PHYSICAL MAP CONSTRUCTION AND PHYSICAL CHARACTERIZATION OF

CHANNEL CATFISH GENOME

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

_____

Peng Xu

Certificate of Approval:

_____	_____
Nannan Liu Associate Professor Entomology and Plant Pathology	Zhanjiang (John) Liu, Chair Professor Fisheries and Allied Aquacultures
_____	_____
Narendra Singh Professor Biology	Covadonga Arias Associate Professor Fisheries and Allied Aquacultures

_____

Joe Pittman
Interim Dean
Graduate School

PHYSICAL MAP CONSTRUCTION AND PHYSICAL CHARACTERIZATION OF CHANNEL CATFISH GENOME

Peng Xu

A Dissertation

Submitted to

the Graduate Faculty of

Auburn University

in Partial Fulfillment of the

Requirements for the

Degree of

Doctor of Philosophy

Auburn, Alabama
August 4, 2007

PHYSICAL MAP CONSTRUCTION AND PHYSICAL CHARACTERIZATION OF CHANNEL CATFISH GENOME

Peng Xu

_____
Signature of Author

_____
Date of Graduation

VITA

Peng Xu, son of Shunze Xu and Xiaoqing Wang, was born March 26, 1977, in Minquan, China. He graduated with a Bachelor of Science degree in 1999 from Xiamen University, China majoring in Biology, and with a Master degree of Science in 2002 from the Institute of Oceanology, Chinese Academy of Sciences at Qingdao, China majoring in Marine Biology. After working for one year in the Institute of Oceanology, Chinese Academy of Sciences as research scientist, he entered Graduate School of Auburn University in the Department of Fisheries and Allied Aquacultures, to pursue a Doctor of Philosophy degree in August 2003.

DISSERTATION ABSTRACT

PHYSICAL MAP CONSTRUCTION AND PHYSICAL CHARACTERIZATION OF CHANNEL CATFISH GENOME

Peng Xu

Doctor of Philosophy, August 4, 2007
(M.Sc., Chinese Academy of Sciences, China 2002)
(B.Sc., Xiamen University, 1999)

111 Typed Pages

Directed by Zhanjiang (John) Liu

Catfish is the major aquaculture species in the United States. To enhance genome studies involving linkage mapping, comparative mapping and linkage map and physical map integration, over 20,000 Bacterial Artificial Chromosome (BAC) end sequences were generated and a BAC-based physical map of the channel catfish (*Ictalurus punctatus Rafinesque*) genome was constructed using four color fluorescence-based fingerprinting.

A total of 25,195 BAC ends were sequenced, generating 20,366 clean BAC end sequences (BES) with an average reading length of 557 bp. The total reading length of 11,414,601 bp represented approximately 1.2% of the catfish genome. Based on this survey, the catfish genome was found to be highly AT-rich with 60.7% A/T. Approximately 12% of the catfish genome consisted of dispersed repetitive elements with

the *TC1* transposons making up the largest percentage by base pair (4.57%).

Microsatellites were detected in 17.5% of BAC end sequences, providing valuable resources for marker development and map integration. BAC end sequences were anchored to the zebrafish and Tetraodon genome by Basic Local Alignment Search Tool (BLAST) search, revealing 16% and 8.2% significant hits ($E < e^{-5}$), respectively. The mate-paired BAC end sequences were used to compare with zebrafish and Tetraodon genome and identified 23 conserved syntenies.

A total of 40,416 BAC clones were fingerprinted, generating 34,580 (84.3% success rate, 5.6X genome coverage) validated fingerprints for the FPC (Fingerprinted Contig) assembly. A total of 3,307 contigs were assembled using a cutoff value of $1e^{-20}$ and size deviation tolerance of 0.4 bp. Each contig contained an average of 9.25 clones, with an average size of 292 kb. The combined contig size for all contigs represents approximately 1X genome size of the channel catfish. The accuracy of the contig assembly was assessed by hybridizations on BAC library high-density filters using overgo probes from a set of genes, followed by determination of positive clones in comparison with their locations in the assembly. The constructed physical map should provide an important and powerful tool for genomic studies in catfish, including comparative mapping among various fish species.

ACKNOWLEDGMENTS

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

Style manual used Genome Research

Computer software used Microsoft Word 2003, Microsoft Excel 2003, Adobe Photoshop CS2, Genoprofiler, Genemapper, FPminer, Repeatmasker, Phred.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

I. INTRODUCTION

Catfish is the leading aquaculture industry in the United States accounting for over 60% of all US aquaculture production. Catfish possesses a combination of biological and cultural attributes that make them an excellent fish for commercial aquaculture, e.g., easy to produce seeds, manipulation of spawning, artificial spawning, easy to culture, tolerance to crowding and wide range of environmental conditions including salinities, temperature and dissolved oxygen, and efficient feed conversion. Channel catfish have firm, white flesh with a mild flavor that retains high sensory quality after a variety of processing methods on industrial scale.

Of the cultured catfish in the US, channel catfish (*Ictalurus punctatus Rafinesque*) is the major cultured catfish species. The closely related specie, blue catfish (*I. furcatus Valenciennes*) is also considered important because of its ability to produce hybrid catfish with channel catfish, which has the characteristics including improved capture by seining, resistance to some bacterial diseases, and tolerance to low concentrations of dissolved oxygen. In 2005, catfish production reached almost 700 million pounds in the US. Catfish aquaculture is also rapidly growing in Asia. In addition to its great economic importance for aquaculture, catfish is also one of the top sport fishing species in North America.

Channel catfish has long been also a research model for comparative immunology and toxicology. Its unique characteristics such as its ability to adapt to changing

environment, e.g. low oxygen, high pressure and changing temperature, are well beyond the capability of higher vertebrates such as mammals. Therefore, genomic studies of an aquaculture fish species might provide new insight addressing genetic mechanisms of performance traits in aquatic environments as well as genome evolution. Catfish is an excellent model organism for genomic studies, particularly for aquaculture important issues. Its high reproductive ability/fertility allows breeding of large families with thousands of progenies. This offers a great opportunity for quantitative trait locus (QTL) scans and extensive phenotype selection using selective genotyping thereby reducing time, money, and efforts in QTL analysis.

Rapid progress in catfish genomics has been made in the last several years. Large numbers of molecular markers have been developed and evaluated for linkage mapping (Serapion et al. 2004; Xu et al. 2006) and framework genetic linkage maps have been constructed (Liu et al. 2003; Waldbieser et al. 2001). Genome repeat structure has been characterized and several novel repetitive elements, e.g. *Xba* elements, TC-1 like elements tip1, tip2 and tipnon, short interspersed elements (SINE) *mermaid* and *merman*, were identified from catfish genome (Kim et al. 2000; Liu et al. 1999; Nandi et al. 2006; Xu et al. 2006). More than 55,000 expressed sequence tags (ESTs) have been generated (Cao et al. 2001; Ju et al. 2000; Karsi et al. 2002; Kocabas et al. 2004), and an ongoing large-scale EST project by the Joint Genome Institute of the Department of Energy will further significantly expand the EST resources in both channel catfish and blue catfish (He et al. 2003). Microarrays have been used to study genome-wide expression in catfish (Ju et al. 2002; Karsi et al. 2002; Li and Waldbieser 2006; Peatman et al. 2007). Two bacterial artificial chromosome (BAC) libraries using different restriction endonucleases

have been previously constructed and characterized (Quiniou et al. 2003, Wang et al, 2007). The channel catfish library CHORI-212 was used in my dissertation research for physical mapping and BAC end sequence generating. The genomic DNA isolated from one normal male channel catfish was used for BAC library CHORI-212's construction. Comparing with another BAC library CCBL1 which using the genomic DNA isolated from one genogen (duplicated haploid genome) female fish, CHORI-212 has the clones from Y chromosome, thus cover more genomic regions for channel catfish.

BAC LIBRARY CONSTRUCTION

Large insert genomic libraries are important genomic resources for many genomic studies, especially in the species with large and complex genome. There are several types of large insert genomic library, e.g. yeast artificial chromosome (YAC), bacterial artificial chromosome (BAC), bacteriophage P-1 derived artificial chromosome (PAC). YAC is an artificially constructed chromosome and contains the telomeric, centromeric, and replication origin sequences needed for replication and preservation in yeast cells. YAC can clone genomic DNA segments up to 3,000 kb (Burke et al. 1987; Larin et al. 1991); however, it has problems with clone stability. Recombination of noncontiguous DNAs may occur between YACs, and as a result, chimeric non-contiguous DNA fragments can be cloned (Green et al. 1991; Larionov et al. 1994). In addition, it is difficult to obtain enough YAC DNA for analysis by restriction digestion or for the production of small fragment shotgun libraries that are required for DNA sequencing. BAC and PAC are better alternative libraries for physical mapping, shot-gun genomic sequencing and other genomic analyses. Comparing with YAC, BAC and PAC have greater stability and have no problem with chimeric recombination. BAC DNA can be isolated by high-throughput

method (e.g. Qiagen R.E.A.L. Prep 96 kit), almost like regular plasmid isolation.

BACs are circular DNA molecules which contain a replicon based on the F factor. BACs have *ori*S and *rep*E which encode an ATP-driven helicase and *par*A, *par*B and *par*C for partitioning (Shizuya et al. 1992). Genes present in the F factor regulate BAC replication and control its copy number in the bacterial cells, which can efficiently prevent the rearrangement and recombination between those large insert DNA fragments, especially those DNA fragments retaining repetitive elements from eukaryotic organisms. Similar to plasmids, the cloning site is usually flanked by T7 and SP6 promoters for generating RNA probes for chromosome walking, and for DNA sequencing of the inserted segment at the vector-insert junction, known as BAC end sequencing. Drug resistant genes such as chloramphenicol or ampicillin resistance genes are used in the vector for selection. In some BAC vectors, the *LacZ* gene is also incorporated for blue/white colony selection system to exclude the empty clones. In some recently developed vectors, e.g. *pBACe3.6* and *pTARBAC* series, the *sacBII* gene is included as a positive-selection marker for the selection of recombinant clones (Osoegawa and de Jong 2004; Osoegawa et al. 1998; Zeng et al. 2001). The *sacBII* gene from *Bacillus amyloliquefaciens* was derived from Nat Sternberg's P1 vector (Pierce et al. 1992). The toxicity of the sacBII gene is a result of the conversion of fructose (derived from sacharose) into polyfructose (levan), which is toxic in *Escherichia coli*. The cloning vector and recombinant clones do not express the sacBII gene and hence are sucrose-resistant. Undesirable clones derived from the deleted cloning vector are sucrose-sensitive and, hence, do not form colonies on sucrose-containing media. The original BAC vectors are 7.4 kb while others are around 8 kb to 9 kb in length. BACs can

retain 80-300 kb DNA inserts.

High-molecular-weight (HMW) genomic DNA is important for the construction of high quality BAC library. Usually, intact cells are embedded in the agarose plugs and treated by proteinase K. The genomic DNA is dialyzed and collected from the agarose plug and partially digested by restriction enzyme (e.g. *Eco*RI and *Eco*RI methylase). The DNA segments with prospective sizes, usually in the range from 100 kb to 500 kb, are recovered from the agarose gels and collected for the library construction. Ligation and transformation are conducted to clone the DNA segments into prepared BAC vectors (Osoegawa et al. 1998). The following evaluation experiments are also necessary to access the BAC library quality and basic information such as average insert size, insert size distribution, non-recombination rate, and so on. The positive clones are picked from the agar plates and maintained in 384-well plates at -80^{o}C. High-density BAC library nylon membranes are usually made by automatic systems e.g. Genetix Qbot robot based on these 384-well plates, which can be used to screen and locate specific genes in the BAC library.

PHYSICAL MAP CONSTRUCTION

Physical map is the map of the locations of identifiable landmarks on DNA molecules (e.g. restriction enzyme cutting sites, genes), regardless of inheritance. Distance is measured in base pairs. The highest resolution map would be the complete nucleotide sequence of the genome.

In situ hybridization (ISH) or fluorescent in situ hybridization (FISH) is one of the physical mapping methods. This technique is used to detect the positions of the markers

or genes on chromosomes. The fluorescent labeled probes are hybridized to chromosomal DNA on the slides. The position of the fluorescence labeled probes is visualized on the chromosomes using the microscope or other methods (Jin and Lloyd 1997).

The major approach to physical mapping is restriction fragment overlapping based method. The large insert genomic libraries are digested using one or several restriction enzymes. The restriction fragments are separated using electrophoresis (agarose, polyacrylamide gels or capillary electrophoresis) (Coulson et al. 1986). The restriction profiles are recorded and submitted to the computing software. The overlapping clones are identified and built as contigs. Currently, the BAC-based physical mapping is probably the most popular method currently used because of the BAC advantages as described above. Compared with FISH, BAC-based physical mapping does not require the preparation of the metaphase chromosome slides. Also BAC-based physical mapping can map hundreds or thousands of genes to the contigs using BES or hybridization on BAC library. These genes will be located on the chromosomes once the physical map is integrated with the linkage map and the karyotypes. Obviously, FISH will never provide the same level of resolution as BAC-based physical mapping.

The traditional and widely used method is the agarose fingerprinting method, usually involves use of only one restriction enzyme (Gonzalez et al. 2005; Han et al. 2007; McPherson et al. 2001; Ng et al. 2005). This method does not require expensive equipment. The reagents are less expensive and the techniques are relatively easy. The agarose gel can detect most of the restriction fragments from 100 bp to more than 10 kb. However, this method is labor-intensive and time-consuming, which is hard to apply on high throughput physical mapping projects. Currently, fluorescence labeling is used in

more and more physical mapping projects. Automatic capillary electrophoresis equipments such as ABI PRISM series of genetic analyzers are used to analyze the fluorescence-labeled restriction segments. Up to four restriction enzymes can be used in this method; each kind of restriction fragments is labeled by one kind of fluorescent dye using SnapShot kit (Luo et al. 2003). This method can generate many more fragments for each clone than agarose method, which makes the overlapping assembly more efficient and reliable. The automatic genetic analyzer and supporting software, such as Genoprofiler (http://wheat.pw.usda.gov/PhysicalMapping) and FPminer (http://bioinforsoft.com/fpminer.html), make this method a high-throughput method for physical mapping.

