COMPARATIVE GENOMIC ANALYSIS OF CATFISH LINKAGE GROUP 27 WITH TELEOST SPECIES GENOMES

by

Xiaoyu Gao

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Auburn, Alabama
August 4, 2012

Keywords: Comparative genomics, catfish, linkage group 27

Approved by

Zhanjiang Liu, Chair, Alumni Professor, Fisheries and Allied Aquacultures
Nannan Liu, Professor, Entomology and Plant Pathology
Claude E. Boyd, Professor, Fisheries and Allied Aquacultures

Abstract

Comparative genomics is a powerful tool to transfer genomic information from model species to related non-model species. Channel catfish (*Ictalurus punctatus*) is the primary aquaculture species in the United States. Its existing genome resources such as genomic sequences generated from next generation sequencing, BAC end sequences (BES), physical maps, linkage maps, and integrated linkage and physical maps using BES-associated markers provide a powerful platform for comparative genomic analysis between catfish and other model teleost fish species. This study aimed to gain understanding of similarities and conserved syntenies between the catfish genome and the genomes of several well-characterized fish genomes such as those of zebrafish, medaka, stickleback, and *Tetraodon* by examining one linkage group.

Acknowledgments

I would like to express my deep and sincere gratitude and appreciation to my advisor, Dr. Zhanjiang(John) Liu for his attention, guidance, insight, encouragement and support during this research and the preparation of this thesis.

I warmly thank my committee: Dr. Claude E. Boyd and Dr. Nannan Liu for their advice. I would also like to thank Dr. Huseyin Kucuktas, Mrs. Ludmilla Kaltenboeck and Mrs. Dongya Gao for their generous help and warm caring.

My thanks also go to all the colleagues in the lab: Shikai Liu, Jianguo Lu, Hong Liu, Yu Zhang, Jiaren Zhang, Yanliang Jiang, Hao Zhang, Chao Li, Ruijia Wang, Luyang Sun, Fanyue Sun, Parichart Ninwichian and Li Lian Wong for their help, collaboration, and friendship. I am grateful to my parents, my boyfriend and my friends for their constant support.

TABLE OF CONTENTS

List of tables

List of figures

Introduction

The genome mapping of the fish and aquatics were developing at the end of 20th century, which was much later than the mammals (Kocher et al. 1998, Moore et al. 1999, Young et al. 1998). But the development of the fish and aquatics has presented promising results for the research of molecular genetics of fish and aquatics. The comparative genomics is one of the most powerful tools to detect and discover the targeted genes. In this thesis, comparative genomics is used to find the relative genes between channel catfish (*Ictalurus punctatus*) and other model teleost fish species.

Background

According to the agricultural marketing resource center, catfish is the sixth most popular fish or seafood purchased and consumed in the United States (http://www.agmrc.org/commodities__products/aquaculture/catfish/), which is also one of the most important freshwater species cultured in the United States. Hence, the importance of the catfish genomics research attracts the big attention from the U.S. government, which listed the genomic research of catfish in its National Animal Genome Project NRSP-8 since 2003 (http://www.animalgenome.org/bioinfo/updates/). Therefore, the genomic research of catfish has been extensively studied by many scientific research methods including molecular genetics tools and comparative genomics (He et al. 2003, Serapion et al. 2004, Somridhivej et al. 2008, Xu et al. 2006, Kucuktas et al. 2009). Among these methods, comparative genomics is utilized as an effective way to compare the current model species genomic resources with catfish (Kucuktas et al. 2009).

Since the 1960s, the cell genetic research of more than 1,600 species of fish and aquatics showed their various differences of genomes (Kocher et al. 1998). For instance, the genome groups of fish are 20 to 25. The content of DNA is 684.6 to 2640.6 Mbp, which is 50% to 70% of human genome group. The genome project of aquatics was firstly begun at 1997 (Liu 2008). Ongoing genome projects include programs for Atlantic salmon, Atlantic cod, Pacific oysters, tilapia, catfish and shrimp (Liu 2010). After 15 years of development in the molecular genetics of fish, the gene sequence of four typical model species have been explored and partly researched. They are zebrafish *(Danio rerio)*, medaka *(Oryzias latipes)*, three-spined stickleback (*Gasterosteus aculeatus)*, and green spotted puffer (*Tetraodon)(REF)*. In this study, these model species were used as the genomic resource to find the related functional gene group on the catfish linkage group 27.

Contributions

The whole genome sequence of catfish is absence. However, the promising results of comparative genomic analysis between catfish and other model fish species would be acquired by using the existing genome resources such as genomic sequences generated from next generation sequencing, BAC end sequences (BES), physical maps, linkage maps, and integrated linkage and physical maps using BES-associated markers. Therefore, in this study, one linkage group of catfish, LG 27 was specifically analyzed to obtain the similarities and conserved syntenies between the catfish genome and those fully sequenced fish genomes (zebrafish, medaka, three-spined stickleback and green spotted puffer). The analyzed results showed that 18 most related genes were found between catfish and chromosome 6 of zebrafish, two of which were related to the immune system of catfish. Other significant similar results were found in the blast results against three-spined stickleback and green spotted puffer. All of these results

suggested the presence of conserved syntenies and intense rearrangements of chromosomal segments during the evolution.

Dissertation Organization

This dissertation is organized in the order that related research in literatures is briefly reviewed in Chapter 2 and the comparative genetics of catfish linkage group 27 and other four model species of teleost fish is conducted and the results are presented and discussed in Chapter 3. In Chapter 4, we summarize the main advancement for catfish genomic research of this dissertation and discuss the future directions.

Literature review

In the last decade, with the development of molecular biotechnology, the genomic resource has been extensively researched and provided. Massive genetic database of animals and plants has been served as public resource, which dramatically increased recently. However, the efficient tool lacked to deal with huge data resource. Therefore, bioinformatics (Zhao G P et al. 2002) was created for acquisition, processing, storage, distribution, analysis and interpretation of biological information as well as a method to make the comprehensive utilization of mathematics, computer science and biological tools. Because of the assistance of bioinformatics, human genome project is completed in 2003 and the speed of splicing and assembly of the genome sequence data have gradually improved.

Fish accounted for about 50% of vertebrate species during the evolutionary history (Nelson J S et al. 1994). Therefore, the research of fish resources, evolutionary conserved region between species and molecular mechanisms becomes important. But compared to higher vertebrates, the research of fish molecular biology lags behind. As for the fish, bioinformatics works continuously to improve the genome sequencing of gene splicing by the new fish functional genomics and molecular biology, this technology is using the principle of reverse biology, from gene to protein and then researchers could establish the methods to discover the new bioactive molecules (Liu H T et al. 1995). Scientists have been screening to a number of developmental, reproductive, and immune-related functional and domestic genes. Although the development of fish new functional genes start late, the bioinformatics combined with molecular biology methods greatly explored the genetic database information of fish. The development and the

study of important fish functional genes make massive achievements in this field. In this section, current research status of fish genome database and the methods of using these databases are summarized.

A public database of the fish gene resources

Zebrafish (*Danio rerio)*

Zebrafish belongs to the minnow family *(Cyprinidae*) of order *Cypriniformes*. Because of its unique characteristics, such as: small size, short spawning cycle, high fecundity, large egg, in vitro fertilization, transparent embryos, early embryonic development, fast and easy access to a large number of samples, zebrafish becomes one of the world-recognized mode of vertebrates.

Zebrafish genome is composed by the 25-chromosome (1.7×10^3Mb, 1Mb = 10^6 base pair or bp), about half of the mammalian genome numbers (Postlethwait J H et al. 2004). The British Sanger Institute started the whole genome of the zebrafish sequencing work in February 2001, the project has used the BAC (Bacterial artificial chromosome), the PAC (P1, derived artificial chromosome) cloning library and then sequenced the whole genome by the shotgun sequencing (Whole genome the shotgun sequencing).

Meanwhile, University of Washington also carried out a different scale of EST sequencing, this resource of the zebrafish could be compared to the Internet database (Table 2.1). It integrates a large number of zebrafish research resources, including genomic data, EST data, research reference text, researchers and units. In addition, it provides the zebrafish mutants, EST of cDNA, as well as monoclonal antibodies and other resources.

