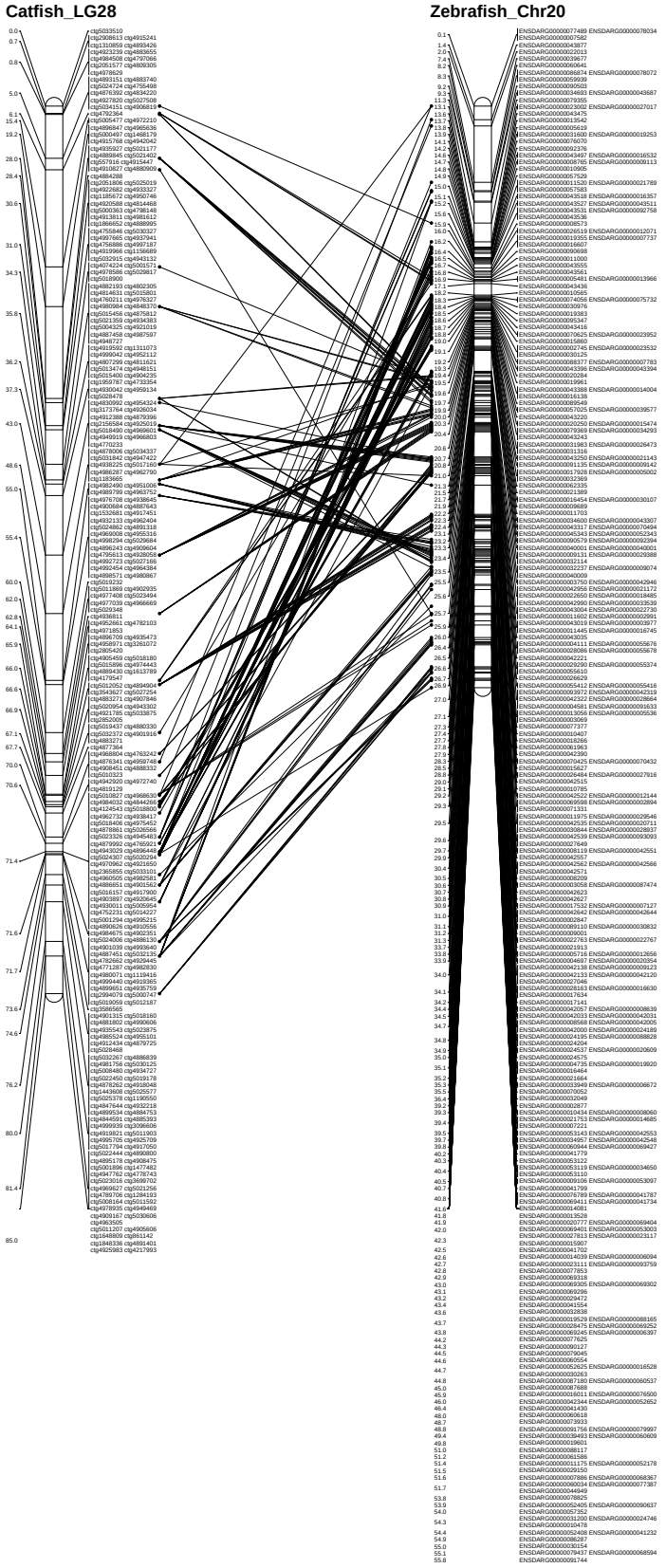