Physical map assembly will be the next step after the fingerprint data are collected by either high-throughput methods or agarose electrophoresis methods. FingerPrinted Contigs (FPC) (Soderlund et al. 1997) is the most popular computer program for physical map assembly. The fingerprint data of chromatogram file from automatic genetic analyzers and image file from agarose gels need to be converted to proper format that can be recognized by FPC. Software IMAGE (http://www.sanger.ac.uk/Software/Image) was developed by Sanger Institute for fragment size calling from agarose gels. Software Genoprofiler was developed to process chromatogram data with multiple fluorescence labeling from automatic genetic analyzer. Commercial software FPminer had been developed by bioinforsoft LLC (http://bioinforsoft.com) for high-throughput automated DNA fragment analysis, which is very suitable for large scale physical mapping project. The accurate peak finding algorithms and peak size calling algorithms in FPminer can process fingerprint data and generate size files for FPC assembly in batch processing. The

quality score cutoff can remove low quality data and retain desired high quality fingerprints to ensure the accuracy of FPC assembly. The built-in database in FPminer makes it possible to process a large number of samples (http://www.bioinforsoft.com/).

FPC is an interactive program for building contigs from fingerprinted clones, where the fingerprint for a clone is a set of restriction fragments. FPC considers fragments to be shared by two BAC clones if they have the same size within a given tolerance. The algorithm in FPC will assemble clones together if they satisfy a user-defined cutoff value for fingerprint similarity based on the Sulston score (Sulston et al. 1988).

Proper tolerance and Sulston cutoff score must be determined before the formal FPC contig assembly. The tolerance level dictates how closely two fragments must match to be considered the same fragments. The fingerprinting experiments may be processed in different electrophoresis machines, or analyzed on different gels or at different time. The same fragment may have different called sizes in different runs due to these fluctuations even they should be considered as one fragment. Thus, the proper tolerance value in the FPC assembly must be determined to override the interference from those technical fluctuations. Generally the tolerance value for the fingerprint data from automatic genetic analyzer or capillary electrophoresis is around 0.4 bp (Katagiri et al. 2005; Luo et al. 2003; Nelson et al. 2005). For the fingerprint from agarose gel, the tolerance may be around several base pairs depending on electrophoresis condition (Han et al. 2007; Ng et al. 2005). The tolerance value can be determined experimentally. One or more specific fragments are analyzed several hundred times. The sizes are called from different electrophoresis and submitted for statistic analysis. The fragment size deviation on certain confidence level (e.g. 95% confidence interval) can determine the tolerance value

which should be used in FPC assembly (Katagiri et al. 2005; Nelson et al. 2005).

Sulston score cutoff is another parameter need to be determined before FPC assembly. The Sulston score corresponds to the probability that two fingerprints share similar fragment patterns by chance. It's always a dilemma on choosing a proper cutoff value. Too low a cutoff value splits true contigs into multiple contigs or singletons, whereas too high a cutoff value creates false contigs. Generally, the larger the genome, the lower the cutoff should be used. While a Sulston score of $3x\ 10^{-12}$ was used for the human genome (3×10^9 bp) for automated assembly (McPherson et al. 2001), a larger score of 1×10^{-9} was used for the smaller Arabidopsis genome (Mozo et al. 1999). The total number of fingerprints used in the assembly and the average fragment number for each fingerprint should also be considered while determining the cutoff value.

A BAC-based physical map is important for the understanding of genome structure and organization, and for position-based cloning of economically important genes. A well characterized physical map can often be an important foundation for whole genome sequencing. A BAC-based physical map would also allow exploitation of existing genomic information from map-rich species using comparative mapping. Physical maps had been constructed in aquaculture fish Nile tilapia (*Oreochromis niloticus Linnaeus*) and Atlantic salmon (*Salmo salar Linnaeus*) and gynogenetic catfish (Katagiri et al. 2005; Ng et al. 2005; Quiniou et al., 2007). This dissertation will report the construction of a BAC contig-based physical map of the normal channel catfish genome.

BAC END SEQUENCE ANALYSIS

BAC end sequences can be generated by direct sequencing of BAC clones using

sequencing primers based on the BAC vector sequences at the border of the genomic insert. Typically, Sp6 and T7 sequencing primers can be used because these promoter sequences have been incorporated into the BAC vectors. The sequencing reactions are straightforward using the dideoxy chain termination sequencing reactions (Sanger's sequencing method) except that a large number of cycles is required for cycle sequencing, usually 80–100 cycles (Xu et al. 2006) because of low copy number of BAC DNA.

BAC end sequences can provide an unbiased survey of genomic sequences. The genomic DNA is partially digested and inserted in the BAC vector. Assuming the restriction sites are randomly distributed in the genome, then the BAC end sequences can represent random genome sequences and provide an unbiased genomic survey, allowing estimation of A/T (G/C) content of the genome, assessment of the repeat structure of the genome, discovery of the microsatellite markers for linkage mapping, production of genomic resources for comparative mapping, and virtually mapping genes to BACs.

Microsatellite sequences can be mined from BAC end sequences for marker development. These BAC-derived microsatellite markers are also important genetic markers for linkage mapping and QTL mapping. Once these BAC-derived markers are mapped in linkage map, anchor points are built between physical map and linkage map. Collecting enough BAC-derived microsatellite markers is essential for physical map and linkage map integration (Barbosa et al. 2004; Danzmann et al. 2005; Faivre et al. 2002; Kiuchi et al. 2002; Varshney et al. 2006). In addition, BAC end sequences are also a potential resource to develop other genetic markers, such as SNP, which can be also used in the map integration.

BAC end sequences allow genes to be mapped to BAC clones virtually. The

BLASTX search of BAC end sequences against Non-redundant DNA database can identify potential gene contents in the BAC clones. Comparing with hybridization, BAC end sequencing and BLASTX analysis are more efficient and less labor-intensive. The different e-values in the BLASTX search may give us different levels of confidence. The higher the stringency, the more likely the significant putative genes are truly in the BAC clones. Once the physical map is constructed, these genes are mapped to the BAC contigs.

BAC end sequences also provide the sequence tag segments (STS) for comparative mapping which allow the anchoring of BACs or BAC contigs to those completely sequenced genomes (Childers et al. 2006). The conserved syntenies between closely-related species can be identified. Genomic organization of the species of interest can be inferred from that of a completely sequenced species.

BAC end sequences can also be used to build minimal tiling path (MTP) in the whole genome sequencing. MTP is a set of minimally overlapping BAC clones in the physical map picked for sequencing of the entire genome using the clone-by-clone approach. For most efficient whole genome sequencing, the ideal situation is to have minimally repeated sequencing and also cover all gaps so that the entire genome sequences can be assembled. A set of seed BAC clones are picked and completely sequenced. The BAC end sequences are queried for hits to the finished sequence. The BAC clones with minimal overlaps are picked for extending sequencing. The new set of MTP clones is sequenced, and new clones are picked from the ends of this set based on minimal overlapping with BAC end sequences. This process is repeated until the entire region is sequenced. Comparing with whole genome shotgun sequencing, sequencing using MTP

provide much more reliable genome sequence assembly.

BAC end sequencing is simple enough not to refer it as a genome technology. However, BAC end sequences are extremely rich in genome information. The analysis of BAC end sequences allows an unbiased sampling of the genome as to its composition and architecture. Bioinformatic mining of BAC end sequences allows the uncovering of the repeat structure of the genome and the identification of polymorphic microsatellites. Many markers discovered from BAC end sequences can be used to integrate the genetic linkage maps and the physical map as they can be placed on both maps. BAC end sequencing is one of the most efficient approaches to locate genes to physical maps. BAC end sequences can be exploited for comparative genome analysis including characterization of evolutionarily conserved syntenies. For many aquaculture species, the assessment of their genomes by BAC end sequencing is probably as good as they can get as their genomes may never be sequenced.

REFERENCES

Barbosa, A., O. Demeure, C. Urien, D. Milan, P. Chardon, and C. Renard. 2004. A physical map of large segments of pig chromosome 7q11-q14: comparative analysis with human chromosome 6p21. *Mamm Genome* 15: 982-995.

Burke, D.T., G.F. Carle, and M.V. Olson. 1987. Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science* **236:** 806-812.

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

Childers, C.P., H.L. Newkirk, D.A. Honeycutt, N. Ramlachan, D.M. Muzney, E. Sodergren, R.A. Gibbs, G.M. Weinstock, J.E. Womack, and L.C. Skow. 2006. Comparative analysis of the bovine MHC class IIb sequence identifies inversion breakpoints and three unexpected genes. *Anim Genet* **37:** 121-129.

Coulson, A., J. Sulston, S. Brenner, and J. Karn. 1986. Toward a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proc Natl Acad Sci U S A* **83:** 7821-7825.

Danzmann, R.G., M. Cairney, W.S. Davidson, M.M. Ferguson, K. Gharbi, R. Guyomard, L.E. Holm, E. Leder, N. Okamoto, A. Ozaki, C.E. Rexroad, 3rd, T. Sakamoto, J.B. Taggart, and R.A. Woram. 2005. A comparative analysis of the rainbow trout genome with 2 other species of fish (*Arctic charr* and *Atlantic salmon*) within the tetraploid derivative Salmonidae family (subfamily: Salmoninae). *Genome* **48:** 1037-1051.

Faivre, M., A.P. Rattink, B. Harlizius, R.P. Crooijmans, and M.A. Groenen. 2002. Porcine BAC derived microsatellites linked to ADRBK1, CNTF and GAL on SSC2. *Anim Genet* 33: 72-73.

Gonzalez, J., M. Nefedov, I. Bosdet, F. Casals, O. Calvete, A. Delprat, H. Shin, R. Chiu, C. Mathewson, N. Wye, R.A. Hoskins, J.E. Schein, P. de Jong, and A. Ruiz. 2005. A BAC-based physical map of the *Drosophila buzzatii* genome. *Genome Res* 15: 885-892.

Green, E.D., H.C. Riethman, J.E. Dutchik, and M.V. Olson. 1991. Detection and characterization of chimeric yeast artificial-chromosome clones. *Genomics* **11:** 658-669.

Han, Y., K. Gasic, B. Marron, J.E. Beever, and S.S. Korban. 2007. A BAC-based physical map of the apple genome. *Genomics*.

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

Jin, L. and R.V. Lloyd. 1997. In situ hybridization: methods and applications. *J Clin Lab Anal* **11:** 2-9.

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

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

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

Katagiri, T., C. Kidd, E. Tomasino, J.T. Davis, C. Wishon, J.E. Stern, K.L. Carleton, A.E. Howe, and T.D. Kocher. 2005. A BAC-based physical map of the Nile tilapia genome. *BMC Genomics* **6:** 89.

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

Kiuchi, S., Y. Inage, H. Hiraiwa, H. Uenishi, and H. Yasue. 2002. Assignment of 280 swine genomic inserts including 31 microsatellites from BAC clones to the swine RH map (IMpRH map). *Mamm Genome* 13: 80-88.

Kocabas, A., R. Dunham, and Z.J. Liu. 2004. Alterations in gene expression in the brain of white catfish (*Ameirus catus*) in response to cold acclimation. *Marine Biotechnology* **6:** 431-438.

Larin, Z., A.P. Monaco, and H. Lehrach. 1991. Yeast artificial chromosome libraries containing large inserts from mouse and human DNA. *Proc Natl Acad Sci U S A* **88:** 4123-4127.

Larionov, V., N. Kouprina, N. Nikolaishvili, and M.A. Resnick. 1994. Recombination during transformation as a source of chimeric mammalian artificial chromosomes in yeast (YACs). *Nucleic Acids Res* **22:** 4154-4162.

Li, R.W. and G.C. Waldbieser. 2006. Genomic organisation and expression of the natural

killer cell enhancing factor (NKEF) gene in channel catfish, *Ictalurus punctatus (Rafinesque)*. *Fish Shellfish Immunol* **20:** 72-82.

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

Liu, Z., P. Li, H. Kucuktas, and R. Dunham. 1999. Characterization of nonautonomous TC1-like transposable elements of channel catfish (*Ictalurus punctatus*). *Fish Physiology and Biochemistry* **21:** 65-72.

Luo, M.C., C. Thomas, F.M. You, J. Hsiao, S. Ouyang, C.R. Buell, M. Malandro, P.E. McGuire, O.D. Anderson, and J. Dvorak. 2003. High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. *Genomics* **82:** 378-389.

McPherson, J.D. M. Marra L. Hillier R.H. Waterston A. Chinwalla J. Wallis M. Sekhon K. Wylie E.R. Mardis R.K. Wilson R. Fulton T.A. Kucaba C. Wagner-McPherson W.B. Barbazuk S.G. Gregory S.J. Humphray L. French R.S. Evans G. Bethel A. Whittaker J.L. Holden O.T. McCann A. Dunham C. Soderlund C.E. Scott D.R. Bentley G. Schuler H.C. Chen W. Jang E.D. Green J.R. Idol V.V. Maduro K.T. Montgomery E. Lee A. Miller S. Emerling Kucherlapati R. Gibbs S. Scherer J.H. Gorrell E. Sodergren K. Clerc-Blankenburg P. Tabor S. Naylor D. Garcia P.J. de Jong J.J. Catanese N. Nowak K. Osoegawa S. Qin L. Rowen A. Madan M. Dors L. Hood B. Trask C. Friedman H. Massa V.G. Cheung I.R. Kirsch T. Reid R. Yonescu J. Weissenbach T. Bruls R. Heilig E. Branscomb A. Olsen N. Doggett J.F. Cheng T. Hawkins R.M. Myers J. Shang L. Ramirez J. Schmutz O. Velasquez K.

Dixon N.E. Stone D.R. Cox D. Haussler W.J. Kent T. Furey S. Rogic S. Kennedy S. Jones A. Rosenthal G. Wen M. Schilhabel G. Gloeckner G. Nyakatura R. Siebert B. Schlegelberger J. Korenberg X.N. Chen A. Fujiyama M. Hattori A. Toyoda T. Yada H.S. Park Y. Sakaki N. Shimizu S. Asakawa K. Kawasaki T. Sasaki A. Shintani A. Shimizu K. Shibuya J. Kudoh S. Minoshima J. Ramser P. Seranski C. Hoff A. Poustka R. Reinhardt and H. Lehrach. 2001. A physical map of the human genome. *Nature* **409:** 934-941.

Mozo, T., K. Dewar, P. Dunn, J.R. Ecker, S. Fischer, S. Kloska, H. Lehrach, M. Marra, R. Martienssen, S. Meier-Ewert, and T. Altmann. 1999. A complete BAC-based physical map of the *Arabidopsis thaliana* genome. *Nat Genet* **22:** 271-275.