Medaka (*Oryzias latipes)*

Medaka is one of the small freshwater needle fish (*Beloniformes*). It has many advantages such as the small body, the healthy short period, strong breeding ability, transparent eggs, and adaptability to the variance in dissolved oxygen and temperature of the environment (Mitani H et al. 2004). Medaka has 24 pairs of chromosomes, containing (0.65 - 1) $\times 10^3$Mb bases; its size is equivalent to one-third of the human genome which is less than the zebrafish genome size (Wittbrodt J et al. 2002). Japanese scientists have succeeded in developing many natural or human-induced (Ethylnitrosourea ENU) mutants. including a transparent medaka mutant (Wakamatsu Y et al. 2001). The Japanese research organizations started the Medaka Genome Project in 2002, including building the large-scale EST measured sequence, the BAC library construction sequencing, whole genome shotgun sequencing and genetic map production and so on. At present, nearly 10 different medaka cDNA library composed by millions of EST sequences become a public database at the University of Tokyo, Japan. The Mbase (Table 2.1) provides researchers with efficient analysis tools and database. Approximately 90% of the genome sequence has been determined and the stitching work is constantly being improved; Japanese genetic Institute (National Institute of Genetics,) provided the genome Sequence database (Table 2.1). The unique characteristics of Medaka and the current research results gradually make it become a new model vertebrate.

Green spotted puffer (*Tetraodon)*

In 2004, Jaillon, et al. reported the genome sequence sketch of the black *tetraodon* (Jaillon, et al. 2004). *Tetraodon* is a kind of freshwater fish with small body, which could be fed in small tank. The genome size of *tetraodon* is about 350 Mb which is known as the smallest vertebrate genome. There are 39 long splice sequences (Ultracontigs) has been located to the types of

chromosome 21, covering 64.2% of the genes in the group (http://www.genoscope.cns.fr/externe/ tetraodon/). In addition, Jaillon, et al. also provided a large number of EST sequences information (Jaillon, et al. 2004), you can find them through the GENOSCOPE ENSEMBL or the NCBI database work tools and extraction (Table 2.1).

Other species

In addition, the genome centers of the United States, Britain, Japan and other countries started the sequence of variety of fish genome and EST sequencing. At present, Three-spine stickleback *(Gasterosteus aculeatus)*, channel catfish (*Ictalurus punctatus)*, Rainbown trout (*Oncorhynchus mykiss)*, Atlantic salmon *(Salmo salar)*, Nile bichir (*Polypterus bichir)*, *Astatotilapia burtoni*, red tail sheller *(Haplochromis sp.)*, killifish *(Fundulus heteroclitus)*, and Sea lamprey (*Petromyzon marinus)* including many of the fish gene sequence information, could be found in the TGI, ENSEMBL or NCBI database server for query and extraction (Table 2.1). With the improvement of these types of genetic sequence information, these databases will provide an extremely important platform for Comparative genomics and comparative biology.

Comparative genomic analysis of channel catfish linkage group 27

Introduction

Comparative genomics allows the importance of genetic drift to be studied on a large scale across the genome (Hardison RC et al. 2003). After the genome of the bacterium *Haemophilus influenza* became the first completely sequenced genome, a massive of completely sequenced genomes were discovered by the scientists (Fleischmann RD et al. 1995). Hence, comparative genomics has been widely utilized in the study of the similarities and difference in structure and function of hereditary information on many kinds of species (Hardison RC et al. 2003, Fleischmann RD et al. 1995, Paterson AH et al. 2000, Castresana J. 2001, Souciet J-L et al. 2009, Sarropoulou E et al. 2011). The power of comparative genomics on the human genomes has been proved by many scientists; this makes it possible to enhance the functional studies in the non-model species of interest (Gregory et al. 2002, Larkin et al. 2003, Meyers et al. 2005, Rogatcheva et al, 2008, Leeb et al. 2006).

Presently, the comparative genomics was effectively applied in fish molecular genetics. The whole genomes fully sequenced of five bony fish have been presented by scientists, which are *Takifugu rubripes* (Puffer fish) (http://www.fugu-sg.org/), *Tetraodon nigroviridis* (Puffer fish) (Jaillon O et al. 2004), *Gasterosteus aculeatus* (Three-spined stickleback) (http://useast.ensembl.org/Gasterosteus_aculeatus/Info/Index), *Danio rerio* (Zebrafish) (http://useast.ensembl.org/Danio_rerio/Info/Index), *Gadus morhua* (Atlantic cod)

(http://codgenome.no/). The sequenced genomes provide vast availability of information for the utilization of comparative genomics. Most recently, the genomic resources for a massive number of fish species has been developed by many genomic research scientists, such as Moen T. etc (Moen T et al. 2008), lee BY etc (Lee BY et al. 2006) and Liu Z. etc (Liu Z et al. 2011). These researches have showed promising future for the application of comparative genomics in fish genomics research.

According to agricultural marketing resource center, catfish is the sixth most popular fish or seafood consumed in the United States (http://www.agmrc.org/commodities__products/aquaculture/catfish/), which is also one of the most important freshwater species cultured in the United States. Hence, the importance of the catfish genomics research attracts the big attention from the U.S. government, which listed the genomic research of catfish in its National Animal Genome Project NRSP-8 since 2003(http://www.animalgenome.org/bioinfo/updates/). Therefore, the genomic research of catfish has been extensively studied by many scientific research methods including molecular genetics and comparative genomics etc. (He et al. 2003, Serapion et al. 2004, Somridhivej et al. 2008, Xu et al. 2006, Kucuktas et al. 2009). Among these methods, comparative genomics is utilized as an effective way to compare the current model species genomic resources with catfish (Kucuktas et al. 2009).

For catfish and other teleosts, comparative genomics can be used to detect the sequences of their uniqueness (Hardison, R.C. et al. 2003). Currently, the available genome resources for non-model fish species are limited, but they have been used in the first comparative genomics approaches (Danzmann, R.G et al. 2008). The genomes of channel catfish and zebrafish have 29 and 25 chromosomes, respectively; a rearrangement of chromosomes has to be made when

comparing the two genomes. As well as genomic information, this was achieved by the construction of a genetic linkage map of the channel catfish using EST-based microsatellites and single nucleotide polymorphism (SNP) markers (Kucuktas et al. 2009). Functional information such as expression profiles is also available for catfish, forming the way for comparative transcriptomics between fish (Bilodeau-Bourgeois L et al. 2008, Peatman, E et al. 2008). For catfish, between channel catfish and channel x blue catfish backcross populations, two BAC contig-based physical maps that are to be integrated to reduce the number of contigs, a large database of BAC-end sequences, integrated genetic and physical maps, and a large database of EST sequences and structure of interspersed repeats in the catfish genome have been characterized (Liu Z et al. 2011).

The whole genome sequence of catfish is absence. However, the promising results of comparative genomic analysis between catfish and other model fish species would be acquired by using the existing genome resources such as genomic sequences generated from next generation sequencing, BAC end sequences (BES), physical maps, linkage maps, and integrated linkage and physical maps using BES-associated markers. Therefore, in this study, one linkage group of catfish, LG 27 was specifically analyzed to obtain the similarities and conserved syntenies between the catfish genome and those fully sequenced fish genomes (zebrafish, medaka, stickleback and green spotted puffer). The analyzed results showed that 18 most related genes were found between catfish and chromosome 6 of zebrafish, two of which were related to the immune system of catfish. Other significant similar results were found in the blast results against medaka, stickleback and green spotted puffer. All of these results suggested the presence of conserved syntenies and intense rearrangements of chromosomal segments during the evolution.

Methods

Identification of genes of the catfish linkage group 27

All the BAC end sequences (BES) on the linkage group 27 (LG 27) had been collected before the identification of genes of LG27 on catfish. 60 physical map contigs contains 60 BAC end-associated microsatellites in LG27. The combination of these BES and physical map presented a total of 1,899 BES in LG27. After collecting these BES, BLASTN (E-value -10) searches were conducted to obtain longer sequences of LG27 by using these 1,899 BES as queries against the exiting catfish genomic contigs (average length is 3kb) (Table 3.1).

BLASTN searches using these 1,899 BES against the existing catfish genomic sequence contigs resulted in 786 significant hits. A total of 196 unique genes were identified using significant hit of genomic contigs as queries for BLASTX (E-value $\leq$-5) searches against the ENSEMBL zebrafish protein database (Zv9) (Table 3.1). Using the top hits from zebrafish, a total of 364 coding sequences were obtained from the zebrafish genome by Biomart searches, which were used as queries for BLASTX searches against ENSEMBL protein database of medaka (HdrR), three-spined stickleback (BROADS1), and Green spotted puffer (TETRAODON8.0) to locate their chromosomal location in these model fish.

Comparative genome analysis

Syntenies were established based on microsatellites-based linkage mapping and zebrafish chromosomal locations. Putative conserved syntenies were identified when the genes were located in the same chromosome and the same linkage group. The whole process is briefly shown in Figure 3.1.