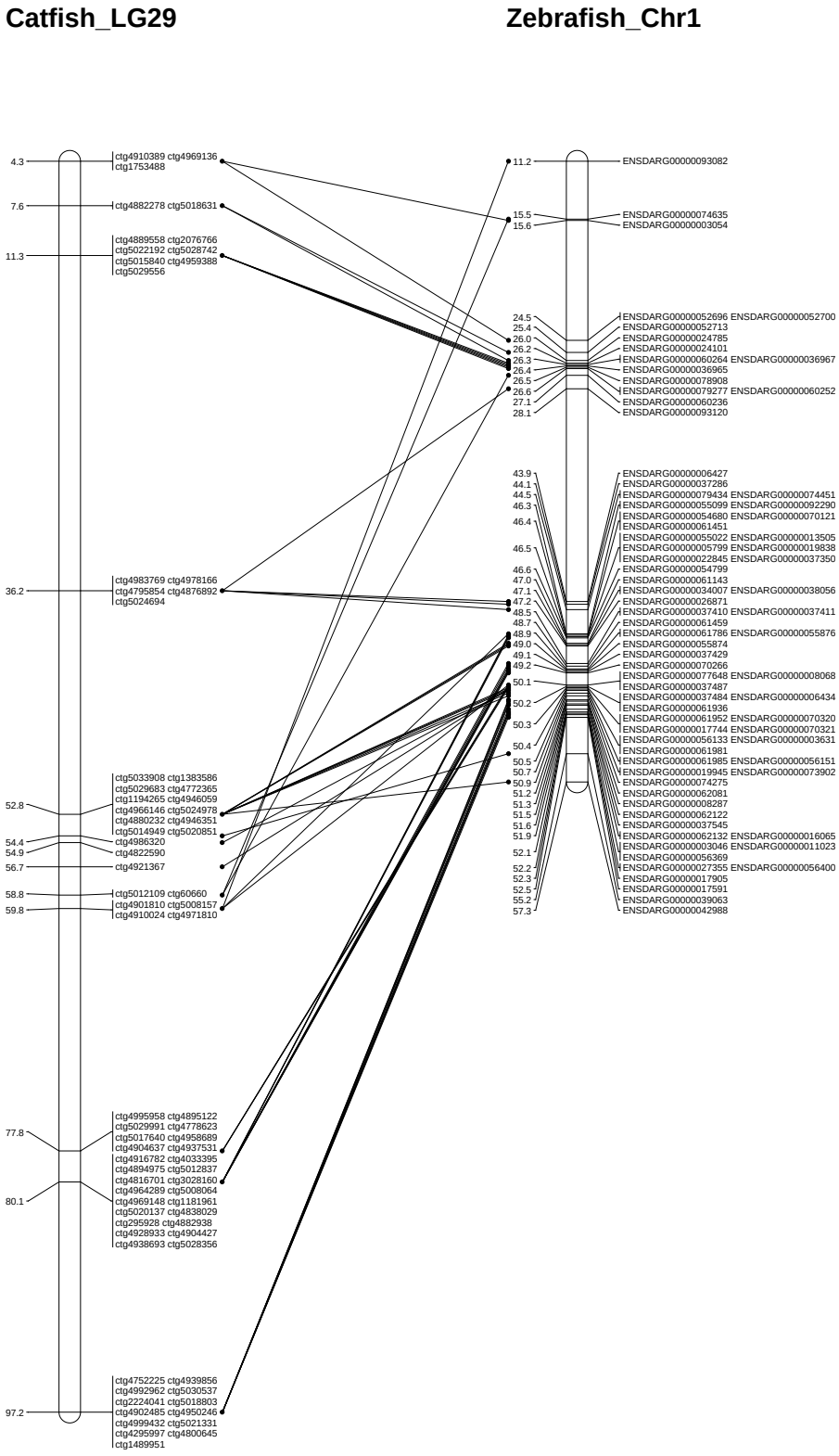