Nandi, S., E. Peatman, P. Xu, S. Wang, P. Li, and Z. Liu. 2006. Repeat structure of the catfish genome: a genomic and transcriptomic assessment of Tc1-like transposon elements in channel catfish (*Ictalurus punctatus*). *Genetica*. published online.

Nelson, W.M., A.K. Bharti, E. Butler, F. Wei, G. Fuks, H. Kim, R.A. Wing, J. Messing, and C. Soderlund. 2005. Whole-genome validation of high-information-content fingerprinting. Plant Physiol 139: 27-38.

Ng, S.H., C.G. Artieri, I.E. Bosdet, R. Chiu, R.G. Danzmann, W.S. Davidson, M.M. Ferguson, C.D. Fjell, B. Hoyheim, S.J. Jones, P.J. de Jong, B.F. Koop, M.I. Krzywinski, K. Lubieniecki, M.A. Marra, L.A. Mitchell, C. Mathewson, K. Osoegawa, S.E. Parisotto, R.B. Phillips, M.L. Rise, K.R. von Schalburg, J.E. Schein, H. Shin, A. Siddiqui, J. Thorsen, N. Wye, G. Yang, and B. Zhu. 2005. A physical map of the genome of Atlantic salmon, *Salmo salar*. *Genomics* **86:** 396-404.

Osoegawa, K. and P.J. de Jong. 2004. BAC library construction. *Methods Mol Biol* **255:** 1-46.

Osoegawa, K., P.Y. Woon, B. Zhao, E. Frengen, M. Tateno, J.J. Catanese, and P.J. de Jong. 1998. An improved approach for construction of bacterial artificial chromosome libraries. *Genomics* **52:** 1-8.

Peatman, E., P. Baoprasertkul, J. Terhune, P. Xu, S. Nandi, H. Kucuktas, P. Li, S. Wang, B. Somridhivej, R. Dunham, and Z.J. Liu. 2007. Global assessment of the acute phase response in channel catfish (*Ictalurus punctatus*) after infection with Gram negative bacterium *Edwardsiella ictaluri*. *Physiological Genomics* **in review**.

Pierce, J.C., B. Sauer, and N. Sternberg. 1992. A positive selection vector for cloning high molecular weight DNA by the bacteriophage P1 system: improved cloning efficacy. *Proc Natl Acad Sci U S A* **89:** 2056-2060.

Quiniou, S.M., T. Katagiri, N.W. Miller, M. Wilson, W.R. Wolters, and G.C. Waldbieser. 2003. Construction and characterization of a BAC library from a gynogenetic channel catfish *Ictalurus punctatus*. *Genet Sel Evol* **35:** 673-683.

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

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

Shizuya, H., B. Birren, U.J. Kim, V. Mancino, T. Slepak, Y. Tachiiri, and M. Simon. 1992. Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. Proc Natl Acad Sci U S A 89:

8794-8797.

Soderlund, C., I. Longden, and R. Mott. 1997. FPC: a system for building contigs from restriction fingerprinted clones. *Comput Appl Biosci* **13:** 523-535.

Sulston, J., F. Mallett, R. Staden, R. Durbin, T. Horsnell, and A. Coulson. 1988. Software for genome mapping by fingerprinting techniques. *Comput Appl Biosci* **4:** 125-132.

Varshney, R.K., I. Grosse, U. Hahnel, R. Siefken, M. Prasad, N. Stein, P. Langridge, L. Altschmied, and A. Graner. 2006. Genetic mapping and BAC assignment of EST-derived SSR markers shows non-uniform distribution of genes in the barley genome. *Theor Appl Genet* **113:** 239-250.

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

Wang, S., P. Xu, J. Thorsen, B. Zhu, P. de Jong, G. Waldbieser, and Z. Liu. 2007. Characterization of a BAC library from channel catfish *Ictalurus punctatus*: indications of high rates of evolution among teleost genomes. *Marine Biotechnology* **in press**.

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

Zeng, C., N. Kouprina, B. Zhu, A. Cairo, M. Hoek, G. Cross, K. Osoegawa, V. Larionov, and P. de Jong. 2001. Large-insert BAC/YAC libraries for selective re-isolation of

genomic regions by homologous recombination in yeast. *Genomics* **77:** 27-34.

II. RESEARCH OBJECTIVES

Two BAC libraries of channel catfish had been constructed and the library CHORI 212 (http://bacpac.chori.org) was available for my dissertation research: BAC-based physical mapping and BAC end sequencing.

My dissertation was focused on the following objectives:

1. BAC end sequencing on CHORI 212 library:
    a) Generating over 20,000 BAC end sequences from both ends of fingerprinted BAC clones. Considering the success rate, around 25,000 sequencing reactions need to be conducted on 12,500 BAC clones;
    b) Repetitive element analysis using software Repeatmasker: assessing the repetitive element distribution and masking repeat sequences for BLAST analysis of these BAC end sequences;
    c) Isolation and evaluation of microsatellites in BAC end sequences: this is a very important resource for BAC-anchored microsatellite markers, which will be used in the physical map and linkage map integration;
    d) Mapping genes to BAC clones using BLASTX analysis;
    e) Comparative analysis with zebrafish genome and Tetraodon genome using BLAST analysis of BAC end sequences.
2. Construction of channel catfish BAC-based physical map using CHORI 212 library:
    a) Generating valid fingerprints from over 34,000 BAC clones (equivalent to 5X genome coverage) from the library CHORI 212 for physical map assembly. Around 42,000 BAC clones need to be isolated and fingerprinted using High Information Content Fingerprint (HICF) technique, assuming 80% success rate in the isolation and fingerprinting experiments;

b) Constructing the first version of channel catfish physical map using software Fingerprint Contig (FPC). Proper assembly parameters (tolerance value and Sulston score cutoff value) need to be determined for the assembly;
c) Validation of the newly assembled physical map. Overgo hybridizations will be conducted in the physical map contigs for the validation.

III. GENERATION OF CHANNEL CATFISH BAC END SEQUENCES FOR MARKER DEVELOPMENT AND ASSESSMENT OF SYNTENIC CONSERVATION WITH FISH MODEL SPECIES

ABSTRACT

BAC end sequences (BES) are a valuable resource in the genome research of any organism. Sequencing of BAC ends generates gene markers for comparative genome analysis, reveals microsattelites useful for map integration, allows an assessment of genomic architecture, and provides a framework for efficient genome sequencing. To develop such a resource in the aquaculture species channel catfish (*Ictalurus punctatus Rafinesque*), 25,195 BAC ends were sequenced generating 20,366 clean BES with an average read length of 557 bp after trimming. A total of 11,414,601 bp were obtained representing approximately 1.2% of the catfish genome. Based on this survey, the catfish genome was found to be highly AT-rich with 60.7% AT and 39.3% GC. Approximately 12% of the catfish genome consisted of dispersed repetitive elements with the Tc1/mariner transposons making up the largest percentage by base pair (4.57%). Microsatellites were detected in 17.5% of BES, providing valuable resources for map integration. Catfish BACs were anchored to the zebrafish and *Tetraodon* genome sequences by BLASTN search, revealing 16% and 8.2% significant hits ($E < e^{-5}$), respectively. A total of 1,074 and 773 significant hits were unique to the zebrafish and *Tetraodon* genomes, respectively, of which 417 and 406 were identified as known genes.

A total of 141 mate-paired BES were found to include genes on both sides of the BAC insert, from which 23 conserved syntenies were identified (~16.3%) among catfish, zebrafish, and *Tetraodon* genomes. Several of these syntenies were successfully extended by comparative analysis and sequencing within catfish BAC clones, indicating that comparative genomic mapping using genome resources of the fish model species will be highly effective for dissecting the genomes of aquaculture species.

INTRODUCTION

Teleost fish, accounting for over half of all vertebrate species, have in recent years been a testing ground for theories of genome duplication, divergence, and speciation (Crollius and Weissenbach 2005; Venkatesh 2003). The sequencing of the zebrafish *Danio rerio (Hamilton)* and the spotted green pufferfish *Tetraodon nigroviridis (Marion de Procé)* genomes, on top of the previously sequenced mammalian genomes, has provided a molecular platform for the analysis of vertebrate evolution. Genome-wide analyses of zebrafish and pufferfish, as well as medaka *Oryzias latipes (Temminck & Schlegel)* (Chiu et al. 2004; Naruse et al. 2004; Taylor et al. 2003; Woods et al. 2005), have lent support to the theory that two rounds of genome duplications occurred early in vertebrate history, followed by another whole genome duplication event in ray-finned fishes after the split from mammals and before the teleost radiation (Amores et al. 1998; Holland et al. 1994). The predictive value of this theory appears to be restricted, however, by differing rates of gene retention and/or duplication and divergence not only among teleost species, but also among the gene families within a given teleost species (Dugas and Ngai 2001; Gloriam et al. 2005; Kountikov et al. 2005; Peatman et al. 2006). The traditional model species alone may provide an insufficient context to resolve complex questions of duplication and divergence.

Genome enablement of aquaculture fish species holds unique promise in that it not only allows researchers to deepen understanding of genome evolution, but also to make genetic progress in economically-important food species (Crollius and Weissenbach 2005). For example, advances in salmonid genomics within the Genome Research of Atlantic Salmon Project (GRASP), including the construction of a physical map and

development of other genome resources, have allowed QTL analysis of important economic traits as well as providing a better understanding of tetraploid genomes (Artieri et al. 2006; Ng et al. 2005; Rise et al. 2004; Thorsen et al. 2005; Woram et al. 2003). Channel catfish, as the primary aquaculture species in the United States, has long been a research model for comparative immunology and toxicology, but has until recently lacked genomic resources. Development and evaluation of large numbers of molecular markers (Karsi et al. 2002; Liu et al. 2001; Serapion et al. 2004), construction of framework genetic linkage maps (Liu et al. 2003; Waldbieser et al. 2001), a successful EST project (Cao et al. 2001; Ju et al. 2002; Karsi et al. 2002; Kocabas et al. 2002), and the production of two BAC libraries using different restriction endonucleases (Quiniou et al. 2003; Wang et al. 2007) within the last five years have provided the foundation for large-scale genome research in catfish. The objective of this project, therefore, was to produce a resource of catfish BAC end sequences needed for physical mapping, comparative genome analysis, map integration, and better utilization of the existing genomic information.

Sequencing of BAC ends generates gene markers for comparative genome analysis, reveals microsatellite useful for map integration, allows an assessment of genomic architecture, and provides a framework for the development of a minimal tiling path for efficient genome sequencing (Chen et al. 2004; Fujiyama et al. 2002; Gregory et al. 2002; Larkin et al. 2003; Venter et al. 1996; Winter et al. 2004). While the long-term goals are to produce a high-quality BAC-based physical map and lay the groundwork for whole genome sequencing of channel catfish, BES from catfish could immediately be utilized for comparative genomics, gene mapping and microsatellite marker development, map

integration, and analysis of genomic architecture. In this project I sequenced 25,195 BAC ends and generated 20,366 clean BES. This project produced over 11.4 million base pairs of genomic sequences representing approximately 1.2% of the catfish genome and allowing, for the first time in catfish, an extensive survey of gene content, repeat status and arrangements, and syntenic conservation. A large number of genes and microsatellites were revealed in the BES, providing potential markers for integration of the physical map with linkage maps. Several conserved syntenies discovered in genes from paired BAC ends were successfully extended by comparative analysis and sequencing, indicating that genomic comparisons with the fish model species will be fundamentally important in future research in catfish.

RESULTS

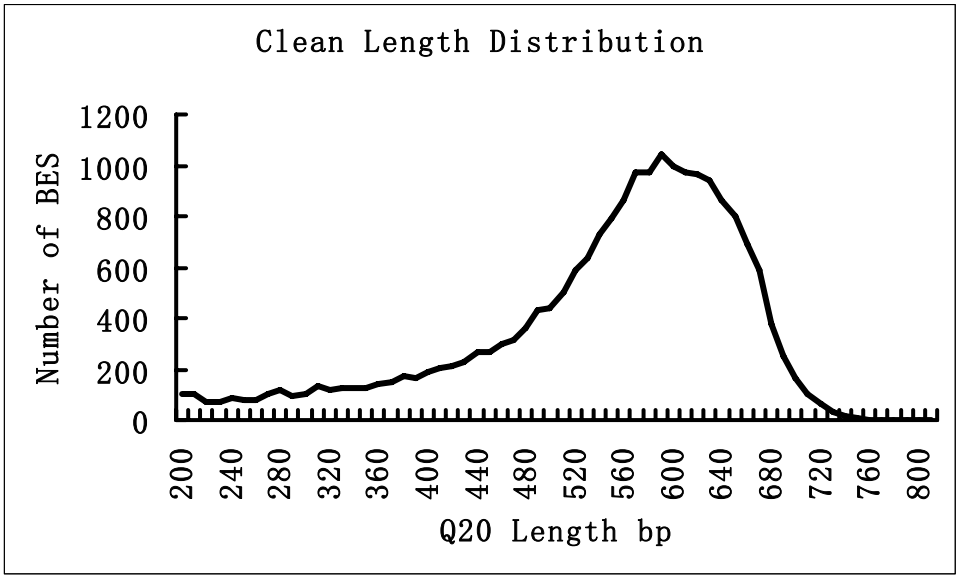
BAC end sequencing statistics

A total of 25,195 BAC ends were sequenced from 12,672 BAC clones (1.84X coverage of the channel catfish genome) from both ends, resulting in 20,493 high quality sequences (BES) >200 bp in length (81.3% overall success rate). Sequencing failures were largely due to empty wells and low BAC DNA yield. As detailed below, high numbers of simple sequence repeats appeared to be one of the additional major reasons reducing sequencing read length as 980 BES terminated after a round of microsatellite sequences. After filtering and trimming for *E. coli* and vector sequences, a total of 20,366 BES were generated. The BES were of high quality as the Q20 length ranged from 200 to 810 bp, with an average Q20 read length of 557 bp (Figure 1), comparing favorably to BES produced from human (Zhao et al. 2000), mouse (Zhao et al. 2001), and cattle

Table 1. Channel catfish BAC end sequencing statistics

Total BAC clones in CHORI 212 library	72,067 (10.60X genome coverage)
BAC clones sequenced	12,672 (1.84X genome coverage)
% BAC clone ends sequenced	17.58%
BAC end sequencing reads	25,195
BAC end sequences	20,493
BES after trimming and filtering	20,366
Average read length, bp	557 bp
Paired BES	17,478
Total bases sequenced	11,414,601
Percent of genome sequenced	1.27%
Repeat masked bases	1,524,504

Figure 1. Distribution of clean read length of the channel catfish BAC end sequences (BES). Quality assessment was performed using Phred software using Q≥20 as a cutoff as detailed in Methods.