Results and Discussion

Establishing chromosome-scale scaffolds

In order to conduct comparative genomic analysis, the first required step without a whole genome sequence is to establish large scaffolds that can then be compared to chromosome segments of other species with rich genomic resources. Here, we started with the 68 BAC end sequence-derived microsatellites that has been mapped to linkage group 27 (LG27) (Ninwichian. Et al., 2012). As shown in Table 3.2, these 68 mapped BAC end sequence-derived microsatellites were from 60 BAC contigs of the physical map that include 1,899 BAC end sequences (BES). BLASTN searches using these 1,899 BES against the existing catfish genomic sequence contigs resulted in 786 significant hits (Table 3.1). The 786 genomic sequence contigs were then used as queries to determine what genes are associated with these genomic sequence contigs using BLASTX searches against ENSEMBL zebrafish protein dabatase. The BLASTX searches resulted in 196 unique gene hits.

Because the position of the 68 mapped BES-derived microsatellites are known on LG27, the BLASTX analysis allowed the anchor of the 196 genes associated with the BAC clones on LG27 to be positioned on LG27, forming the LG27 scaffold for comparative analysis.

Identification of orthologous chromosomes of catfish LG27

The 196 genes identified on LG27 were used as queries to search the genomes of the four sequenced teleost species, zebrafish, medaka, stickleback, and green spotted puffer. As summarized in Table 3.2, the largest number of genes had hits on chromosome 6 (118 hits) in zebrafish, although significant hits existed for all chromosomes, as well as for unassigned scaffolds (Table 3.2). Similarly, the 196 genes had largest number of hits on three chromosomes

in medaka (chromosome 4, 5 and 17), two chromosomes in stickleback (chromosome 8 and 17), and three chromosomes in green spotted puffer (chromosome 1, 11 and 15). As catfish is most closely related to zebrafish phylogenetically, the number of the genes had significant hits were also largest in zebrafish. In spotted green pufferfish, a large number of these genes have not been assigned to chromosomes, and that is part of the reason that the numbers of gene hits on the relevant chromosomes were low (Table 3.2).

Annotation of genes on LG27

Annotation in teleost species is often difficult because of the complication of genes from a non-model species require detailed phylogenetic analysis, or analysis if evolutionarily conserved syntenic blocks. Here we have annotated 196 genes on catfish LG27 through comparative analysis of conserved microsyntenies on zebrafish chromosome 6 (Table 3.3).

As shown in Table 3.4, various length of conserved syntenies were identified. For instance, the physical contig 0716, 4 genes were identified on zebrafish chromosome 6, ie. , High-density lipoprotein-binding protein (vigilin), FERM, RhoGEF and pleckstrin domain protein 2, ATG4 autophagy related 4 homolog B (S. cerevisiae), and Deoxythymidylate kinase (thymidylate kinase). These four genes are clustered together on zebrafish chromosome 6 with a space (without consideration of the gene sizes) of 76Kb, 140Kb, and 14Kb between them, strongly support the orthologues relationship of these genes in catfish and zebrafish. A total of 13 microsyntenic blocks were identified between catfish LG27 and zebrafish chromosome 6. The largest conserved syntenic block involves a total of 2 genes spanning a distance of 906Kb.

Conclusions and Future Work

In this dissertation, we finished the comparative genetics analysis between catfish genome group 27 and other four teleost fish. In addition, the conserved regions between them were developed and shown. These works were described in chapter 3. The goal of this chapter is to summarize the main advancement of the dissertation and concentrate on some future directions, which could be the extensions of our past and current research on the comparative genetics research of catfish.

Main Contributions

These results suggested the presence of many conserved syntenies but also intense rearrangements of chromosomal segments during evolution. It is interesting to observe a higher level of genome scale structural conservation between catfish and zebrafish than between catfish and the other three fish species. In addition, these findings are consistent with our previous study that high levels of conservation were found within small genomic scale, whereas high levels of genome reshuffling and rearrangements were found on large genomic scale between catfish and zebrafish.

Future Work

Due to the huge amount of research works are being contributed to this field, which make the future work of this field show the different developmental directions. To the research in this dissertation, what we can develop in related areas for future work is also complicated, but several of them are very obvious and clear. They are shown as follows:

1. Providing a way for development of the fish genes (such as leukocyte interleukin-2, -4, -15, etc.).
2. Updating fish genome database continuously.
3. Making more effective use of biological Bioinformatics tools, comparative genomics and other research methods.
4. Developing and controlling the immune regulation related genes.

Reference

Bei J X, Suetake H, Araki K, et al . Two interleukin (IL) -15 homologues in fish from two distinct origins. Mol Immunol, 2006, 43 :860 —869

Bilodeau-Bourgeois L, Bosworth BG, Peterson BC. Differences in mortality, growth, lysozyme, and toll-like receptor gene expression among genetic groups of catfish exposed to virulent Edwardsiella ictaluri. Fish Shellfish Immunol. 2008, 24(1):82–9.

Bird S , Zou J , Kono T, et al . Characterization and expression analysis of interleukin 2 (IL22) and IL221 homologues in the Japanese pufferfish , Fugu rubripes , following their discovery by synteny. Immunogenetics , 2004 , 56 : 909 —923

Bourque G, Pevzner P A, Tesler G. Reconstructing the Genomic Architecture of Ancestral Mammals : Lessons From Human , Mouse , and Rat Genomes. Genome Res , 2004 , 14 : 507 —516

Castresana J. Comparative genomics and bioenergetics. Biochim. Biophys. Acta. 2001, 1506(3):147–62.

Danzmann RG, Davidson EA, Ferguson MM, Gharbi K, Hoyheim B, Koop BF, et al. Distribution of ancestral proto-Actinopterygian chromosome arms within the genomes of 4R-derivative salmonid fishes (Rainbow trout and Atlantic salmon). BMC Genomics. 2008, 9(1):557.

Ellegren H. Comparative genomics and the study of evolution by natural selection. Mol. Ecol. 2008, 17(21):4586–96.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae Rd. Science. 1995, 269(5223):496–512.

Gregory SG, Sekhon M, Schein J, Zhao S, Osoegawa K, Scott CE, et al. A physical map of the mouse genome. Nature. 2002, 418(6899):743–50.

Harrison G, Wedlock D. Marsupial cytokines: structure , function and evolution. Dev Comp Immunol, 2000 , 24 : 473 —484

Hardison RC. Comparative Genomics. PLoS Biol. 2003, 1(2):e58.

He C, Chen L, Simmons M, Li P, Kim S, Liu ZJ. Putative SNP discovery in interspecific hybrids of catfish by comparative EST analysis. Anim. Genet. 2003, 34(6):445–8.

Huseyin Kucuktas SW. Construction of Genetic Linkage Maps and Comparative Genome Analysis of Catfish Using Gene-associated Markers. Genetics. 2009;

Jaillon O, Aury J-M, Brunet F, Petit J-L, Stange-Thomann N, Mauceli E, et al. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. Nature. 2004, 431(7011):946–57.

Kasahara M, Naruse K, Sasaki S, Nakatani Y, et al. The medaka draft genome and insights into vertebrate genome evolution. Nature 2007, 447: 714-9.

Ravi V, Venkatesh B: Rapidly evolving fish genomes and teleost diversity. Curr Opin Genet Dev 2008, 18: 544-50

Kurokawa T, Uji S, Suzuki T. Identification of cDNA coding for a homologue to mammalian leptin from pufferfish , Takifugu rubripes. Peptides, 2005 , 26 : 745 —750

Lee B-Y, Lee W-J, Streelman JT, Carleton KL, Howe AE, Hulata G, et al. A Second-Generation Genetic Linkage Map of Tilapia (Oreochromis Spp.). Genetics. 2005, 170(1):237–44.

Liu H T. Cloning and the strategy of new cytokines. International Journal of Immunology, 1995 , 5 :256 —261

Liu SM, YangL. Pufferfish2An ideal model for genetic studies of vertebrate. Int J Genet, 1998 , 21 : 1 —4

Liu Z. Aquaculture genomics---progress in identifying species genetics continues. Global Aquaculture Advocate, November/December 2008, 75-78.

Liu Z, Peatman E. Whole Genome Sequencing of Aquaculture Species: Identifying the Genotypes behind Production Phenotypes. Global Aquaculture Advocate, March/April (2010):68-70.

Liu Z. Development of genomic resources in support of sequencing, assembly, and annotation of the catfish genome. Comp. Biochem. Physiol. Part D Genomics Proteomics. 2011, 6(1):11–7.