Figure 7 continued.



VI. REFERENCES

Andreassen R, Lunner S, Høyheim B. Characterization of full-length sequenced cDNA inserts (FLIcs) from Atlantic salmon (*Salmo salar*). BMC Genomics. 2009 Oct 30;10:502.

Aury JM, Cruaud C, Barbe V, Rogier O, Mangenot S, Samson G, Poulain J, Anthouard V, Scarpelli C, Artiguenave F et al. High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. BMC genomics 2008, 9:603.

Boffelli D, Nobrega MA, Rubin EM. Comparative genomics at the vertebrate extremes. Nat Rev Genet. 2004 Jun;5(6):456-65. Review.

Bruton, Michael N. Alternative life-history strategies of catfishes. Aquat. Living Resour.1996 (9): 35–41.

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

Castelli V, Aury JM, Jaillon O, Wincker P, Clepet C, Menard M, Cruaud C, Quétier F, Scarpelli C, Schächter V, et al. Whole genome sequence comparisons and "full-length" cDNA sequences: a combined approach to evaluate and improve Arabidopsis genome annotation. Genome Res. 2004 Mar;14(3):406-13.

Chen F, Lee Y, Jiang Y, Wang S, Peatman E, Abernathy J, Liu H, Liu S, Kucuktas H, Ke

C et al. Identification and characterization of full-length cDNAs in channel catfish (*Ictalurus punctatus*) and blue catfish (*Ictalurus furcatus*). PLoS One 2010, 5(7):e11546.

Chevreux B, Wetter T, Suhai S. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) 1999, 45-56.

Chistiakov DA, Hellemans B, Haley CS, Law AS, Tsigenopoulos CS, Kotoulas G, Bertotto D, Libertini A, Volckaert FA.A microsatellite linkage map of the European sea bass *Dicentrarchus labrax L*. Genetics 2005, 170:1821-1826.

Chistiakov DA, Tsigenopoulos CS, Lagnel J, Guo YM, Hellemans B, Haley CS, Volckaert FA, Kotoulas G. A combined AFLP and microsatellite linkage map and pilot comparative genomic analysis of European sea bass *Dicentrarchus labrax L*. Anim Genet. 2008 Dec;39(6):623-34.

Chou HH, Holmes MH.DNA sequence quality trimming and vector removal. Bioinformatics 2001, 17:1093-1104.

Clark MS, Edwards YJ, Peterson D, Clifton SW, Thompson AJ, Sasaki M, Suzuki Y, Kikuchi K, Watabe S, Kawakami K et al. Fugu ESTs: new resources for transcription analysis and genome annotation. Genome Res 2003, 13(12):2747-53.

Dalrymple BP, Kirkness EF, Nefedov M, McWilliam S, Ratnakumar A, Barris W, Zhao S, Shetty J, Maddox JF, O'Grady M et al. Using comparative genomics to reorder

the human genome sequence into a virtual sheep genome. Genome Biol 2007, 8:R152.

Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 1998a, 8:186-194.

Ewing B, Hillier L, Wendl MC, Green P. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 1998b, 8:175-185.

Ferraris CJ Jr, de Pinna MCC. Higher-level names for catfishes (*Actinopterygii: Ostariophysi: Siluriformes*). Proceedings of the California Academy of Sciences.1999 (51): 1–17.

Franch R, Louro B, Tsalavouta M, Chatziplis D, Tsigenopoulos CS, Sarropoulou E, Antonello J, Magoulas A, Mylonas CC, Babbucci M et al. A genetic linkage map of the hermaphrodite teleost fish *Sparus aurata L*. Genetics 2006, 174:851-861.

Fujiyama A, Watanabe H, Toyoda A, Taylor TD, Itoh T, Tsai S-F, Park H-S, Yaspo M-L, Lehrach H, Chen Z et al. Construction and Analysis of a Human-Chimpanzee Comparative Clone Map. Science 2002, 295:131-134.

Gregory SG, Sekhon M, Schein J, Zhao S, Osoegawa K, Scott CE, Evans RS, Burridge PW, Cox TV, Fox CA et al. A physical map of the mouse genome. Nature 2002, 418:743-750.

Guyomard R, Mauger S, Tabet-Canale K, Martineau S, Genet C, Krieg F, Quillet E. A type I and type II microsatellite linkage map of rainbow trout (*Oncorhynchus mykiss*) with presumptive coverage of all chromosome arms. BMC Genomics 2006, 7:302.

Hardison RC. Comparative genomics. PLoS Biol. 2003 Nov;1(2):E58. Epub 2003 Nov 17. Review.