(Larkin et al. 2003). All these BES have been deposited to GenBank GSS database with consecutive accession numbers of DX083364-DX103729. Of the 20,366 BES, 17,478 were mate-paired BES. The paired BES sequenced from both ends of the BACs should be a valuable resource for anchoring them to the zebrafish or *Tetraodon* genome sequence

for comparative analysis. A total of 11,414,601 bp genomic sequences were generated from this project, representing approximately 1.2% of the catfish genome (Table 1). The channel catfish genome appeared to be highly A/T rich with 60.7% A/T and 39.3 % G/C.

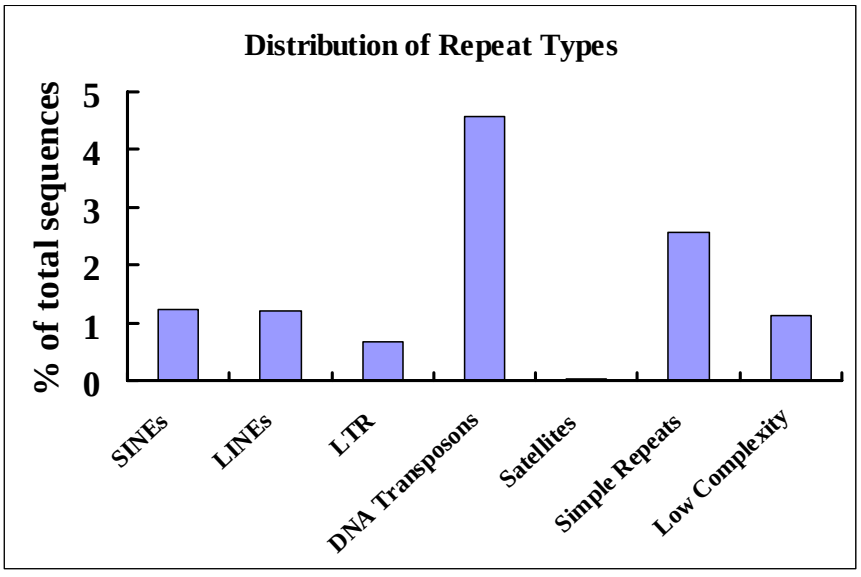
Table 2. Repeat composition of the channel catfish genome as assessed by RepeatMasker of the BAC end sequences using the combined repeat database of zebrafish and *Takifugu rubripes.*

Repetitive Elements	**Number of Elements**	**Length Occupied**	**Percentage of Sequence**
Retroelements	1972	349917 bp	3.13%
SINEs:	1083	139373 bp	1.24%
Penelope	1	269 bp	0%
LINEs:	643	135520 bp	1.21%
L2/CR1/Rex	525	107885 bp	0.96%
R1/LOA/Jockey	14	2773 bp	0.02%
R2/R4/NeSL	1	50 bp	0%
RTE/Bov-B	37	6137 bp	0.05%
L1/CIN4	65	18406 bp	0.16%
LTR elements:	246	75024 bp	0.67%
BEL/Pao	12	3604 bp	0.03%
Gypsy/DIRS1	179	60047 bp	0.54%
Retroviral	21	3726 bp	0.03%
DNA transposons:	2591	563923 bp	4.57%
hobo-Activator	220	23205 bp	0.16%
Tc1-IS630-Pogo	2077	507735 bp	4.12%
En-Spm	3	160 bp	0%
PiggyBac	57	6399 bp	0.06%
Tourist/Harbinger	77	7982 bp	0.07%
Unclassified:	9	751 bp	0.01%
Total interspersed repeats:		914591 bp	8.17%
Satellites:	25	2602 bp	0.02%
Simple repeats:	5798	289211 bp	2.58%
Low complexity:	3202	127064 bp	1.13%

Repeat status of the catfish genome

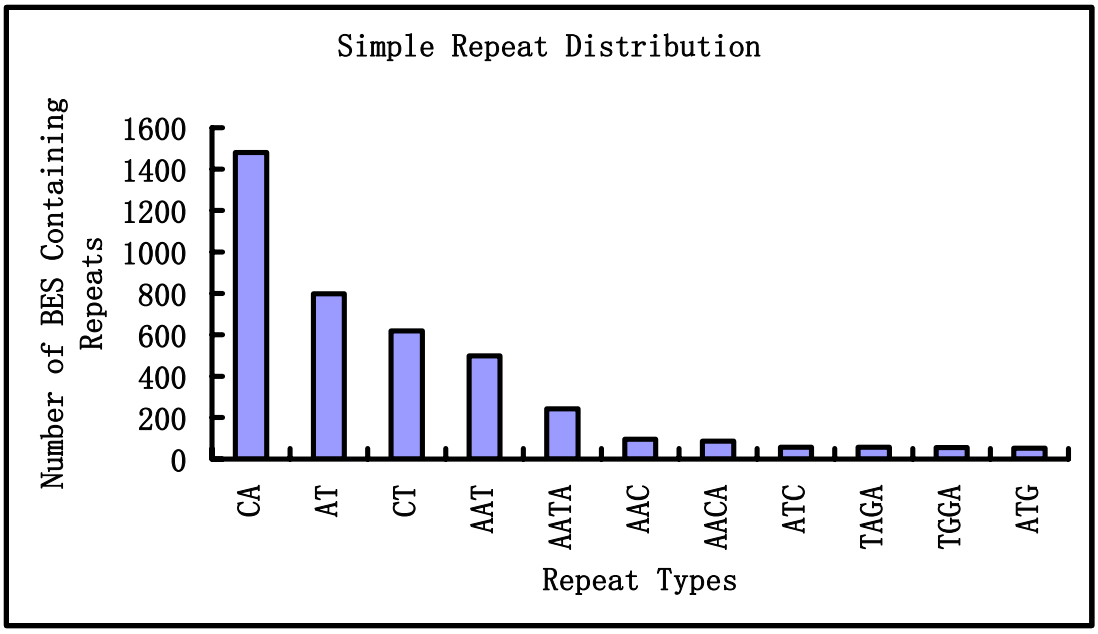
The overall status of repetitive elements in the catfish genome was assessed by repeatmasking. Analysis using the 11,414,601 bp BES resulted in 10.86% of base pairs masked using the *Danio* repeat database, and 7.31% of base pairs masked using the *Takifugu* repeat database. Use of both *Danio* and *Takifugu* repeat databases masked 11.91% (Table 2). These results suggested that while teleost fish share a high level of repetitive elements, a significant fraction of taxa-specific repeats exist. It appeared from the analysis that the zebrafish and catfish genomes shared a larger pool of repetitive elements than do the *Takifugu* and catfish genomes. It was also clear that the catfish and zebrafish genomes harbored a larger percentage of repetitive elements than the *Takifugu* genome, an expected result considering the compact size of the *Takifugu* genome.

Figure 2 Distribution of major repeat types as revealed by Repeatmasker analysis of the channel catfish BAC end sequences (BES). The percentage indicates the percentage of total base pairs.



The most abundant type of repeat encountered in the catfish genome was the DNA transposons, the vast majority consisting of Tc1/mariner transposon related sequences (Figure 2 and Table 2). The Tc1/mariner transposons accounted for 4.12% of the base pairs of the BES. Retroelements were the second largest fraction of repetitive elements accounting for 3.13% of the base pairs of the BES. Simple repeats such as microsatellites accounted for 2.58% of the catfish BES base pairs with CA, AT, and CT type repeats accounting for 68% of all microsatellite types (Figure 3).

Figure 3 Distribution of major microsatellite types identified from the channel catfish BAC end sequences (BES).



A total of 3,748 BES were found to contain one or more stretches of microsatellite sequences. Of these, 2,365 (63%) had sufficient flanking sequences on both sides, making them potentially useful as markers for genetic mapping; 403 BES harbor microsatellite sequences at the immediate beginning of the BES, making them more

difficult to develop as markers; the remaining 980 clones had microsatellite sequences at the end of BES (Table 3). For this last category, additional sequencing can be conducted using primers close to the end of the BES to generate sufficient flanking sequences on both sides. These microsatellites should be useful for genetic linkage mapping and for integration of the catfish linkage maps with the BAC-based physical map. In order to assess the proportion of microsatellites useful for linkage mapping in the resource families, the polymorphism analysis on 80 microsatellite loci was conducted. Approximately 55% of the tested microsatellites exhibited polymorphism in one of the resource families in the lab, F_1-2 (female) X Ch-6 (male). This result indicated that the majority of the identified microsatellites can be developed into polymorphic markers for genetic linkage mapping in one or more of the catfish resource families (Liu et al. 2003). Based on the conservative 55% polymorphic rate by testing in only one resource family, approximately 1,300 microsatellite markers can be developed from this BES resource without additional sequencing.

Table 3. Microsatellite contents of the catfish BAC end sequences and their characteristics.

Total number of microsatellites contained in the BES	4,262
Number of di-nucleotide microsatellite loci	2896 (68%)
Number of tri-nucleotide microsatellite loci	789 (18.5%)
Number of tetra-nucleotide microsatellite loci	577 (13.5%)
Total number of BES containing microsatellite	3,748 (18.4% of total BES)
Number of microsatellite loci with sufficient flanking sequences for primer design	2,365 (63%)
Number of microsatellites located at the beginning of BES	403 (11%)
Number of microsatellites located at the end of BES	980 (26%)

Analysis of the catfish BES against the human repeat database suggested the presence of a significant fraction of similar repetitive elements between teleost fish and mammals. In addition to the expected simple repeats (2.7%) and low complexity repeats (1.1%), the Repeatmasker masked 1.08% of the channel catfish BES in the category of DNA repetitive elements, of which the vast majority (1.06%) were the MER2 type of repeats.

Identification of novel repeats in the catfish genome

In order to identify novel repetitive elements in the catfish genome, the catfish BES were analyzed by sequence comparison against themselves. A catfish BES database was established and BLASTN was used to search against the catfish BES. After repeatmasking, self-BLAST searches were conducted. The BLAST searches resulted in the identification of two major groups of sequences with large numbers of hits (>50). The first group of repeats included 87 BES that had significant hits on one another while the second group of repeats included 76 BES that had significant hits on one another. Sequence analysis using BLASTX indicated that the first repeat was most similar to rRNA intron-encoded homing endonuclease from chimpanzee (accession number XP_525925), suggesting a long evolutionary history for this class of repeats in vertebrates. The second repeat was most similar to an unknown gene sequence from *Schistosoma japonicum* (*Katsurada*) (accession number AAX30301), and to the cytochrome P450 genes identified from plants (e.g. a cytochrome-like gene from tobacco, accession number BAA10929).

Table 4. Mapping of genes to BACs through BAC end sequencing as assessed by BLASTX searches. Listed are number of BLASTX hits of genes by BAC end sequences, excluding redundant hits. P-values, alignment length range, average alignment length, and percentage of identities are provided as an indication of the level of similarities.

p-value	Number of hits	Alignment length (amino acid)	Average alignment length (amino acids)	% Identity
$<10^{-50}$	58	101-228	167	48-99
10^{-40}-10^{-50}	54	81-207	134	43-97
10^{-30}-10^{-40}	77	66-217	103	40-100
10^{-20}-10^{-30}	253	45-175	75	34-100
10^{-15}-10^{-20}	275	37-199	62	30-100
10^{-10}-10^{-15}	413	30-186	54	31-100
Subtotal	1130	30-228	73	31-100
10^{-5}-10^{-10}	747	19-193	47	23-100
Total	1,877	19-228	63	23-100

Mapping genes to BACs

BLASTX searches of the 20,366 BES resulted in 2,351 BES with significant hits to 1,877 unique genes. As establishing orthologies is difficult by simple BLASTX searches, I attempted to characterize the top hit genes by their p-values as well as alignment length and percentage of identity to better assess hit quality. Of the 1,877 gene hits, 1,130 had a p-value smaller than 1×10^{-10} in BLASTX searches with average alignment length of 73 amino acids and a range of 31-100% identity (Table 4). While it is difficult to conclude what level of p-values would provide a stringent confidence on the putative gene identities of the BES, it is obvious that the lower the p-values, the more likely the BES are to be related to the putatively identified genes. An additional 747 gene hits were

identified with a p-value between 10^{-5} to 10^{-10}, but these hits in many cases involved either short, high quality alignments or long, highly divergent alignments (Table 4). If a p-value of 10^{-10} is sufficiently low to determine the putative identities of the genes contained within the BES, this project should have effectively mapped 1,130 genes to BAC clones, an important resource for comparative genome analysis.