Meyers SN, Rogatcheva MB, Larkin DM, Yerle M, Milan D, Hawken RJ, et al. Piggy-BACing the human genome II. A high-resolution, physically anchored, comparative map of the porcine autosomes. Genomics. 2005, 86(6):739–52.

Mitani H, Shima A , Naruse K, et al . Medaka genome mapping to functional genomics. In: Gong Z, Korzh V (Eds.) , Fish Development and Genetics. World scientific, 2004 , 612 — 636

Moore SS, Whan V, Davis GP, Byrne K, Hetzel DJ., Preston N. The development and application of genetic markers for the Kuruma prawn Penaeus japonicus. Aquaculture. 1999, 173(1-4):19–32.

Nelson J S. Fishes of the world. New York: John Wiley and Sons Inc. 1994 , 2

Otto SP: The evolutionary consequences of polyploidy. Cell 2007, 131: 452–462.

Paterson AH, Bowers JE, Burow MD, Draye X, Elsik CG, Jiang C-X, et al. Comparative Genomics of Plant Chromosomes. Plant Cell. 2000, 12(9):1523–39.

Peatman E, Terhune J, Baoprasertkul P, Xu P, Nandi S, Wang S, 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(2):553–66.

Postlethwait J H. Evolution of the Zebrafish Genome. In: Gong Z, Korzh V (Eds.) , Fish Development and Genetics. World Scientific, 2004 ,581 —611

Sarropoulou E, Fernandes JMO. Comparative genomics in teleost species: Knowledge transfer by linking the genomes of model and non-model fish species. Comparative Biochemistry and Physiology Part D: Genomics and Proteomics. 2011, 6(1):92–102.

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

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. 2004, 6(4):364–77.

Somridhivej B, Wang S, Sha Z, Liu H, Quilang J, Xu P, et al. 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(1-4):76–80.

Souciet J-L, Dujon B, Gaillardin C, Johnston M, Baret PV, Cliften P, et al. Comparative genomics of protoploid Saccharomycetaceae. Genome Res. 2009, 19(10):1696–709.

Wakamatsu Y, Pristyazhnyuk S, Kinoshita M, et al . The see-through medaka: a fish model that is transparent throughout life. Proc Natl Acad Sci USA, 2001 , 98 : 10046 —10050

Wittbrodt J, Shima A, Schartl M. Medaka2a model organismfrom the far East. Nat Rev Genet, 2002, 3 : 53 —64

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

Young WP, Wheeler PA, Coryell VH, Keim P, Thorgaard GH. A Detailed Linkage Map of Rainbow Trout Produced Using Doubled Haploids. Genetics. 1998, 148(2):839–50.

Zhao G P. Bioinformatics. Beijing : Academic Press. 2002, 3 — 12

Zou J, Clark M S, Secombes C J. Characterization, expression and promoter analysis of an interleukin 10 homologue in the puffer fish, Fugu rubripes. Immunogenetics, 2003, 55: 325 — 335

Table 2.1. Major fish genomic databases

Service Names	Detail	Uniform Resource Locator
NCBI (National Center for Biotechnology Information)	MEGABLAST Discontiguous MEGABLAST Blastp (Talapia, fugu, zebrafish)	http :/ / www.ncbi.nlm.nih.gov/ Traces/ trace.cgi ? cmd = show&f = blast&m = blast&s = blast
The Sanger Institute	Zebrafish	http :/ / www.sanger.ac.uk/ cgibin/ blast/ submitblast/ d rerio/ ad vanced
ENSEMBL (European Bioinformatics Institute and The Sanger Institute)	Medaka stickleback fugu zebrafish and other fish, mammals	http :/ / www.ensembl.org/ index.html
TIGR (USA , The Institute for Genomic Research)	Catfish, Atlantic salmon, medaka, fugu EST sequence comparison	http :/ / tigrblast.tigr.org/ tgi/
Mbase (medaka genome database , university of Tokyo	Medaka and its EST sequence including BAC database	http :/ / mbase.bioweb.ne.jp/ ～ dclust/ medaka_top.html
(Washington University)	Zebrafish EST sequence	http :/ / www.genetics.wustl.edu/ fish lab/ frank/ cgibin/ fish/
(Japan , National Institute of Genetics)	Medaka gene database	http :/ / dolphin.lab.nig.ac.jp/ medaka

Table 3.1. Identification of genes in linkage group 27 (LG27) using information from mapping of BAC end sequences, the physical map, and BLAST analysis (E-value≦1e-10)

Category	Numbers
BAC-associated markers mapped in LG 27	68
Physical map contigs containing the BAC-associated markers	60
All available BAC-end sequences (BES) within the 60 contigs	1,899
Number of BES with significant hits to draft genome sequences	1,183
Draft genome contigs with significant hits by BES	786
Significant hits to zebrafish genes	196

Table 3.2. Chromosome distribution of the 196 genes on catfish LG27 in four teleost fish species with whole genome sequences. (The largest number of genes of catfish LG27 located on a single chromosome is underlined. Genes on unassigned scaffolds were not included)

Chromosome/ Group	**No. of gene hits in zebrafish**	**No. of gene hits in medaka**	**No. of gene hits in stickleback**	**No. of gene hits in *Tetraodon***
1	4	4	2	**32**
2	8	0	5	2
3	1	3	9	0
4	8	**48**	4	2
5	4	**43**	0	3
6	**118**	5	3	0
7	4	6	5	2
8	2	3	**54**	1
9	3	4	1	2
10	3	1	2	2
11	3	0	3	**40**
12	3	2	4	6
13	2	1	3	4
14	6	3	3	0
15	0	1	2	**12**
16	3	2	2	0
17	3	**16**	**37**	3
18	0	2	2	1
19	0	0	5	1
20	2	0	1	2
21	0	4	0	
22	5	5		
23	2	3		
24	2	1		
25	5			
Total	191	157	147	115

Table 3. 3. Annotation of catfish genes mapped in LG27 with significant hits to zebrafish chromosome 6. (All gene IDs start with ENSDARG000000- and therefore this string is omitted in the table)

Physical map Contig No.	Gene ID	Zebrafish gene start on Chr6	Gene description
Contig0017	75297	43,157,095	tRNA nucleotidyl transferase, CCA-adding, 1
Contig0017	14181	43,529,979	Forkhead box P1b
Contig0032	20979	51,663,895	Family with sequence similarity 65, member C
Contig0032	13078	52,163,393	Tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, beta polypeptide a
Contig0032	1557	52,177,501	Translocase of outer mitochondrial membrane 34
Contig0032	14239	52,214,055	Extended synaptotagmin-like protein 1b
Contig0122	19405	45,430,455	Contactin 4
Contig0204	8064	2,237,708	Mediator complex subunit 8
Contig0204	17339	19,951,106	Glutamate-ammonia ligase (glutamine synthase) b
Contig0204	61923	27,700,171	Angiomotin like 2a
Contig0204	56924	27,726,618	Sin3A-associated protein a
Contig0204	44441	27,762,654	Myosin VIIB
Contig0204	62030	27,832,621	G protein-coupled receptor 17
Contig0204	74283	27,849,193	Inositol polyphosphate-5-phosphatase, 145kDa
Contig0204	70864	27,924,715	B-cell CLL/lymphoma 6a (zinc finger protein 51)
Contig0204	23578	28,007,995	LIM domain containing preferred translocation partner in lipoma
Contig0204	56026	28,466,739	Tumor protein p63 regulated 1
Contig0204	44356	28,509,576	Tumor protein p63
Contig0230	42123	54,109,965	Nudix (nucleoside diphosphate linked moiety X)-type motif 3b
Contig0230	29930	54,235,523	SAM pointed domain containing ets transcription factor
Contig0230	78075	54,277,985	Chromosome 6 open reading frame 106
Contig0230	78210	54,494,106	Tubby like protein 1
Contig0230	63649	54,531,678	TEA domain family member 3
Contig0230	59854	55,251,293	Uncharacterized protein
Contig0230	75815	55,332,429	Solute carrier family 12 (potassium/chloride transporter), member 5
Contig0230	77028	55,725,962	Uncharacterized protein
Contig0230	18984	55,881,326	Eyes absent homolog 2 (Drosophila)
Contig0230	40607	56,048,378	Chromosome 20 open reading frame 43
Contig0230	40608	56,069,653	Glucosaminyl (N-acetyl) transferase family member 7
Contig0230	40606	56,113,382	Transcription factor AP-2 gamma (activating enhancer binding protein 2 gamma)
Contig0230	87149	58,590,477	Uncharacterized protein
Contig0247	39213	33,624,474	PRP38 pre-mRNA processing factor 38 (yeast) domain containing A
Contig0247	73768	33,780,758	Transmembrane protein 125
Contig0247	59575	34,685,403	U2AF homology motif (UHM) kinase 1
Contig0247	51913	34,714,078	Discoidin domain receptor tyrosine kinase 2a
Contig0247	41072	34,986,514	Mesoderm induction early response 1 homolog a (Xenopus laevis)
Contig0247	70074	35,013,405	Low density lipoprotein receptor-related protein 8, apolipoprotein e receptor
Contig0247	70013	35,599,989	Doublesex and mab-3 related transcription factor 2b
Contig0301	74690	48,367,149	Protein phosphatase, Mg2+/Mn2+ dependent, 1J
Contig0415	21032	31,784,202	Forkhead box D3
Contig0415	44373	31,992,762	Autophagy-related 4C (yeast)
Contig0415	78675	32,018,244	Dedicator of cytokinesis 7
Contig0425	74148	52,710,149	Recombination signal binding protein for immunoglobulin kappa J region-like
Contig0425	15947	52,742,715	Matrilin 4
Contig0425	92236	52,880,526	Odorant receptor, family H, subfamily 136, member 2, pseudogene
Contig0579	58675	3,522,792	Phosphatase, orphan 2
Contig0579	27419	3,642,627	Glutamate decarboxylase 1b
Contig0579	86043	8,555,706	Uncharacterized protein
Contig0716	9830	26,583,067	High density lipoprotein-binding protein (vigilin)
Contig0716	16429	26,658,793	FERM, RhoGEF and pleckstrin domain protein 2
Contig0716	52104	26,799,101	ATG4 autophagy related 4 homolog B (S. cerevisiae)
Contig0716	52103	26,813,190	Deoxythymidylate kinase (thymidylate kinase)
Contig0767	74625	59,823,389	Rho GTPase activating protein 40
Contig1339	91831	57,386,390	DEAD (Asp-Glu-Ala-Asp) box polypeptide 27
Contig1606	87900	20,031,685	Zinc finger protein 648