Harhay GP, Sonstegard TS, Keele JW, Heaton MP, Clawson ML, Snelling WM, Wiedmann RT, Van Tassell CP, Smith TP. Characterization of 954 bovine full-CDS cDNA sequences. BMC Genomics. 2005 Nov 23;6:166.

Hayashizaki Y. The Riken mouse genome encyclopedia project. C R Biol. 2003 Oct-Nov;326(10-11):923-9. Review.

He C, Chen L, Simmons M, Li P, Kim S, Liu ZJ.Putative SNP discovery in interspecific hybrids of catfish by comparative EST analysis. Anim Genet 2003, 34:445-8.

Hurowitz EH, Drori I, Stodden VC, Donoho DL, Brown PO. Virtual Northern analysis of the human genome. PLoS One. 2007 May 23;2(5):e460.

Jiang Y, Lu J, Peatman E, Kucuktas H, Liu S, Wang S, Sun F, Liu Z. A pilot study for channel catfish whole genome sequencing and *de novo* assembly. BMC Genomics. 2011 Dec 22;12:629.

Ju Z, Dunham RA, Liu Z. Differential gene expression in the brain of channel catfish (*Ictalurus punctatus*) in response to cold acclimation. Mol Genet Genomics 2002, 268:87-95.

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

Karsi A, Cao D, Li P, Patterson A, Kocabas A, Feng J, Ju Z, Mickett KD, Liu Z. Transcriptome analysis of channel catfish (*Ictalurus punctatus*): initial analysis

of gene expression and microsatellite-containing cDNAs in the skin. Gene 2002, 285:157-168.

Katagiri T, Kidd C, Tomasino E, Davis JT, Wishon C, Stern JE, Carleton KL, Howe AE, Kocher TD.A BAC-based physical map of the Nile tilapia genome. BMC Genomics 2005, 6:89.

Kim S, Karsi A, Dunham R, Liu Z.The skeletal muscle alpha-actin gene of channel catfish (*Ictalurus punctatus*) and its association with piscine specific SINE elements. Gene 2000, 252:173-181.

Kocabas AM, Li P, Cao D, Karsi A, He C, Patterson A, Ju Z, Dunham RA, Liu Z. Expression profile of the channel catfish spleen: analysis of genes involved in immune functions. Mar Biotechnol (NY) 2002, 4:526-536.

Kucuktas H, Wang S, Li P, He C, Xu P, Sha Z, Liu H, Jiang Y, Baoprasertkul P, Somridhivej B et al. Construction of Genetic Linkage Maps and Comparative Genome Analysis of Catfish Using Gene-associated Markers. Genetics 2009, 181:1649-1660.

Kuhl H, Tine M, Beck A, Timmermann B, Kodira C, Reinhardt R. Directed sequencing and annotation of three Dicentrarchus labrax L. chromosomes by applying Sanger- and pyrosequencing technologies on pooled DNA of comparatively mapped BAC clones. Genomics 2011, 98(3):202-212.

Larkin DM, Everts-van der Wind A, Rebeiz M, Schweitzer PA, Bachman S, Green C, Wright CL, Campos EJ, Benson LD, Edwards J et al. A cattle-human comparative map built with cattle BAC-ends and human genome sequence.

Genome Res 2003, 13:1966-1972.

Lee BY, Lee WJ, Streelman JT, Carleton KL, Howe AE, Hulata G, Slettan A, Stern JE, Terai Y, Kocher TD. A second-generation genetic linkage map of tilapia (*Oreochromis spp*.). Genetics 2005, 170:237-244.

Leeb T, Vogl C, Zhu B, de Jong PJ, Binns MM, Chowdhary BP, Scharfe M, Jarek M, Nordsiek G, Schrader F et al. A human-horse comparative map based on equine BAC end sequences. Genomics 2006, 87:772-776.