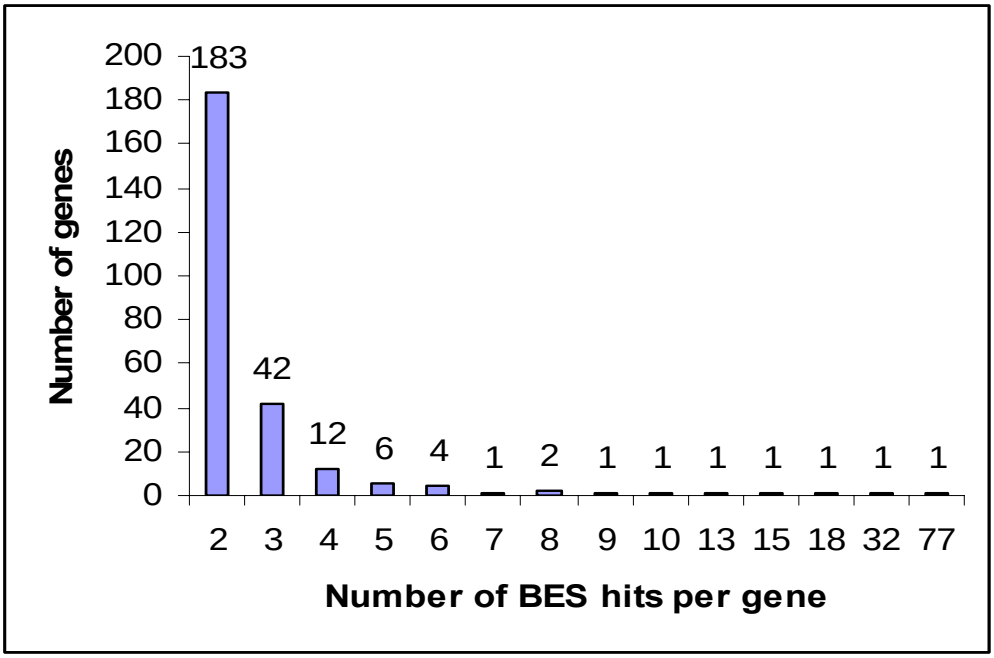
In addition to unique BLASTX hits, multiple BES (2-77) hit a single gene were also found (Figure 4). At the lower range, this result can be explained by redundant sequencing of the same gene harbored in different BAC clones in the 10 X coverage libraries or by gene duplication. However, for some gene hits, a large number of BES was involved. For instance, six genes were hit by 10 or more BES, indicating the presence of large families or multiple copies of these genes in the catfish genome. BLASTX searches indicated their putative identities were neurotactin (10 hits), CCHC type and RNA-directed DNA polymerase (15 hits), gephyrin (18 hits), senescence-associated protein (45 hits), and rRNA intron-encoded homing endonuclease (77 hits). Searches of the zebrafish and *Tetraodon* genomes revealed that all these genes existed in large copy numbers, suggesting conservation of large gene copy numbers or gene families in catfish as well.

Anchoring of channel catfish BES to zebrafish and *Tetraodon* genome sequences

BLASTN searches of the catfish BES against zebrafish and *Tetraodon* genome sequences resulted in 3,251 (16%) and 1,670 (8.2%) significant hits ($p<10^{-5}$), respectively. However, many BES had hits in different genomic regions of the zebrafish genome sequence, suggesting that they are repetitive in nature. In order to obtain unique

significant hits that are meaningful as resources for comparative genome analysis, the significant hits were tabulated in Excel with hit ID, chromosome information, and beginning and end of the region of similarity. Repeated hits were then removed, resulting in 1,074 unique significant hits to the zebrafish genome sequence. Similar BLAST searches against the *Tetraodon* genome resulted in 773 unique significant hits suggesting that the number of genomic regions containing evolutionarily conserved sequences was greater between catfish and zebrafish than between catfish and *Tetraodon*, consistent with their phylogenetic relationships.

Figure 4 Distribution of multiple BLASTX hits by using channel catfish BAC end sequences (BES) as queries as an indication of potential gene duplications and multigene families. For instance, 42 genes were hit by three BES each, and one gene was hit by 77 BES.



In order to understand the nature of the conserved genomic sequence blocks between catfish and zebrafish or *Tetraodon*, the BES with unique BLASTN significant hits were searched against the NR database using BLASTX to assess the number of genes present among the conserved genomic sequences. Of the 1,074 unique significant hits to the zebrafish genome, 417 (38.8%) had significant BLASTX hits. Similarly, with *Tetraodon*, of the 773 unique significant hits, 406 had significant BLASTX hits. The similarity in number of significant BLASTX hits with the zebrafish and the *Tetraodon* genomes suggested that the majority of genes were conserved among catfish, zebrafish, and *Tetraodon*. The greater number of unique BLASTN hits to the zebrafish genome was accounted for, in the most part, by non-gene genomic sequences reflecting the level of conservation between the catfish and zebrafish genomes.

To develop a set of gene markers for potential comparative genome analysis, the catfish BES were searched against the chromosomes of the zebrafish and *Tetraodon* genomes. BLASTN and BLASTX search results were tabulated according to chromosome locations (Table 5). The catfish BES had significant BLASTN hits to each of the 25 zebrafish chromosomes with a range of 16-66 hits per chromosome. BLASTX searches against the NR database using catfish BES with unique BLASTN hits to the zebrafish genome resulted in 4-30 significant hits per chromosome. Similar but fewer unique hits were found with the *Tetraodon* genome. Notably, only a single significant hit with BLASTN or BLASTX was found with *Tetraodon* chromosome 20, likely due to the small size of the chromosome and its low sequence coverage to date (http://www.genoscope.cns.fr/externe/tetranew/).

Table 5. Distribution of unique BES significant BLASTN hits in the *Danio rerio* and *Tetraodon nigroviridis* genomes. The cut-off value was set at $p=10^{-5}$

Chromosome	Zebrafish		*Tetraodon*	
	Unique BLASTN hits	BLASTX hits of catfish BES with unique BLASTN hits	Unique BLASTN hits	BLASTX hits of catfish BES with unique BLASTN hits
1	58	22	44	32
2	39	13	55	27
3	58	28	36	19
4	36	14	13	10
5	66	27	23	13
6	40	21	15	4
7	56	22	20	11
8	36	15	20	14
9	59	24	23	12
10	42	16	32	21
11	38	15	21	11
12	37	16	24	16
13	45	16	74	17
14	54	16	14	9
15	31	9	30	25
16	53	18	23	15
17	47	15	16	9
18	28	4	21	14
19	44	20	10	8
20	52	30	1	1
21	34	10	19	14
22	28	11		
23	54	20		
24	23	7		
25	16	8		
Undesignated			239	104
Total	1074	417	773	406

Conserved syntenies between catfish, zebrafish, and *Tetraodon*

Particular attention was given to evaluating the level of conservation between the catfish and zebrafish or between the catfish and *Tetraodon* genomes. Of the 20,366 BES,

17,478 BES were mate-paired sequences from 8,739 BAC clones. BLASTX searches indicated that 141 sequenced BACs harbor genes on both ends, allowing us to compare whether the same set of genes were located in similar environs in the zebrafish and *Tetraodon* genomes. Of the 141 paired BAC ends with genes, 43 (30.5%) were located on the same chromosomes of zebrafish or *Tetraodon*, of which 23 (16.3%) appeared to exhibit a high level of conserved synteny (Table 6). The number of conserved syntenies was greater between the catfish genome and the zebrafish genome than between the catfish genome and the *Tetraodon* genome. Of the 23 conserved syntenies, 21 were present between the catfish and zebrafish genomes. Comparison of syntenic conservation was more difficult with the *Tetraodon* genome because of its incomplete assembly. Of the 23 conserved syntenies among catfish, zebrafish, and *Tetraodon*, 14 were present in *Tetraodon*; four were absent in *Tetraodon*; and the status of five could not be determined because one or both of the two genes involved in the paired BAC ends were not yet assigned to specific chromosomes in *Tetraodon*. Although the exact distances between the paired genes sequenced from catfish BAC ends were not determined, the average insert size of the catfish CHORI 212 BAC library is 161 kb (http://bacpac.chori.org/library.php?id=103). In most cases, the distances between the sets of two genes were larger in zebrafish than in *Tetraodon* (Table 6), consistent with the compact genome of *Tetraodon*.

Table 6. A summary of conserved syntenies identified by comparison of 141 mate-paired genes of channel catfish with genomic locations of those within the zebrafish and *Tetraodon* genomes**.** The putative identities of the mate-paired genes are provided as

GenBank accession numbers of their top BLASTX hits. Sp6 or T7 hits indicate the gene identities of the BES using the Sp6 or T7 sequencing primer. Chr indicates chromosome on which the genes are located, and distance indicates the distance found between the two genes in zebrafish or *Tetraodon* as appropriate. Shaded rows are syntenies conserved among all three species of catfish, zebrafish, and *Tetraodon*. For "cadherin cluster" and "pheromone receptor cluster", the distance could not be determined because these genes are arranged in tandem as gene clusters.

Catfish BAC	**Sp6 hits**	**T7 hits**	**Zebrafish**		***Tetraodon***	
			Chr	**Distance**	**Chr**	**Distance**
001_L07	CAG01022.1	CAG01025.1			2	130 kb
003_J07	XP_691151.1	Q96Q40	6	2400 kb	3	1087 kb
007_H08	XP_690677.1	AAH90180.1	12	430 kb		
007_K16	XP_544816.1	CAG08989.1	10	30 kb	16	100 kb
008_I11	NP_001019337.1	NP_001007763.1	10	Cadherin cluster	7	Cadherin cluster
013_P16	XP_691920.1	XP_690925.1	18	545 kb	13	74 kb
014_D16	XP_428910.1	XP_698664.1	2	510 kb	Un/19	
018_H11	AAH56818.1	AAH44562.1	13	453 kb	Un/17	
018_I19	CAF99682.1	CAF99686.1	7	508 kb	5	53 kb
020_H13	AAH85663.1	CAF92624.1	13	424 kb	3/Un	
020_I23	BAD90503.1	CAF96508.1	9	8 mb	11	91 kb
020_L17	XP_690830.1	AAX46593.1	5	190 kb	Un/12	
021_L13	CAF95098.1	XP_688911.1	19	602 kb	Un/Un	
022_D09	XP_685117.1	P21359	15	376 kb	16/Un	
022_G21	NP_072140.1	XP_535054.2	17	562 kb		
023_O10	XP_693773.1	XP_685853.1	24	580 kb	6/Un	
025_I21	XP_706772.1	XP_684635.1	11	260 kb		
028_J11	AAW38963.1	AAC64076.1	17	Pheromone receptor cluster	16	Pheromone receptor cluster
028_M04	XP_687685.1	XP_695804.1	5	122 kb	16	206 kb
029_M13	CAG04452.1	CAG04458.1	12	1.9 mb	2	158 kb
031_O02	AAL66362.1	XP_696942.1			6	309 kb
032_F20	XP_691291.1	AAH78367.1	7	270 kb		
033_A11	XP_693134.1	AAH97450.1	17	600 kb	10	57 kb

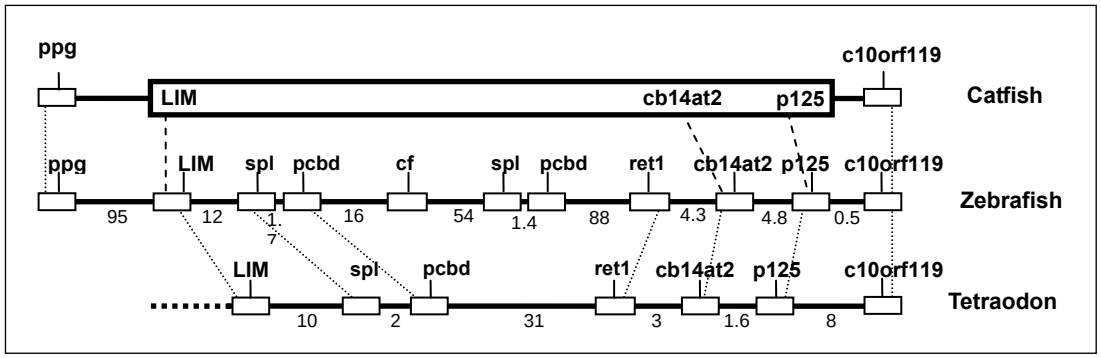
Validation and extension of conserved syntenies

In order to evaluate the extent of the conserved syntenies, I attempted to extend the observed syntenies by additional experiments. Three BACs with paired genes on their ends, 018_H11, 022_D09, and 028_M04, were further evaluated by direct BAC sequencing. As shown in Figure 5, the zebrafish and *Tetraodon* genomes were searched in order to determine what genes were located between the genes syntenic with the paired catfish genes. The catfish EST database was searched for the presence of sequences orthologous to these intermediate genes. When EST sequences were available, sequencing primers were designed and used to directly sequence the relevant catfish BAC clone. The generated sequences were then analyzed by BLASTX searches or by sequence alignment with the ESTs. As shown in Figure 5, the syntenies were well conserved, though the order of the genes in catfish was not determined. For BAC 018_H11, the genes on the BAC ends were poly A polymerase gamma and the chromosome 10 ORF119 gene. Three sequencing primers designed for *LIM*, *cb14at2* (chondroitin beta1,4 N-acetylgalactosaminyltransferase 2), and *p125* (SEC23 interacting protein p125) all generated correct sequences by direct BAC sequencing, confirming the presence of these genes within the BAC clone. Similarly, four genes [RAB1, wsb1 (SOCS box-containing WD protein SWiP-1), Galectin 9, and neurofibromatosis type 1] were confirmed to be present within the BAC clone 022_D09; and four genes [archain 1, *h2afx* (H2A histone family member X, *mizf* (methyl-CpG-binding protein-interacting zinc finger protein), and *abcg* (ATP-binding cassette, sub-family G, member 4)] were confirmed to be present in the BAC clone 028_M04 by direct BAC sequencing. A brief examination of the conserved syntenies also suggested a higher level of genome conservation between the

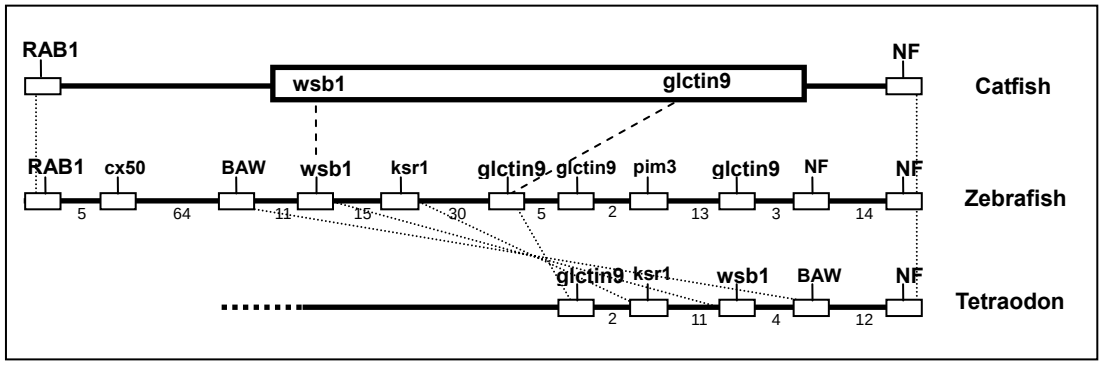
catfish and zebrafish genomes than between the catfish and the *Tetraodon* genomes. In the second and third conserved syntenies as shown in Figure 5, many more gene rearrangements were observed in the *Tetraodon* genome as compared to the zebrafish and the catfish genomes. Because of the lack of the EST sequence information for the design of sequence or PCR primers, many genes within the conserved syntenies could not be easily confirmed in catfish, but the demonstrated extension of conserved syntenies suggests a high level of genome conservation.