Contig1606	52046	20,034,682	Im:7147241
Contig1606	69983	20,292,392	Scinderin like a
Contig1606	61162	20,542,346	Epoxide hydrolase 4
Contig1606	20076	20,578,126	Bromodomain, testis-specific
Contig1606	78670	20,632,844	Transforming growth factor, beta receptor III
Contig1606	52254	20,874,073	Zinc finger protein 644
Contig1638	9677	29,575,056	Discs, large (Drosophila) homolog 1
Contig1638	75159	29,805,607	Antigen p97 (melanoma associated) identified by monoclonal antibodies 133.2 and 96.5
Contig1638	79670	29,838,528	Leucine rich repeat containing 7
Contig1738	15471	36,792,721	Chromodomain helicase DNA binding protein 1-like
Contig1808	76416	30,119,684	Shroom family member 2
Contig1808	44406	30,318,149	Chromosome 1 open reading frame 190
Contig1808	70961	30,876,228	Leptin receptor
Contig1843	60235	42,407,140	Macrophage stimulating 1 receptor (c-met-related tyrosine kinase), a
Contig1895	32868	30,543,718	Phosphodiesterase 4B, cAMP-specific
Contig1895	53447	44,128,649	Protein phosphatase 4, regulatory subunit 2b
Contig1895	73869	44,175,907	PDZ domain containing RING finger 3
Contig1895	53454	44,521,953	Contactin 6
Contig1895	76233	44,850,988	Cell adhesion molecule with homology to L1CAM (close homolog of L1)
Contig1895	77141	45,236,933	Uncharacterized protein
Contig1973	77004	23,758,526	Aldehyde dehydrogenase 1 family, member L1
Contig1978	89886	23,545,425	Wu:fb82g02
Contig1978	17211	23,559,591	Plexin b1b
Contig1978	52351	23,711,125	Zinc finger, X-linked, duplicated A
Contig2049	44146	21,383,889	Leucine rich repeat containing 8 family, member Db
Contig2049	59718	21,400,489	Leucine rich repeat containing 8 family, member C
Contig2159	44365	32,036,367	Angiopoietin-like 3
Contig2431	74028	57,348,406	Zinc finger, NFX1-type containing 1
Contig2431	11125	57,428,413	Small nuclear ribonucleoprotein polypeptides B and B1
Contig2431	74903	57,442,386	Itchy E3 ubiquitin protein ligase homolog (mouse)
Contig2431	74794	57,533,945	N-terminal EF-hand calcium binding protein 3
Contig2431	74337	57,584,825	Core-binding factor, runt domain, alpha subunit 2; translocated to, 2
Contig2431	90781	57,638,387	Syntrophin, alpha 1 (dystrophin-associated protein A1, 59kDa, acidic component)
Contig2431	35773	57,794,691	TOX high mobility group box family member 2
Contig2506	90260	24,749,737	Uncharacterized protein
Contig2602	44601	6,147,827	Reticulon 4a
Contig2602	76351	34,008,034	Family with sequence similarity 5, member C
Contig2602	61051	35,909,603	Family with sequence similarity 5, member C
Contig2602	44326	36,243,919	Uncharacterized protein
Contig2603	70037	34,259,457	Regulator of G-protein signaling 8
Contig2603	70047	34,379,926	Regulator of G-protein signalling 4
Contig2603	78725	34,436,646	Nitric oxide synthase 1 (neuronal) adaptor protein a
Contig2603	88140	34,866,637	Hydroxysteroid (17-beta) dehydrogenase 7
Contig2660	730	41,049,773	Solute carrier family 6 (neurotransmitter transporter, GABA), member 13
Contig2660	70567	41,186,467	Ca++-dependent secretion activator
Contig2660	94792	41,394,649	Twinfilin, actin-binding protein, homolog 1 (Drosophila), like
Contig2660	78894	41,441,001	HemK methyltransferase family member 1
Contig2660	60309	41,486,570	SLIT-ROBO Rho GTPase activating protein 3
Contig2660	27938	41,739,537	RAD18 homolog (S. cerevisiae)
Contig2660	33956	41,888,144	Oxytocin receptor
Contig2660	4150	41,982,782	Glutamate receptor, metabotropic 2
Contig2660	53574	42,077,037	Fanconi anemia, complementation group D2
Contig2701	75268	56,702,053	Cadherin 22, type 2
Contig2703	52025	32,774,484	Protein O-linked mannose beta1,2-N-acetylglucosaminyltransferase
Contig2703	4105	33,680,462	Endothelium-specific receptor tyrosine kinase 1
Contig2797	76810	2,810,030	Protein tyrosine phosphatase, receptor type, f, a
Contig2798	77753	2,262,173	Seizure threshold 2 homolog
Contig2798	30166	2,422,955	Hydroxypyruvate isomerase (putative)
Contig2798	18782	3,089,746	Lysine (K)-specific demethylase 4A
Contig2798	77764	3,324,243	Fatty acid amide hydrolase
Contig2798	7231	27,222,693	Receptor-like tyrosine kinase
Contig2798	61896	27,332,251	Solute carrier organic anion transporter family, member 2A1
Contig2798	34522	27,375,544	RAB6B, member RAS oncogene family a

Table 3.4. Summary of conserved syntenic blocks between catfish LG 27 and zebrafish chromosome 6

Syntenic block	Catfish physical contigs involved	Number of genes involved	Spanning size in zebrafish genome
1	Contig0017	2	373Kb
2	Contig0032	4	550Kb
3	Contig0415	3	234Kb
4	Contig0425	3	170Kb
5	Contig0716	4	230Kb
6	Contig1606	7	842Kb
7	Contig1638	3	263Kb
8	Contig1808	3	757Kb
9	Contig1978	3	166Kb
10	Contig2049	2	17Kb
11	Contig2431	7	446Kb
12	Contig2603	4	607Kb
13	Contig2703	2	906Kb

Table 3.5. Catfish genes identified from the genome sequence contigs that are directly linked to BAC end sequences with markers mapped to LG27