Li C, Zhang Y, Wang R, Lu J, Nandi S, Mohanty S, Terhune J, Liu Z, Peatman E. RNA-seq analysis of mucosal immune responses reveals signatures of intestinal barrier disruption and pathogen entry following *Edwardsiella ictaluri* infection in channel catfish, *Ictalurus punctatus*. Fish Shellfish Immunol. 2012 May;32(5):816-27.

Li P, Peatman E, Wang S, Feng J, He C, Baoprasertkul P, Xu P, Kucuktas H, Nandi S, Somridhivej B et al. Towards the ictalurid catfish transcriptome: generation and analysis of 31,215 catfish ESTs. BMC Genomics 2007, 8:177.

Li RW, Waldbieser GC.Production and utilization of a high-density oligonucleotide microarray in channel catfish, *Ictalurus punctatus*. BMC Genomics 2006, 7:134.

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

Liu S, Zhou Z, Lu J, Sun F, Wang S, Liu H, Jiang Y, Kucuktas H, Kaltenboeck L, Peatman E et al. Generation of genome-scale gene-associated SNPs in catfish for

the construction of a high-density SNP array. BMC Genomics 2011, 12:53.

Liu Z, Karsi A, Dunham RA. Development of Polymorphic EST Markers Suitable for Genetic Linkage Mapping of Catfish. Mar Biotechnol (NY). 1999 Sep;1(5):437-0447.

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

Liu Z, Li P, Dunham R. Characterization of an A/T-rich family of sequences from the channel catfish (*Ictalurus punctatus*). Mol Mar Biol Biotechnol 1998, 7:232-9.

Liu Z, Li RW, Waldbieser GC.Utilization of microarray technology for functional genomics in ictalurid catfish. J Fish Biol 2008, 72:2377-2390.

Liu Z. A review of catfish genomics: progress and perspectives. Comp Funct Genomics 2003, 4:259-265.

Liu Z. Catfish. In Genome Mapping and Genomics in Fishes and Aquatic Animals. Edited by Kocher T, Kole C. Berlin Heidelberg: Springer; 2008: 85-100. [Kole C (Series Editor): Genome Mapping and Genomics in Animals, vol. 2]

Liu Z. Development of genomic resources in support of sequencing, assembly, and annotation of the catfish genome. Comp Biochem Physiol, Part D, Genomics and Proteomics 2011, 6:11-17.

Liu Z. (ed) (2007) Frontmatter, in Aquaculture Genome Technologies, Blackwell Publishing Ltd, Oxford, UK.doi: 10.1002/9780470277560.fmatter.

Luo MC, Thomas C, You FM, Hsiao J, Ouyang S, Buell CR, Malandro M, McGuire PE, Anderson OD, Dvorak J. High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. Genomics. 2003 Sep;82(3):378-89.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 2005, 437(7057):376-380.Apr;181(4):1649-60.

Marra MA, Kucaba TA, Dietrich NL, Green ED, Brownstein B, Wilson RK, McDonald KM, Hillier LW, McPherson JD, Waterston RH. High throughput fingerprint analysis of large-insert clones. Genome Res. 1997 Nov;7(11):1072-84.

Meyer A, Van de Peer Y. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). Bioessays 2005, 27(9):937-945.

Meyers SN, Rogatcheva MB, Larkin DM, Yerle M, Milan D, Hawken RJ, Schook LB, Beever JE. Piggy-BACing the human genome: II. A high-resolution, physically anchored, comparative map of the porcine autosomes. Genomics 2005, 86:739-752.

Miller JR, Koren S, Sutton G. Assembly algorithms for next-generation sequencing data. Genomics 2010, 95(6):315-327.

Moen T, Hayes B, Baranski M, Berg PR, Kjoglum S, Koop BF, Davidson WS, Omholt SW, Lien S. A linkage map of the Atlantic salmon (*Salmo salar*) based on

EST-derived SNP markers. BMC Genomics 2008, 9:223.