Figure 5. Examples of conserved syntenies extended from the mate-paired genes from both ends of the channel catfish BAC end sequences (BES) by comparative analysis of the genes in the catfish genome with those from the zebrafish and *Tetraodon* genomes. Exact gene order, distance, and orientation of the catfish genes internal to the mate-paired genes were not determined. Three syntenies are shown from BAC 018_H11 (A), 022_D09 (B), and 028_M04 (C).

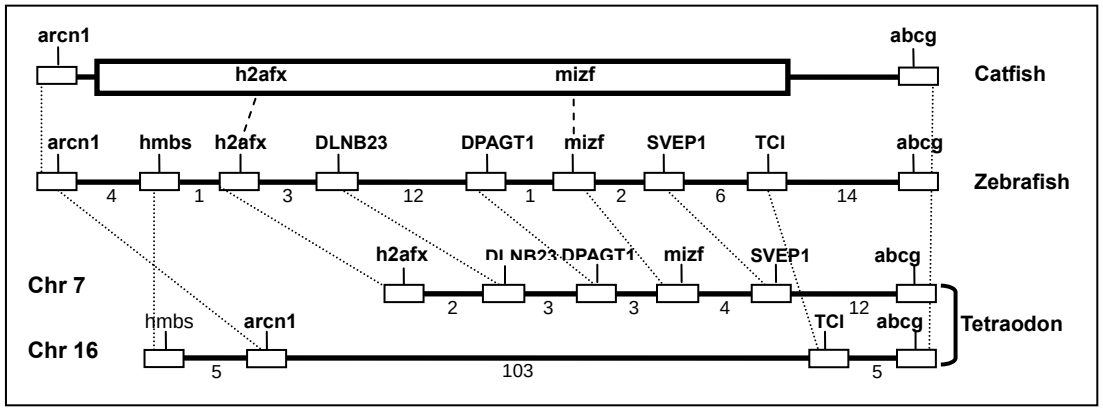
A.



B.



C.



DISCUSSION

Catfish (*Ictalurus sp.*) aquaculture, a billion-dollar industry in the United States, represents over 60% of all US aquaculture production. Catfish has long served as a model for comparative immunology, reproductive physiology, and toxicology among ectothermic vertebrates because of their unique characteristics and natural occurrence in the majority of freshwater ponds, lakes, streams, and rivers of the US (Bao et al. 2006a; Kazeto and Trant 2005; Shen et al. 2004; Tilton et al. 2002). Closely related catfish species of the order *Siluriformes* are cultured and studied throughout the world. Large-scale genome research in catfish, therefore, holds the potential to place decades of important, applied molecular studies into a functional and evolutionary context through comparative genomics. In particular, BAC end sequencing has proved to be an especially efficient approach for the generation of genome resources. While uncovering gene and microsattelite loci, the BAC end sequences themselves provide information needed to merge contigs for physical mapping. A physical mapping project has been recently initiated in channel catfish. Additionally, BAC-anchored markers are crucial for future integration of the physical map with existing catfish linkage maps. The production of 20,366 high quality BAC end sequences from catfish, representing 1.2% of the genome should allow advances in physical mapping, comparative genome analysis, map integration, and better utilization of the existing genomic information. In the long term, the BES will be also important for the identification of a minimal tiling path in preparation for genome sequencing.

Catfish and zebrafish, as Ostariophysian fishes, share much closer evolutionary relationships than does zebrafish with other fish model species, medaka, *Tetraodon*, and

Takifugu. One would predict, therefore, that the catfish and zebrafish genomes would show high levels of conservation. If catfish demonstrates high syntenic conservation with zebrafish, the future tasks of mapping and assembly of the catfish genome would be simplified. Additionally, both catfish and zebrafish researchers could benefit from an exchange of information on gene duplication and divergence. In particular, a major catfish transcriptome project is ongoing with the Joint Genome Institute (JGI) for the sequencing of 600,000 ESTs. Once the physical framework is established, along with the anticipated identification of the vast majority of catfish genes through the JGI project, catfish genome information could be a substantial comparative resource for the studies of teleost genomes. A comparison of gene families from zebrafish and catfish could better reveal the timing of duplication events, as well as phenomena such as chromosomal rearrangements, alternative splicing, and pseudogenes, than could a comparison of either species with the distant and compact genomes of *Tetraodon* and *Takifugu*. Despite these advantages, the only points of genomic comparison between the two species prior to BAC end sequencing have been small gene-based sequencing reports (Peatman et al. 2006; Wang et al. 2006). The catfish BES provided a resource for anchoring catfish genomic sequences to the zebrafish and *Tetraodon* genomes. Over 3200 significant BLASTN hits were generated with the zebrafish genome sequence. Concentrating on unique hits alone, 1,074 hits were generated with the zebrafish genome. Only 417 of the hits appeared to be genes as revealed by BLASTX searches. The majority of these significant hits were to non-gene regions, indicating strong genome similarities between catfish and zebrafish.

The sequencing of 141 BAC clones containing genes on both ends allowed additional evaluation of syntenic conservation with zebrafish as well as *Tetraodon*. Over 16% of the paired genes identified from the mate-paired BES were located proximally to one another in the zebrafish and/or *Tetraodon* genomes. Additional experiments utilizing comparative analysis and direct BAC sequencing in catfish revealed that many of the syntenies could be extended, with catfish EST availability being the limiting factor. These encouraging results suggest that comparative mapping, especially with zebrafish, will be a sound approach for future catfish genomics research. Furthermore, information gained in mapping and gene discovery projects in catfish may help to explain aspects of zebrafish and teleost genome evolution.

The catfish genome harbors a significant amount of dispersed repeats as revealed by BAC end sequencing. Of the various repeat types, DNA transposons are the most abundant. The vast majority of the DNA transposons are Tc1/mariner-related sequences present as transposon remnants. Previous research using hybridization suggested that the catfish genome was rich in Tc1-like transposons, but the present study revealed that catfish contains a larger than previously thought proportion of Tc1-like transposons, accounting for 4.1% of the catfish genome. Various types of retroelements account for 3.1% of the catfish genome. Most importantly for mapping, the catfish genome is rich in simple sequence repeats including microsatellites. A total of 4,262 microsatellites were sequenced in 3,748 BES (18.4% of BES). At least 2,365 of these microsatellites are ready for marker development, with sufficient flanking sequences for primer design. These microsatellites will be an important resource for integrating the catfish physical map with existing linkage maps, as well as for comparative mapping (Barbosa et al. 2004; Kiuchi et

al. 2002).

The catfish genome contains a significantly lower proportion of repetitive elements when compared to the genomes of mammals. The fraction of repeat bases in the set of catfish BES was approximately 12% using Repeatmasker with the zebrafish and *Takifugu* repeat databases. Even using the mammalian repeat database, overall masked repeats were still below 15%. Clearly, this fraction was underestimated, most likely because of the low ability of BAC end sequencing to detect tandem repeats. Previously published paper had described the presence of a major class of tandem repeats named *Xba* elements (Liu et al. 1998) that accounted for about 5% of the catfish genome. These elements were not detected in the BES, because they lack the *Eco*R I restriction sites necessary for insertion into BAC clones. Nevertheless, fish genomes may harbor a significantly lower fraction of repetitive elements than mammals. Repetitive element percentages of 47% and 37% were found in cattle BES (Larkin et al. 2003) and mouse BES (Zhao et al. 2001), respectively. This result appears consistent with the overall smaller genome size of teleost fish in comparison to mammals. The smaller number of dispersed repeats in the catfish genome should alleviate some complications involved in the assembly of a whole-genome shotgun sequence and make comparative genome analysis more attractive.

Integration of physical and linkage maps can be approached from several directions. Markers from a linkage map can be hybridized to BAC libraries used in physical mapping, or markers discovered by sequencing BACs, if polymorphic, can be placed onto linkage maps. Assigning genes to both maps greatly increases their informative value as well as enhances map resolution. Based on past experience, BAC end sequencing appears to be a highly efficient method for marker generation, gene

localization, and eventual map integration. Previous research have been conducted using overgo hybridizations to screen the catfish BAC library for gene localization (Bao et al. 2005; Bao et al. 2006b; Baoprasertkul et al. 2005; Peatman et al. 2006; Wang et al. 2006; Zhao et al. 2001). Despite efforts to increase efficiency by using a two-dimensional design, only several dozen genes have been mapped to BAC clones to date. In contrast, at least 1,130 genes were identified in the catfish BES if a stringent cutoff value of $p=10^{-10}$ was used. I also expect that mapping microsatellite uncovered in the BES to linkage maps will prove to be significantly more efficient than hybridizing microsatellite markers from the linkage maps to BACs. Attempts to use this latter approach in Atlantic salmon proved to be too labor intensive and were complicated by an abundance of repetitive elements in the genome (William Davidson, Simon Fraser University, personal communication). A large resource of BAC end sequences should provide a solid foundation for future physical mapping, map integration, and comparative genomics in catfish.

METHODS

BAC culture and end sequencing

The CHORI-212 Channel Catfish BAC library was used for BAC end sequencing. CHORI-212 BAC library was created in Dr. Pieter de Jong's laboratory by cloning the *Eco*RI/*Eco*RI methylase partially digested high-molecular-weight DNA prepared from a male channel catfish (USDA103 strain) into the pTARBAC2.1 vector between the *Eco*RI sites and transformed into DH10B (T1-resistant) electro-competent cells (Invitrogen, Carlsbad, CA) (http://bacpac.chori.org/library.php?id=103). BAC clones were inoculated into 2.2-ml 96-well culturing blocks containing 1.5 ml 2×YT medium and 12.5 µg/ml

chloramphenicol from 384-well stocking plates using 96-pin replicator (V&P Scientific, Inc., San Diego, CA). Blocks were covered with an air permeable seal (Excel Scientific, Wrightwood, CA) and incubated at 37° C for 24 hours with shaking at 300 rpm. The blocks were centrifuged at 2000g for 10 min in an Eppendorf 5804R bench top centrifuge to precipitate the 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 Perfectprep® BAC kit (Brinkmann Instruments, Westbury, NY) according to the manufacturer's instructions. BAC DNA was collected in 96 plates and stored in -20°C before use.

Dye terminator sequencing reactions were conducted in 96-well semi-skirt plates using the following ingredients: 2 µl 5X Sequencing Buffer (Applied Biosystems, Foster City, CA), 2 µl sequencing primer (3 pmol/µl), 1 µl BigDye v3.1 Dye Terminator, and 5 µl BAC DNA. The cycling reactions were conducted with MJ Research Thermal Cyclers under the following conditions: initial 95°C for 5 min; then 100 cycles of 95°C for 30 sec, 53°C for 10 sec, 60°C for 4 min followed by incubation at 4°C before clean up. The standard T7 and SP6 primers were used for sequencing reactions (T7 primer: TAATACGACTCACTATAGGG; SP6 primer: ATTTAGGTGACACTATAG). After sequencing reactions were completed, 1 µl of 125 mM EDTA and 25 µl pre-chilled 100% ethanol were added to each well. After mixing and incubating at room temperature for 10 min, the plate was spun on 2250g at 4°C for 40 min followed by washing in 50 µl of 70% ethanol at 1650g for 15 min. Hi-Di formamide (10 µl) was added to each well to resuspend DNA. The DNA was denatured at 95°C, and then the samples were analyzed with an ABI 3130XL automated capillary sequencer (Applied Biosystems).

Clone tracking and quality assessment

In order to assure the clone tracking of the BES, eight clones were re-sequenced from each 384-plate from positions A1, A2, B1, B2, C1, C2, D1, and D2. Of a total of 33 384-well plates sequenced, mistakes were found on one plate, CHORI212_004. The sequences of this plate were subjected to a greater level of analysis. Sequencing and BLAST analysis indicated that 288 clones of this plate had correct sequences, while the remaining 96 clones were a direct repeat of the first 96 clones of the 288 clones, suggesting that a mistake was made in the orientation of one 96-well culture plate leading to the repeated growth of the same clones.

Quality assessment was performed using the raw chromatogram files directly before any trimming. The raw, untrimmed files were processed by Phred software (Ewing and Green 1998; Ewing et al. 1998) within the Genome Project Management System (GPMS). Phred quality score cut off value was set at 20 for the acquisition of Q20 values.

Sequence processing and bioinformatics

The BAC end sequences were trimmed of vector sequences and filtered of bacterial sequences, stored in a local Oracle database after base calling and quality assessment using Genome Project Management System (GPMS) (Liu, et al, 2000), a local laboratory information management system for large-scale DNA sequencing projects. Quality assessment was performed using Phred software using Q≥20 as a cutoff. Repeats were masked using Repeatmasker software (http://www.repeatmasker.org) before BLAST analysis.

BLASTX searches of the repeat masked BES were conducted against the Non-Redundant Protein database. A cut off value of 10^{-5} was used as the significance

similarity threshold for the comparison. The BLASTX result was parsed out in a tab-delimited format. In order to anchor the catfish BES to zebrafish and *Tetraodon* genomes, BLASTN searches of the repeat masked catfish BES were conducted against zebrafish and *Tetraodon* genome sequences. The location and chromosome number of each top hit was collected from the results and parsed out in tab-delimited format.

For the identification of novel repeats in the catfish genome, all 20,366 BES of catfish were collected and formatted to a nucleotide database. Self-BLASTN searches were conducted. The cut off p-value was set at 10^{-2} and the identify threshold was set to 90% with a minimal alignment length of 100 bp. The BLASTN results were parsed to tab-delimited format to count the redundant queries and to gather other statistics.

Identification of microsatellites and determination of their polymorphism

Microsatellites and other simple sequence repeats were analyzed by using Repeatmasker as well as by using Vector NTI Vector NTI Suite 9.0 (Invitrogen) as we previously described (Serapion et al. 2004). PCR primers were designed using the FastPCR software package (http://www.biocenter.helsinki.fi/bi/Programs/fastpcr.htm). In each case, primers were designed at the highest possible stringency. Factors considered included PCR product length, duplex formation, hairpin formation, false priming sites, and the melting temperature. For analysis of small differences of the SSR polymorphism, PCR primers were designed to generate products of about 100 bp, but in some cases primers were designed to generate larger PCR products because of highly repetitive sequences flanking microsatellites. Polymorphism within a resource family, F_1-2 (female) X Ch-6 (male), was determined by PCR analysis. PCR products were run on a LICOR 4200 automated sequencer.