Marker ID	Position	Gene Start position on zebrafish Chr6	Gene identity
AUBES5170	69.569	2,422,955	ENSDARG00000030166
AUBES3963	40.132	3,642,627	ENSDARG00000027419
AUBES5085	50.451	8,555,706	ENSDARG00000086043
AUBES4135	68.49	20,578,126	ENSDARG00000020076
AUBES4550	67.79	23,559,591	ENSDARG00000017211
AUBES5200	67.775	23,711,125	ENSDARG00000052351
AUBES4719	68.197	23,758,526	ENSDARG00000077004
AUBES2504	68.092	27,375,544	ENSDARG00000034522
AUBES4097	68.092	28,007,995	ENSDARG00000023578
AUBES4461	59.924	34,379,926	ENSDARG00000070047
AUBES3409	61.141	34,436,646	ENSDARG00000078725
AUBES3408	78.06	34,685,403	ENSDARG00000059575
AUBES3409	61.141	34,866,637	ENSDARG00000088140
AUBES1145	66.543	45,430,455	ENSDARG00000019405
AUBES4122	67.775	48,367,149	ENSDARG00000074690
AUBES3056	67.773	54,235,523	ENSDARG00000029930
AUBES3488	67.773	57,533,945	ENSDARG00000074794
AUBES3487	67.751	57,794,691	ENSDARG00000035773

Table 3.6: Annotation of genes anchored markers in LG27(Medaka)

Marker ID	Position	Gene Start (bp)	Chromosome	Protein_ID
AUBES5170	69.569	6267703	4	ENSORLP00000003950
AUBES3408	78.06	26271077	4	ENSORLP00000017982
AUBES4097	68.092	29311777	4	ENSORLP00000019533
AUBES3487	67.751	10461340	5	ENSORLP00000007478
AUBES3056	67.773	13666432	5	ENSORLP00000009134
AUBES4719	68.197	17351502	5	ENSORLP00000011903
AUBES5200	67.775	17367655	5	ENSORLP00000011920
AUBES4550	67.79	18506799	5	ENSORLP00000012643
AUBES3963	40.132	15136163	17	ENSORLP00000011549
AUBES2504	68.092	25424974	17	ENSORLP00000020394

Table 3.7: Annotation of genes anchored markers in LG27(Stickleback fish)

Marker ID	Position	Chromosome	Gene Start (bp)	Protein ID
AUBES5170	69.569	8	302175	ENSGACP00000003910
AUBES3408	78.06	8	1556600	ENSGACP00000004623
AUBES2504	68.092	8	1830464	ENSGACP00000004908
AUBES4097	68.092	8	2005357	ENSGACP00000005049
AUBES4135	68.49	8	2431974	ENSGACP00000005692
AUBES3409	61.141	8	4978476	ENSGACP00000007499
AUBES4461	59.924	8	5098755	ENSGACP00000007574
AUBES3056	67.773	17	2990590	ENSGACP00000006489
AUBES4122	67.775	17	3666388	ENSGACP00000007146
AUBES3488	67.773	17	4560969	ENSGACP00000008028
AUBES3487	67.751	17	4657545	ENSGACP00000008056
AUBES4719	68.197	17	12102050	ENSGACP00000014952
AUBES5200	67.775	17	12156506	ENSGACP00000014976
AUBES4550	67.79	17	12285324	ENSGACP00000015048

Table 3.8: Annotation of genes anchored markers in LG27(*Tetraodon*)

Marker ID	Position	Chromosome	Gene Start (bp)	Protein ID
AUBES3409	61.141	1	12850657	ENSTNIP00000016155
AUBES4097	68.092	1	10565639	ENSTNIP00000016774
AUBES3488	67.773	11	4169288	ENSTNIP00000017874
AUBES4122	67.775	11	8423258	ENSTNIP00000021445
AUBES3056	67.773	11	8958564	ENSTNIP00000021475
AUBES4461	59.924	15	3503873	ENSTNIP00000009734
AUBES3963	40.132	15	2607432	ENSTNIP00000010715
AUBES2504	68.092	15	5841583	ENSTNIP00000018316

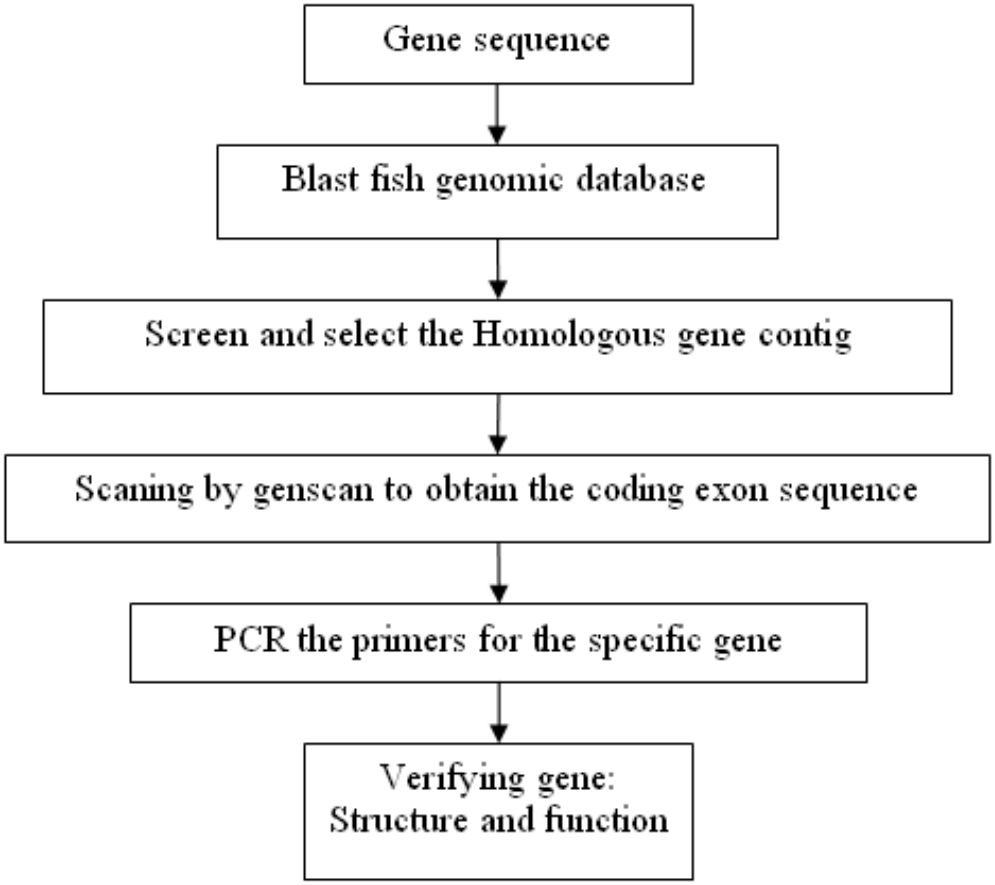

Figure 2.1. The strategy of homology searching for fish genes in genomic database