Moghadam HK, Ferguson MM, Danzmann RG.Whole genome duplication: challenges and considerations associated with sequence orthology assignment in Salmoninae. J Fish Biol 2011 79(3):561-574.

Moghadam HK, Ferguson MM, Danzmann RG. Comparative genomics and evolution of conserved noncoding elements (CNE) in rainbow trout. BMC Genomics 2009, 10:278.

Nagarajan H, Butler JE, Klimes A, Qiu Y, Zengler K, Ward J, Young ND, Methe BA, Palsson BO, Lovley DR et al. De Novo assembly of the complete genome of an enhanced electricity-producing variant of *Geobactersulfurreducens* using only short reads. PloS one 2010, 5(6):e10922.

Nandi S, Peatman E, Xu P, Wang S, Li P, Liu Z. Repeat structure of the catfish genome: a genomic and transcriptomic assessment of Tc1-like transposon elements in channel catfish (*Ictalurus punctatus*). Genetica 2007, 131:81-90.

Ng SHS, Artieri CG, Bosdet IE, Chiu R, Danzmann RG, Davidson WS, Ferguson MM, Fjell CD, Hoyheim B, Jones SJM et al. A physical map of the genome of Atlantic salmon, Salmo salar. Genomics 2005, 86:396-404.

Ninwichian P, Peatman E, Liu H, Kucuktas H, Somridhivej B, Liu S, Li P, Jiang Y, Sha Z, Kaltenboeck M et al. Second generation genetic linkage map and its integration with the BAC-based physical map in channel catfish. Genetics, in review. 2012.

Palti Y, Luo MC, Hu Y, Genet C, You FM, Vallejo RL, Thorgaard GH, Wheeler PA,

Rexroad CE 3rd. A first generation BAC-based physical map of the rainbow trout genome. BMC Genomics.2009 Oct 8;10:462.

Peatman E, Baoprasertkul P, Terhune J, Xu P, Nandi S, Kucuktas H, Li P, Wang S, Somridhivej B, Dunham R et al. Expression analysis of the acute phase response in channel catfish (*Ictalurus punctatus*) after infection with a Gram-negative bacterium. Dev Comp Immunol 2007, 31:1183-1196.

Peatman E, Terhune J, Baoprasertkul P, Xu P, Nandi S, Wang S, Somridhivej B, Kucuktas H, Li P, Dunham R et al. Microarray analysis of gene expression in the blue catfish liver reveals early activation of the MHC class I pathway after infection with *Edwardsiella ictaluri*. Mol Immunol 2008, 45:553-566.

Quiniou SM, Katagiri T, Miller NW, Wilson M, Wolters WR, Waldbieser GC.Construction and characterization of a BAC library from a gynogenetic channel catfish *Ictalurus punctatus*. Genet Sel Evol. 2003 Nov-Dec;35(6):673-83.

Quiniou SM, Waldbieser GC, Duke MV. A first generation BAC-based physical map of the channel catfish genome. BMC Genomics. 2007 Feb 6;8:40.

Quinn NL, Levenkova N, Chow W, Bouffard P, Boroevich KA, Knight JR, Jarvie TP, Lubieniecki KP, Desany BA, Koop BF et al. Assessing the feasibility of GS FLX Pyrosequencing for sequencing the Atlantic salmon genome. BMC Genomics 2008, 9:404.

Reinhardt JA, Baltrus DA, Nishimura MT, Jeck WR, Jones CD, Dangl JL. *De novo* assembly using low-coverage short read sequence data from the rice pathogen

Pseudomonas syringaepv.oryzae. Genome research 2009, 19(2):294-305.

Rexroad CE 3rd, Lee Y, Keele JW, Karamycheva S, Brown G, Koop B, Gahr SA, Palti Y, Quackenbush J. Sequence analysis of a rainbow trout cDNA library and creation of a gene index. Cytogenet Genome Res. 2003;102(1-4):347-54.