Identification and validation of conserved syntenies

Initially, mate-paired BES were analyzed by BLASTX searches to identify mate pairs with genes on both sides of the BAC insert. After their identification, the two mate-paired genes in each BES were used as queries to search their chromosomal locations on the zebrafish and *Tetraodon nigroviridis* genomes. The distance between the two genes was limited to 1.2 Mb unless a smaller distance was found in either the zebrafish genome or the *Tetraodon* genome, in which case we also accepted distances smaller than 2.5 Mb. Once the initial conserved syntenies were identified, further validation for the presence of additional genes found between the two conserved genes of zebrafish or *Tetraodon* within the catfish BACs were determined by direct BAC sequencing. First, the zebrafish and/or *Tetraodon* genes present between the two conserved gene pairs were used to search the catfish EST database to determine if such genes had been identified in catfish. Sequencing primers were then designed based on the catfish EST sequence and used for direct BAC sequencing. The generated sequences were aligned to the catfish EST sequences or subjected to BLASTX searches to determine the putative gene identities of the sequences. The distances and orientations of the conserved genes in catfish were not determined.

ACKNOWLEDGEMENTS

This project was supported by a grant from USDA NRI Animal Genome Tools and Resources Program (award # 2006-35616-16685). I thank Dr. Pieter de Jong at the Children's Hospital of the Oakland Research Institute for providing us with the CHORI 212 channel catfish BAC library.

REFERENCES

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

Artieri, C.G., L.A. Mitchell, S.H. Ng, S.E. Parisotto, R.G. Danzmann, B. Hoyheim, R.B. Phillips, M. Morasch, B.F. Koop, and W.S. Davidson. 2006. Identification of the sex-determining locus of Atlantic salmon (*Salmo salar*) on chromosome 2. *Cytogenet Genome Res* 112: 152-159.

Bao, B., E. Peatman, P. Li, C. He, and Z. Liu. 2005. Catfish hepcidin gene is expressed in a wide range of tissues and exhibits tissue-specific upregulation after bacterial infection. *Dev Comp Immunol* **29:** 939-950.

Bao, B., E. Peatman, X. Peng, P. Baoprasertkul, G. Wang, and Z. Liu. 2006a. Characterization of 23 CC chemokine genes and analysis of their expression in channel catfish (Ictalurus punctatus). *Dev Comp Immunol* **30:** 783-796.

Bao, B., E. Peatman, P. Xu, P. Li, H. Zeng, C. He, and Z. Liu. 2006b. The catfish liver-expressed antimicrobial peptide 2 (LEAP-2) gene is expressed in a wide range of tissues and developmentally regulated. *Mol Immunol* **43:** 367-377.

Baoprasertkul, P., C. He, E. Peatman, S. Zhang, P. Li, and Z. Liu. 2005. Constitutive expression of three novel catfish CXC chemokines: homeostatic chemokines in teleost fish. *Mol Immunol* **42:** 1355-1366.

Barbosa, A., O. Demeure, C. Urien, D. Milan, P. Chardon, and C. Renard. 2004. A physical map of large segments of pig chromosome 7q11-q14: comparative analysis with human chromosome 6p21. *Mamm Genome* 15: 982-995.

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

Chen, R., E. Sodergren, G.M. Weinstock, and R.A. Gibbs. 2004. Dynamic building of a BAC clone tiling path for the Rat Genome Sequencing Project. *Genome Res* **14:** 679-684.

Chiu, C.H., K. Dewar, G.P. Wagner, K. Takahashi, F. Ruddle, C. Ledje, P. Bartsch, J.L. Scemama, E. Stellwag, C. Fried, S.J. Prohaska, P.F. Stadler, and C.T. Amemiya. 2004. Bichir HoxA cluster sequence reveals surprising trends in ray-finned fish genomic evolution. *Genome Res* **14:** 11-17.

Crollius, H.R. and J. Weissenbach. 2005. Fish genomics and biology. *Genome Research* **15:** 1675-1682.

Dugas, J.C. and J. Ngai. 2001. Analysis and characterization of an odorant receptor gene cluster in the zebrafish genome. *Genomics* **71:** 53-65.

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

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

Fujiyama, A., H. Watanabe, A. Toyoda, T.D. Taylor, T. Itoh, S.F. Tsai, H.S. Park, M.L. Yaspo, H. Lehrach, Z. Chen, G. Fu, N. Saitou, K. Osoegawa, P.J. de Jong, Y. Suto, M. Hattori, and Y. Sakaki. 2002. Construction and analysis of a human-chimpanzee comparative clone map. *Science* **295:** 131-134.

Gloriam, D.E., T.K. Bjarnadottir, Y.L. Yan, J.H. Postlethwait, H.B. Schioth, and R.

Fredriksson. 2005. The repertoire of trace amine G-protein-coupled receptors: large expansion in zebrafish. *Mol Phylogenet Evol* **35:** 470-482.

Gregory, S.G., M. Sekhon, J. Schein, S. Zhao, K. Osoegawa, C.E. Scott, R.S. Evans, P.W. Burridge, T.V. Cox, C.A. Fox, R.D. Hutton, I.R. Mullenger, K.J. Phillips, J. Smith, J. Stalker, G.J. Threadgold, E. Birney, K. Wylie, A. Chinwalla, J. Wallis, L. Hillier, J. Carter, T. Gaige, S. Jaeger, C. Kremitzki, D. Layman, J. Maas, R. McGrane, K. Mead, R. Walker, S. Jones, M. Smith, J. Asano, I. Bosdet, S. Chan, S. Chittaranjan, R. Chiu, C. Fjell, D. Fuhrmann, N. Girn, C. Gray, R. Guin, L. Hsiao, M. Krzywinski, R. Kutsche, S.S. Lee, C. Mathewson, C. McLeavy, S. Messervier, S. Ness, P. Pandoh, A.L. Prabhu, P. Saeedi, D. Smailus, L. Spence, J. Stott, S. Taylor, W. Terpstra, M. Tsai, J. Vardy, N. Wye, G. Yang, S. Shatsman, B. Ayodeji, K. Geer, G. Tsegaye, A. Shvartsbeyn, E. Gebregeorgis, M. Krol, D. Russell, L. Overton, J.A. Malek, M. Holmes, M. Heaney, J. Shetty, T. Feldblyum, W.C. Nierman, J.J. Catanese, T. Hubbard, R.H. Waterston, J. Rogers, P.J. de Jong, C.M. Fraser, M. Marra, J.D. McPherson, and D.R. Bentley. 2002. A physical map of the mouse genome. *Nature* **418:** 743-750.

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

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

Karsi, A., D. Cao, P. Li, A. Patterson, A. Kocabas, J. Feng, Z. Ju, K.D. Mickett, and Z. Liu. 2002. Transcriptome analysis of channel catfish (Ictalurus punctatus): initial

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

Kazeto, Y. and J.M. Trant. 2005. Molecular biology of channel catfish brain cytochrome P450 aromatase (CYP19A2): cloning, preovulatory induction of gene expression, hormonal gene regulation and analysis of promoter region. *J Mol Endocrinol* **35:** 571-583.

Kiuchi, S., Y. Inage, H. Hiraiwa, H. Uenishi, and H. Yasue. 2002. Assignment of 280 swine genomic inserts including 31 microsatellites from BAC clones to the swine RH map (IMpRH map). *Mamm Genome* 13: 80-88.

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

Kountikov, E., M. Wilson, S. Quiniou, N. Miller, W. Clem, and E. Bengten. 2005. Genomic organization of the channel catfish CD45 functional gene and CD45 pseudogenes. *Immunogenetics* **57:** 374-383.

Larkin, D.M., A. Everts-van der Wind, M. Rebeiz, P.A. Schweitzer, S. Bachman, C. Green, C.L. Wright, E.J. Campos, L.D. Benson, J. Edwards, L. Liu, K. Osoegawa, J.E. Womack, P.J. de Jong, and H.A. Lewin. 2003. A cattle-human comparative map built with cattle BAC ends and human genome sequence. *Genome Res* **13:** 1966-1972.

Liu, L., Roinishvili, L., Pan, X., Liu, Z., and Kumar, C. 2000. GPMS: A web based genome project management system. *The Proceeding of 4th World Multiconference on Systematics, Cybernectics, and Informatics* SCI2000 62-67.

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

Liu, Z., P. Li, and R.A. Dunham. 1998. Characterization of an A/T-rich family of sequences from channel catfish (Ictalurus punctatus). *Mol Mar Biol Biotechnol* **7:** 232-239.

Liu, Z., P. Li, A. Kocabas, A. Karsi, and Z. Ju. 2001. Microsatellite-containing genes from the channel catfish brain: evidence of trinucleotide repeat expansion in the coding region of nucleotide excision repair gene RAD23B. *Biochem Biophys Res Commun* **289:** 317-324.

Naruse, K., M. Tanaka, K. Mita, A. Shima, J. Postlethwait, and H. Mitani. 2004. A medaka gene map: the trace of ancestral vertebrate proto-chromosomes revealed by comparative gene mapping. *Genome Res* **14:** 820-828.

Ng, S.H., C.G. Artieri, I.E. Bosdet, R. Chiu, R.G. Danzmann, W.S. Davidson, M.M. Ferguson, C.D. Fjell, B. Hoyheim, S.J. Jones, P.J. de Jong, B.F. Koop, M.I. Krzywinski, K. Lubieniecki, M.A. Marra, L.A. Mitchell, C. Mathewson, K. Osoegawa, S.E. Parisotto, R.B. Phillips, M.L. Rise, K.R. von Schalburg, J.E. Schein, H. Shin, A. Siddiqui, J. Thorsen, N. Wye, G. Yang, and B. Zhu. 2005. A physical map of the genome of Atlantic salmon, Salmo salar. *Genomics* **86:** 396-404.

Peatman, E., B. Bao, X. Peng, P. Baoprasertkul, Y. Brady, and Z. Liu. 2006. Catfish CC chemokines: genomic clustering, duplications, and expression after bacterial infection with Edwardsiella ictaluri. *Mol Genet Genomics* **275:** 297-309.

Quiniou, S.M., T. Katagiri, N.W. Miller, M. Wilson, W.R. Wolters, and G.C. Waldbieser. 2003. Construction and characterization of a BAC library from a gynogenetic channel catfish Ictalurus punctatus. *Genet Sel Evol* **35:** 673-683.

Rise, M.L., K.R. von Schalburg, G.D. Brown, M.A. Mawer, R.H. Devlin, N. Kuipers, M. Busby, M. Beetz-Sargent, R. Alberto, A.R. Gibbs, P. Hunt, R. Shukin, J.A. Zeznik, C. Nelson, S.R. Jones, D.E. Smailus, S.J. Jones, J.E. Schein, M.A. Marra, Y.S. Butterfield, J.M. Stott, S.H. Ng, W.S. Davidson, and B.F. Koop. 2004. Development and application of a salmonid EST database and cDNA microarray: data mining and interspecific hybridization characteristics. *Genome Res* **14:** 478-490.

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

Shen, L., T.B. Stuge, E. Bengten, M. Wilson, V.G. Chinchar, J.P. Naftel, J.M. Bernanke, L.W. Clem, and N.W. Miller. 2004. Identification and characterization of clonal NK-like cells from channel catfish (Ictalurus punctatus). *Dev Comp Immunol* **28:** 139-152.

Taylor, J.S., I. Braasch, T. Frickey, A. Meyer, and Y. Van de Peer. 2003. Genome duplication, a trait shared by 22000 species of ray-finned fish. *Genome Res* **13:** 382-390.

Thorsen, J., B. Zhu, E. Frengen, K. Osoegawa, P.J. de Jong, B.F. Koop, W.S. Davidson, and B. Hoyheim. 2005. A highly redundant BAC library of Atlantic salmon (*Salmo salar*): an important tool for salmon projects. *BMC Genomics* 6: 50.

Tilton, F., W.H. Benson, and D. Schlenk. 2002. Evaluation of estrogenic activity from a municipal wastewater treatment plant with predominantly domestic input. *Aquat Toxicol* **61:** 211-224.

Venkatesh, B. 2003. Evolution and diversity of fish genomes. *Curr Opin Genet Dev* **13:** 588-592.

Venter, J.C., H.O. Smith, and L. Hood. 1996. A new strategy for genome sequencing. *Nature* 381: 364-366.

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

Wang, Q., Y. Wang, P. Xu, and Z. Liu. 2006. NK-lysin of channel catfish: gene triplication, sequence variation, and expression analysis. *Mol Immunol* **43:** 1676-1686.

Wang, S., P. Xu, J. Thorsen, B. Zhu, P. de Jong, G. Waldbieser, and Z. Liu. 2007. Characterization of a BAC library from channel catfish Ictalurus punctatus: indications of high rates of evolution among teleost genomes. *Marine Biotechnology*.

Winter, A., A. Alzinger, and R. Fries. 2004. Assessment of the gene content of the chromosomal regions flanking bovine DGAT1. *Genomics* **83:** 172-180.

Woods, I.G., C. Wilson, B. Friedlander, P. Chang, D.K. Reyes, R. Nix, P.D. Kelly, F. Chu, J.H. Postlethwait, and W.S. Talbot. 2005. The zebrafish gene map defines ancestral vertebrate chromosomes. *Genome Res* **15:** 1307-1314.

Woram, R.A., K. Gharbi, T. Sakamoto, B. Hoyheim, L.E. Holm, K. Naish, C. McGowan,

M.M. Ferguson, R.B. Phillips, J. Stein, R. Guyomard, M. Cairney, J.B. Taggart, R. Powell, W. Davidson, and R.G. Danzmann. 2003. Comparative genome analysis of the primary sex-determining locus in salmonid fishes. *Genome Res* **13:** 272-280.

Zhao, S., J. Malek, G. Mahairas, L. Fu, W. Nierman, J.C. Venter, and M.D. Adams. 2000. Human BAC ends quality assessment and sequence analyses. *Genomics* **63:** 321-332.

Zhao, S., S. Shatsman, B. Ayodeji, K. Geer, G. Tsegaye, M. Krol, E. Gebregeorgis, A. Shvartsbeyn, D. Russell, L. Overton, L. Jiang, G. Dimitrov, K. Tran, J. Shetty, J.A. Malek, T. Feldblyum, W.C. Nierman, and C.M. Fraser. 2001. Mouse BAC ends quality assessment and sequence analyses. *Genome Res* **11:** 1736-1745.