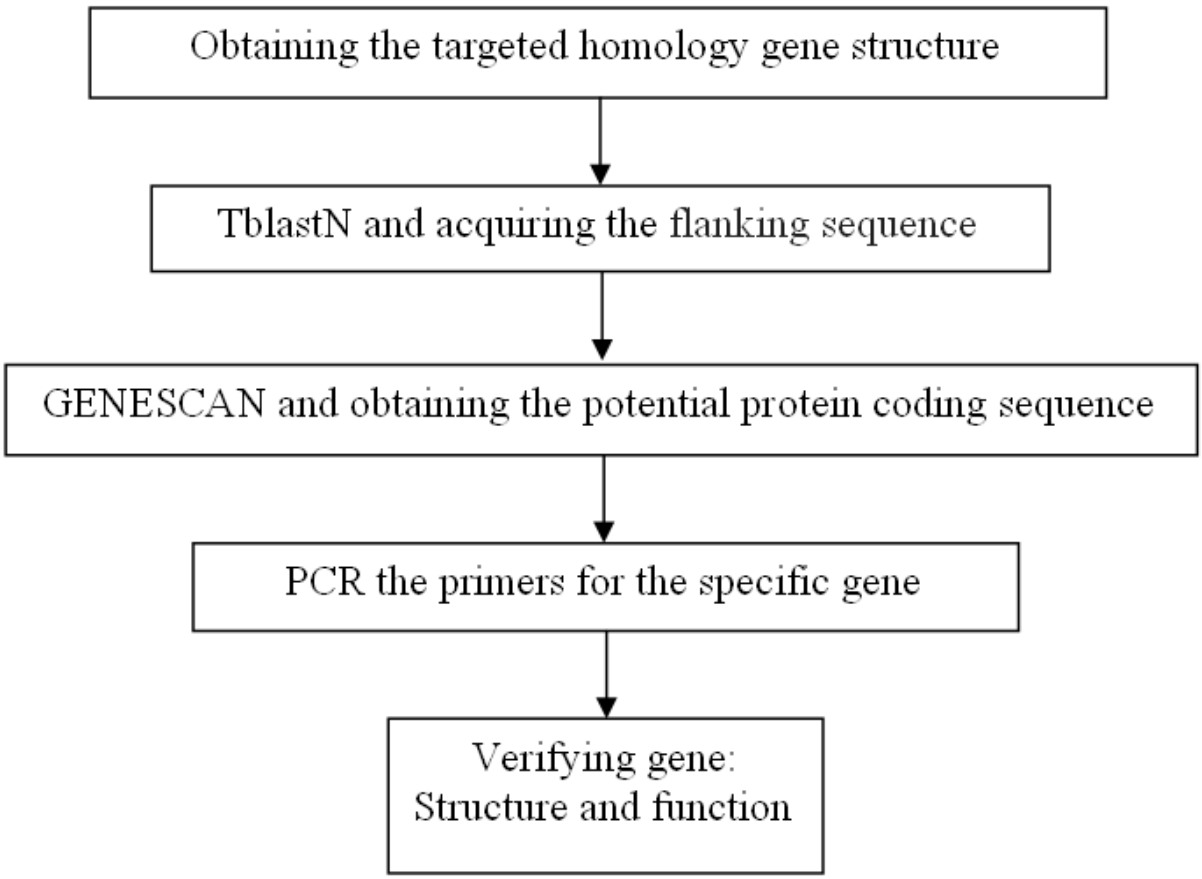

Figure 2.2. The strategy of conserved synteny analysis of fish genes in genomic database

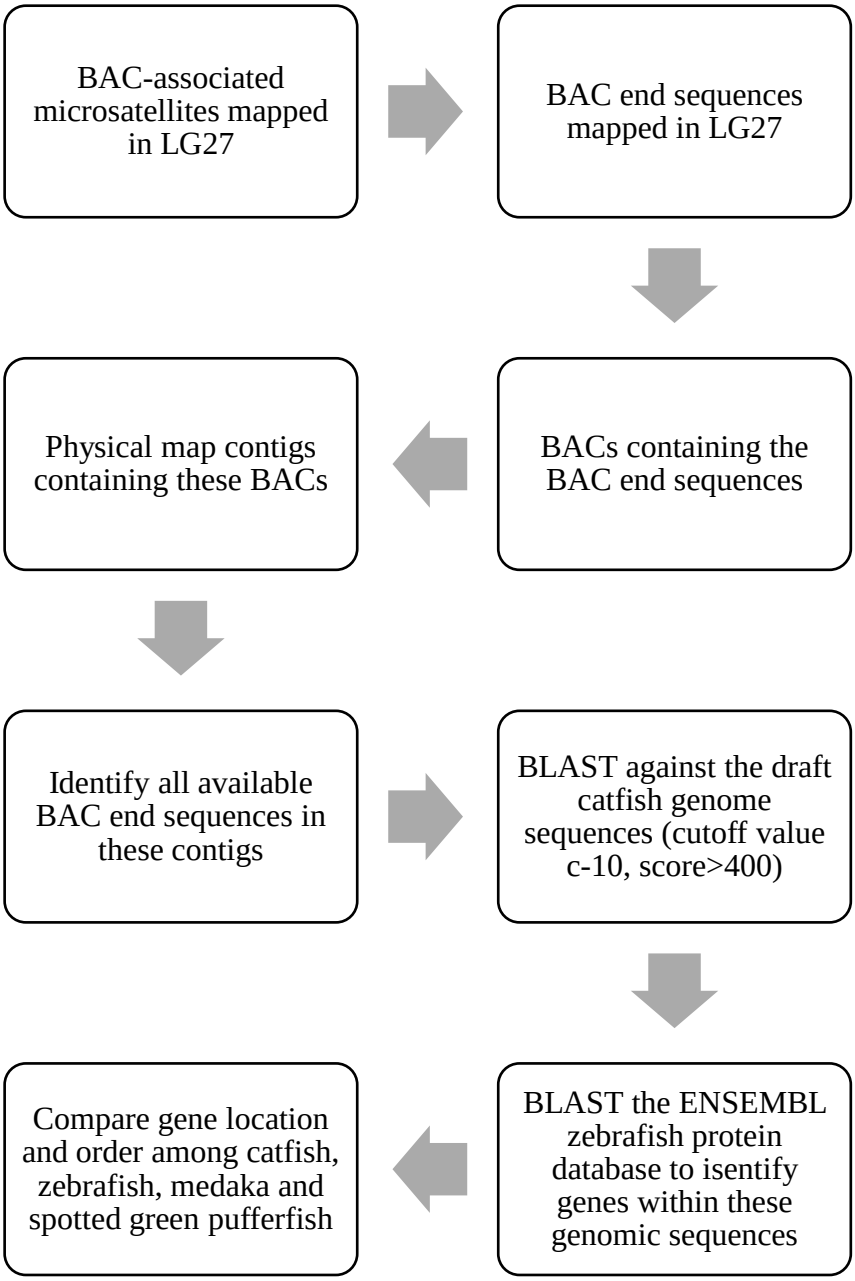

Figure 3.1. The process of experimental methods.

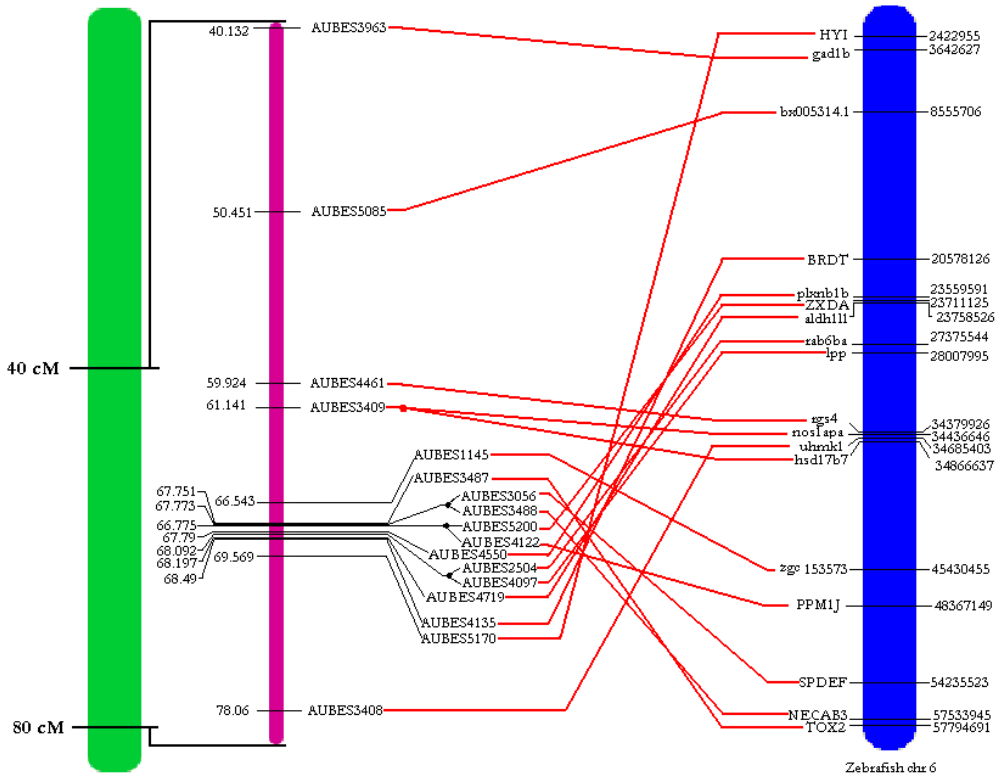

Figure 3.2: Conserved syntenic regions between catfish LG27 and zebrafish chromosome 6.

Conserved syntenic regions were established by genetic linkage mapping of BES-associated microsatellites. The catfish LG27 is presented with markers and markers positions on the furthest left. The numbers on left of LG27 are markers positions (cM) and the ID on the right of LG27 are markers ID. The zebrafish 6 (chr 6) is presented on the furthest right. The left column of chromosomes shows the associated gene names that searched by Biomart. The right column of chromosomes shows these gene positions on the chromosomes. The detailed information was listed in table 3.5.