Rise ML, von Schalburg KR, Brown GD, Mawer MA, Devlin RH, Kuipers N, Busby M, Beetz-Sargent M, Alberto R, Gibbs AR et al. Development and application of a salmonid EST database and cDNA microarray: data mining and interspecific hybridization characteristics. Genome Res 2004, 14:478-490.

Sarropoulou E, Franch R, Louro B, Power DM, Bargelloni L, Magoulas A, Senger F, Tsalavouta M, Patarnello T, Galibert F et al. A gene-based radiation hybrid map of the gilthead sea bream Sparus aurata refines and exploits conserved synteny with *Tetraodon nigroviridis*. BMC Genomics 2007, 8:44.

Sarropoulou E, Nousdili D, Magoulas A, Kotoulas G. Linking the genomes of nonmodel teleosts through comparative genomics. Mar Biotechnol (NY). 2008 May-Jun;10(3):227-33.

Sarropoulou E, Power DM, Magoulas A, Geisler R, Kotoulas G. Comparative analysis and characterization of expressed sequence tags in gilthead sea bream (*Sparus aurata*) liver and embryos. Aquaculture 2005, 243:69-81.

Satoh K, Doi K, Nagata T, Kishimoto N, Suzuki K, Otomo Y, Kawai J, Nakamura M, Hirozane-Kishikawa T, Kanagawa S et al. Gene organization in rice revealed by full-length cDNA mapping and gene expression analysis through microarray.

PLoS One. 2007 Nov 28;2(11):e1235.

Sekino M, Kobayashi T, Hara M. Segregation and linkage analysis of 75 novel microsatellite DNA markers in pair crosses of Japanese abalone (*Haliotis discus hannai*) using the 5'-tailed primer method. Mar Biotechnol (NY) 2006, 8:453-466.

Senger F, Priat C, Hitte C, Sarropoulou E, Franch R, Geisler R, Bargelloni L, Power D, Galibert F. The first radiation hybrid map of a perch-like fish: The gilthead seabream (*Sparus aurata L*). Genomics 2006, 87:793-800.

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

Sha Z, Abernathy WJ, Wang S, Li P, Kucuktas H, Liu H, Peatman E, Liu Z. NOD-like subfamily of the nucleotide-binding domain and leucine-rich repeat containing family receptors and their expression in channel catfish. Dev Comp Immunol 2009, 31:991-999.

Shimizu N, Sasaki T, Asakawa S, Shimizu A, Ishikawa SK, Imai S, Murayama Y, Himmelbauer H, Mitani H, Furutani-Seiki M et al. Comparative genomics of medaka and fugu. Comp Biochem Physiol Part D Genomics Proteomics. 2006 Mar;1(1):6-12.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. ABySS: a parallel assembler for short read sequence data. Genome research 2009, 19(6):1117-1123.

Soares MB, Bonaldo MF, Jelene P, Su L, Lawton L, Efstratiadis A. Construction and

characterization of a normalized cDNA library. Proc Natl Acad Sci U S A. 1994 Sep 27;91(20):9228-32.

Soler L, Conte MA, Katagiri T, Howe AE, Lee BY, Amemiya C, Stuart A, Dossat C, Poulain J, Johnson J et al. Comparative physical maps derived from BAC end sequences of tilapia (*Oreochromis niloticus*). BMC Genomics. 2010 Nov 16;11:636.

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

Steinke D, Hoegg S, Brinkmann H, Meyer A. Three rounds (1R/2R/3R) of genome duplications and the evolution of the glycolytic pathway in vertebrates. BMC Biol 2006, 4:16.

Steinke D, Salzburger W, Meyer A. Novel relationships among ten fish model species revealed based on a phylogenomic analysis using ESTs. J Mol Evol 2006, 62:772-784.

Sun F, Peatman E, Li C, Liu S, Jiang Y, Zhou Z, Liu Z. Transcriptomic signatures of attachment, NF-κB suppression and IFN stimulation in the catfish gill following columnaris bacterial infection. Dev Comp Immunol. 2012 Sep;38(1):169-80.

Tan G, Karsi A, Li P, Kim S, Zheng X, Kucuktas H, Argue BJ, Dunham RA, Liu ZJ. Polymorphic microsatellite markers in *Ictalurus punctatus* and related

catfishspecies. Mol Ecol. 1999Oct;8(10):1758-60.

Tiersch TR, Simco KB, Davis KB, Chandler RW, Wachtel SS, and Carmichael GJ. Stability of genome size among stocks of the channel catfish. Aquaculture1990, 87:15-22.

Tucker CS. Channel catfish culture. 1985. Elsevier.

Voorrips RE. MapChart: software for the graphical presentation of linkage maps and QTLs. J Hered 2002, 93:77-8.

Waldbieser GC, Bosworth BG, Nonneman DJ, Wolters WR. A microsatellite-based genetic linkage map for channel catfish, *Ictalurus punctatus*. Genetics. 2001 Jun;158(2):727-34.

Wang S, Peatman E, Abernathy J, Waldbieser G, Lindquist E, Richardson P, Lucas S, Wang M, Li P, Thimmapuram J et al. Assembly of 500,000 inter-specific catfish expressed sequence tags and large scale gene-associated marker development for whole genome association studies. Genome Biol 2010, 11(1):R8.

Wang S, Xu P, Thorsen J, Zhu B, de Jong PJ, Waldbieser G, Kucuktas H, Liu Z. Characterization of a BAC library from channel catfish *Ictalurus punctatus*: indications of high levels of chromosomal reshuffling among teleost genomes. Mar Biotechnol (NY) 2007, 9:701-711.

Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P et al. Initial sequencing and comparative analysis of the mouse genome. Nature 2002, 420(6915):520-562.

Whitaker HA, McAndrew BJ, Taggart JB.Construction and characterization of a BAC library for the European sea bass *Dicentrarchus labrax*. Anim Genet 2006, 37:526.

Xia JH, Feng F, Lin G, Wang CM, Yue GH. A first generation BAC-based physical map of the Asian seabass (*Lates calcarifer*). PLoS One. 2010 Aug 5;5(8):e11974.

Xin D, Hu L, Kong X. Alternative promoters influence alternative splicing at the genomic level. PLoS One. 2008 Jun 18;3(6):e2377.

Xu P, Wang S, Liu L, Peatman E, Somridhivej B, Thimmapuram J, Gong G, Liu Z. Channel catfish BAC-end sequences for marker development and assessment of syntenic conservation with other fish species. Anim Genet. 2006 Aug;37(4):321-6.

Xu P, Wang S, Liu L, Thorsen J, Kucuktas H, Liu Z. A BAC-based physical map of the channel catfish genome. Genomics. 2007 Sep;90(3):380-8.

Xu P, Wang J, Wang J, Cui R, Li Y, Zhao Z, Ji P, Zhang Y, Li J, Sun X. Generation of the first BAC-based physical map of the common carp genome. BMC Genomics. 2011 Nov 2;12:537.

Zerbino DR, Birney E: Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome research 2008, 18(5):821-829.

Zhang X, Zhao C, Huang C, Duan H, Huan P, Liu C, Zhang X, Zhang Y, Li F, Zhang HB et al. A BAC-based physical map of Zhikong scallop (*Chlamys farreri Jones et Preston*). PLoS One. 2011;6(11):e27612. Epub 2011 Nov 16.

Zhang Y, Liu S, Lu J, Jiang Y, Gao X, Ninwichian P, Li C, Waldbieser G, Liu Z. Comparative genomic analysis of catfish linkage group 8 reveals two homologous chromosomes in zebrafish and other teleosts with extensive inter-chromosomal rearrangements. In Press.

Zimin AV, Delcher AL, Florea L, Kelley DR, Schatz MC, Puiu D, Hanrahan F, Pertea G, Van Tassell CP, Sonstegard TS et al. A whole-genome assembly of the domestic cow, *Bos taurus*. Genome Biol 2009;10(4).