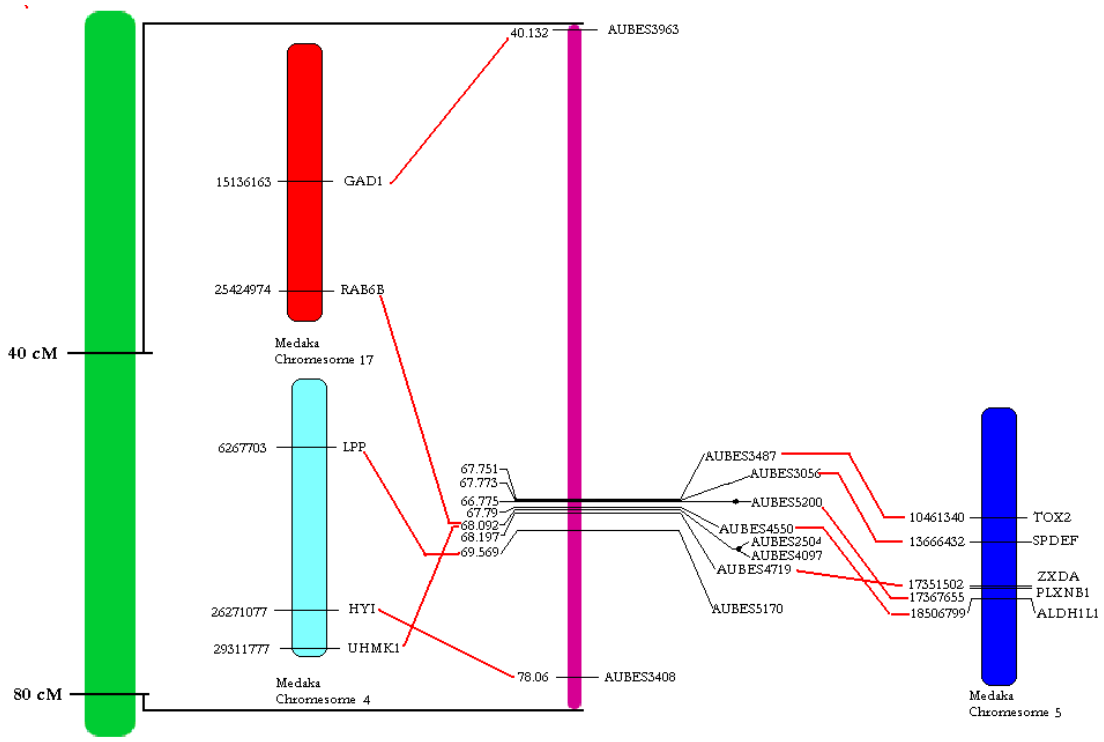

Figure 3.3: Conserved syntenic regions between catfish LG27 and Medaka chromosome 4,5and17

Conserved syntenic regions were established by genetic linkage mapping of BES-associated microsatellites. The catfish LG27 is presented with markers and markers positions on the furthest left. The numbers on left of LG27 are markers positions (cM) and the ID on the right of LG27 are markers ID. The Medaka chromosome 4,5 and17 are presented on the furthest right. The left column of chromosomes shows the associated gene names that searched by Biomart. The right column of chromosomes shows these gene positions on the chromosomes. The detailed information was listed in table 3.6.

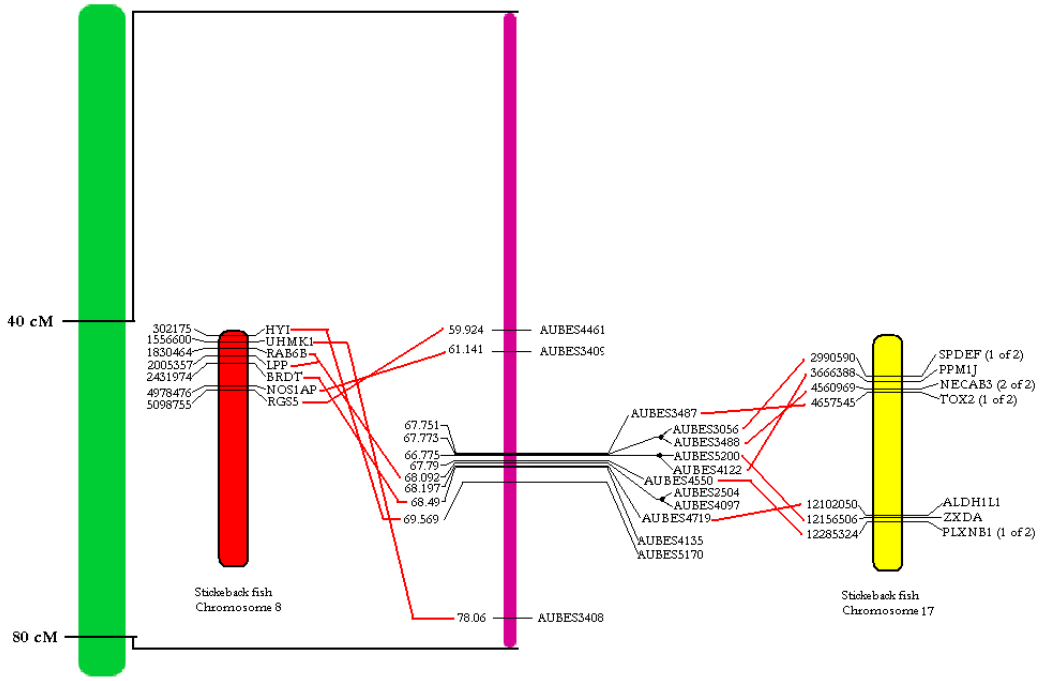

Figure 3.4: Conserved syntenic regions between catfish LG27 and Stickleback fish chromosome 8 and 17

Conserved syntenic regions were established by genetic linkage mapping of BES-associated microsatellites. The catfish LG27 is presented with markers and markers positions on the furthest left. The numbers on left of LG27 are markers positions (cM) and the ID on the right of LG27 are markers ID. The Stickleback fish chromosome 8 and 17 are presented on the furthest right. The left column of chromosomes shows the associated gene names that searched by Biomart. The right column of chromosomes shows these gene positions on the chromosomes. The detailed information was listed in table 3.7.

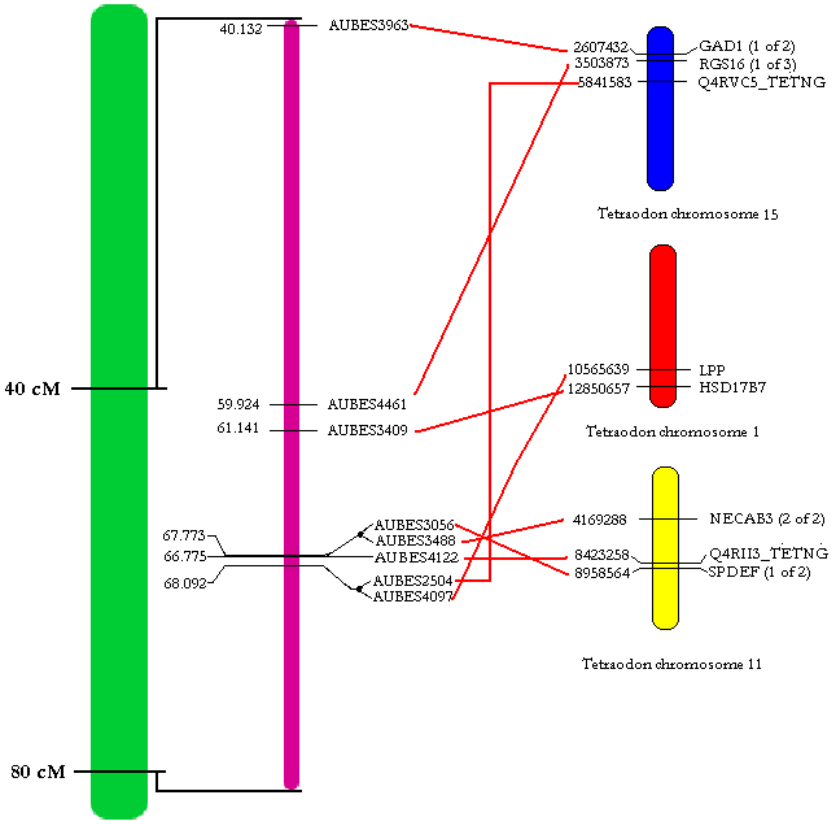

Figure 3.5: Conserved syntenic regions between catfish LG27 and *tetraodon* chromosome 1, 11 and 15

Conserved syntenic regions were established by genetic linkage mapping of BES-associated microsatellites. The catfish LG27 is presented with markers and markers positions on the furthest left. The numbers on left of LG27 are markers positions (cM) and the ID on the right of LG27 are markers ID. The *tetraodon* chromosome 1, 11 and 15 are presented on the furthest right. The left column of chromosomes shows the associated gene names that searched by Biomart. The right column of chromosomes shows these gene positions on the chromosomes. The detailed information was listed in table 3.8.