IV. A BAC-BASED PHYSICAL MAP OF THE CHANNEL CATFISH GENOME

ABSTRACT

Catfish is the major aquaculture species in the United States. To enhance its genome studies involving genetic linkage and comparative mapping, a bacterial artificial chromosome (BAC) contig-based physical map of the channel catfish (*Ictalurus punctatus Rafinesque*) genome was generated using four color fluorescence-based fingerprints. Fingerprints of 34,580 BAC clones (5.6x genome coverage) were generated for the FPC assembly of the BAC contigs. A total of 3,307 contigs were assembled using a cutoff value of $1e^{-20}$. Each contig contains an average of 9.25 clones, and the average size of the contig is 292 kb. The combined contig size for all contigs was 965,279 kb, approximately the genome size of channel catfish. The reliability of the contig assembly was assessed by hybridizations using two approaches: first, all positive BAC clones to 10 gene probes were checked to determine if all positive clones for a specific gene fall into a single contig; second, randomly selected contigs were validated using overgo probes designed from BAC end sequences. This physical map should greatly enhance genome research in catfish, especially for the identification of genomic regions containing genes underlining important performance traits.

INTRODUCTION

Channel catfish (*Ictalurus punctatus Rafinesque*) is the major aquaculture species accounting for over 60% of all US aquaculture production. In 2006, its production reached almost 700 million pounds in the U.S. In addition to its importance for aquaculture, catfish is also one of the top sport fishing species in North America.

Channel catfish has long been also a research model for comparative immunology and toxicology. Its unique characteristics such as its ability to adapt to changing environment, e.g. low oxygen, high pressure and changing temperature, are well beyond the capability of higher vertebrates such as mammals. Therefore, genomic studies of an aquaculture fish species might provide new insight addressing genetic mechanisms of performance traits in aquatic environments as well as genome evolution. Catfish is an excellent model organism for genomic studies, particularly for aquaculture important issues. Its high reproductive ability/fertility allows breeding of large families with thousands of progenies. This offers a great opportunity for QTL scans and extensive phenotype selection using selective genotyping thereby reducing time, money, and efforts in QTL analysis. Channel catfish's evolutionary position allows comparison of its genome information with that of the model species zebrafish, accelerating catfish genome research while facilitating zebrafish genome annotation.

Rapid progress in catfish genomics has been made in the last several years. Large numbers of molecular markers have been developed and evaluated for linkage mapping (Serapion et al. 2004; Xu et al. 2006), framework genetic linkage maps have been constructed (Liu et al. 2003; Waldbieser et al. 2001) and genome repeat structure has been characterized (Nandi et al. 2006; Xu et al. 2006). More than 55,000 ESTs have been

generated (Cao et al. 2001; Ju et al. 2000; Karsi et al. 2002; Kocabas et al. 2002; Li et al. 2007), and an ongoing large-scale EST project by the Joint Genome Institute of the Department of Energy will further significantly expand the EST resources in both channel catfish and blue catfish (He et al. 2003). Microarrays have been used to study genome-wide expression in catfish (Ju et al. 2002; Kocabas et al. 2004; Li and Waldbieser 2006; Peatman et al. 2007). Two bacterial artificial chromosome (BAC) libraries using different restriction endonucleases have been constructed and characterized (Quiniou et al. 2003; Wang et al. 2007). More than 20,000 BAC end sequences (BES) from the channel catfish CHORI-212 library have been generated and characterized (Xu et al. 2006). Of the two BAC libraries, CCBL1 was constructed using DNA from a homozygous gynogenetic female. Typically, gynogens were produced by eggs induced to develop by using sperm killed with radiation; the diploid state was restored by hydrostatic pressure shocks that induce the retention of the second polar body. The other BAC library, CHORI-212, was constructed using DNA from a normal male catfish where the genomic DNA contains all autosomes and sex chromosomes, and the normal level of polymorphism. The two libraries were also constructed using different restriction endonucleases, Hind III for CCBL1, and EcoRI for CHORI-212. BAC contigs have been developed from CCBL1 (Quiniou et al. 2007). In this work, the objective was to construct a BAC contig-based physical map using the CHORI-212 BAC library.

A BAC-based physical map is important for the understanding of genome structure and organization, and for position-based cloning of economically important genes. A well characterized physical map can often be an important foundation for whole genome sequencing. A BAC-based physical map would also allow exploitation of existing

genomic information from map-rich species using comparative mapping. Because of their importance, physical maps have recently been constructed in aquaculture fin fish species including Nile tilapia (*Oreochromis niloticus Linnaeus*) and Atlantic salmon (*Salmo salar Linnaeus*) (Katagiri et al. 2005; Ng et al. 2005). Here we report the construction of a BAC contig-based physical map of the channel catfish genome.

RESULTS

BAC fingerprinting

A total of 40,416 BAC clones was processed from the channel catfish BAC library CHORI-212, and 34,580 (85.6% success) fingerprints were validated and used in the final FPC assembly. The valid fingerprints represent approximately 5.6-fold coverage of

Table 1. Statistics of the BAC contig assembly of the catfish genome

Total number of BAC clones fingerprinted	40,416	6.5X genome equivalent
Valid fingerprints for FPC assembly	34,580	5.6X genome coverage
Total number of contigs assembled	3,307	
Clones contained in the 3,307 contigs	30,582	
Average BAC clones per contig	9.25	
Average estimated size per contig	292 kb	
Number of Q-contigs	517	
Number of Q-clones	1,494	4.3%
Number of singletons	3,998	
Number of clones contained in the top 50% contigs	25,398	83%
Number of top contigs that contained 50% of clones	580	17.5%
Average insert size of the BAC library	161 kb	
Average number of bands per fingerprinted BAC clone	95.2	
Average size each band represents	1.6912 kb	
Total number of bands included in the contigs	570,766	18.7 bands per BAC clone in the consensus map
Total physical length of assembled contigs	965,279 kb	~1 X genome size

the catfish genome, while the total number of processed clones represents approximately 6.5-fold coverage of the catfish genome (Table 1). Each BAC clone contains, on average, 95.2 restriction fragments, with 60 to 120 bands in most of the samples (Figure 1).

Using a tolerance of 4 and a cutoff stringency of $1e^{-20}$ (see below), I tested FPC assembly of 5,000, 10,000, 15,000, 20,000, 25,000 and 30,000 BAC clones respectively to assess if the number of fingerprinted BAC clones was sufficient to cover the entire genome. When the assembled contig numbers were plotted against the clone numbers (Figure 2), the contig number reached a plateau when the total clone number reached 25,000, suggesting that the number of fingerprinted BAC clones was sufficient to cover the catfish genome.

Figure 1. The distribution of the band numbers in the catfish fingerprint.

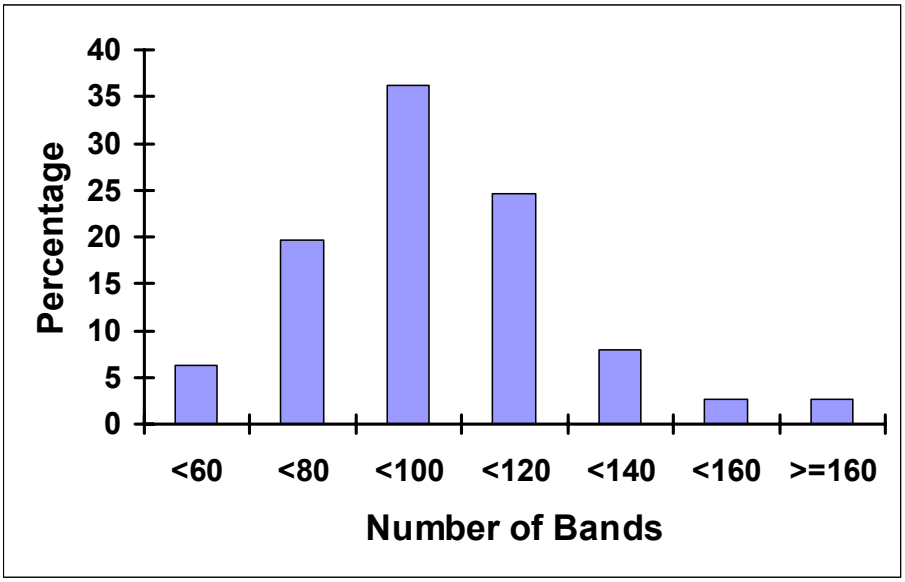
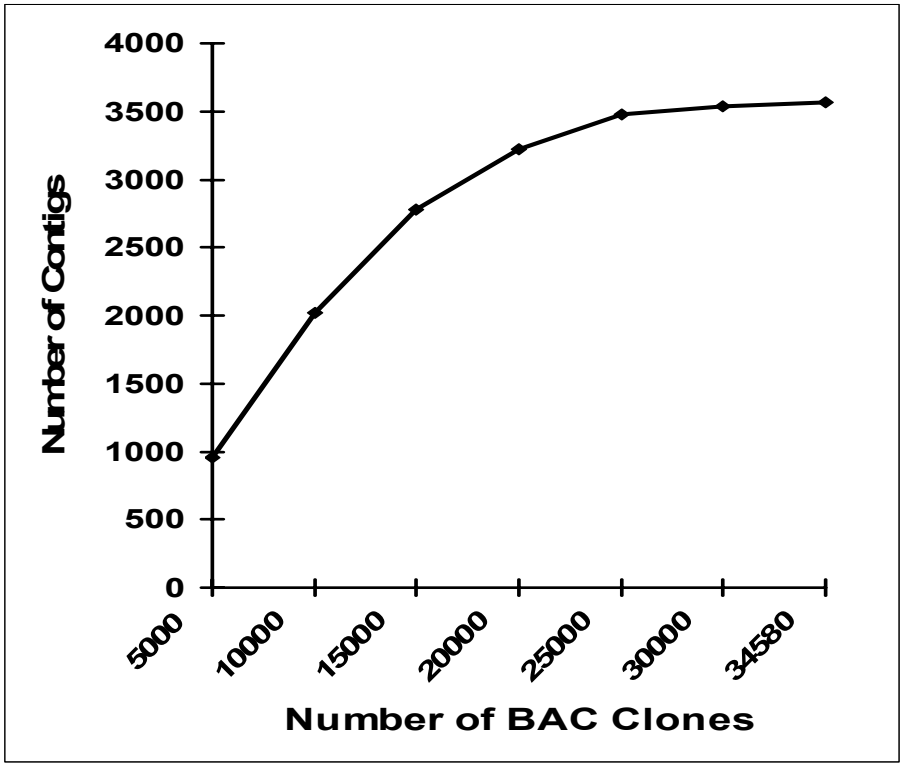


Figure 2. The relationship of the number of fingerprinted BAC clones and the number of BAC contigs assembled using a cutoff value of 1e-20 and a tolerance of 0.4.



Determination of tolerance

The tolerance level dictates how closely two restriction fragments must match to consider them the same fragment across gel runs. The tolerance was determined by identifying identical fragments in many different fingerprints and computing the standard deviation of their sizes in different fingerprints. For this purpose, vector fragments from BAC vector pTARBAC2.1 were identified from 300 randomly picked fingerprints and the standard deviation of each vector fragment was computed. The standard deviations of the three vector fragments (59.1 bp, 157.3 bp and 369.8 bp) were 0.099 bp, 0.081 bp and 0.085 bp, respectively, with an average of 0.088 bp (Figure 3A). Since the three vector fragments did not cover the whole range of fragment sizes (50-500 bp), I also computed

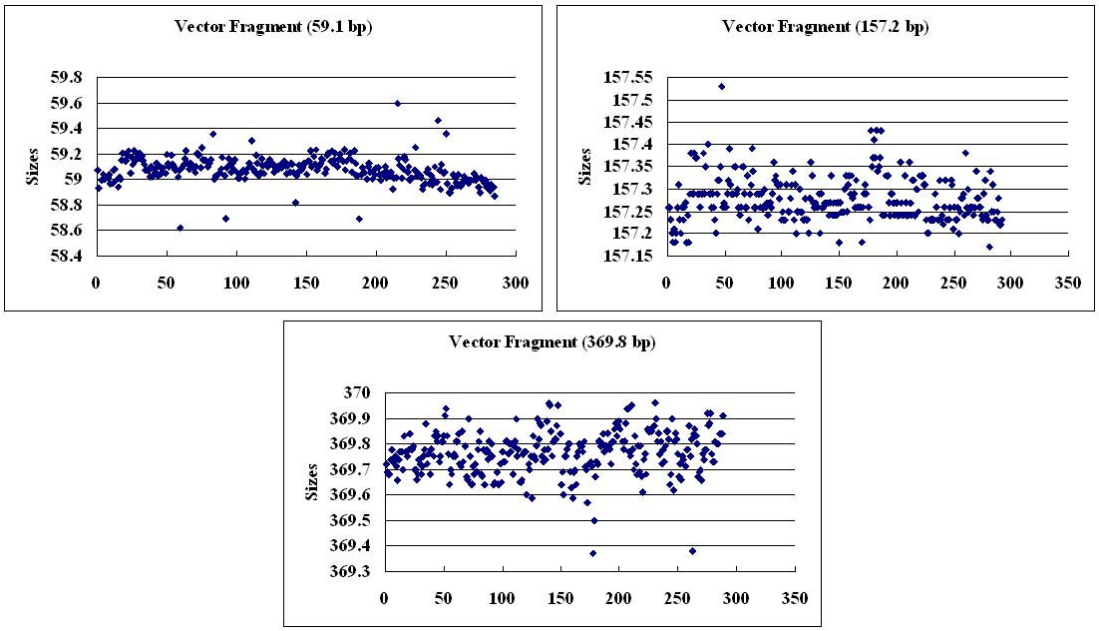
the standard deviations of selected GS500-LIZ internal size standard fragments (Applied Biosystems, Foster City, CA), 100 bp, 160 bp, 340 bp and 490 bp. The standard deviations lay between 0.06 and 0.13, with an overall average of 0.087 bp (Figure 3B). Thus, the tolerance value was estimated at 0.36 according to the size deviation with 95% confidence interval. The tolerance value used in FPC assembly was set at 4 since all the fragments sizes were multiplied by 10, and decimals were not allowed in the FPC program.

Determination of cutoff values for the contig assembly

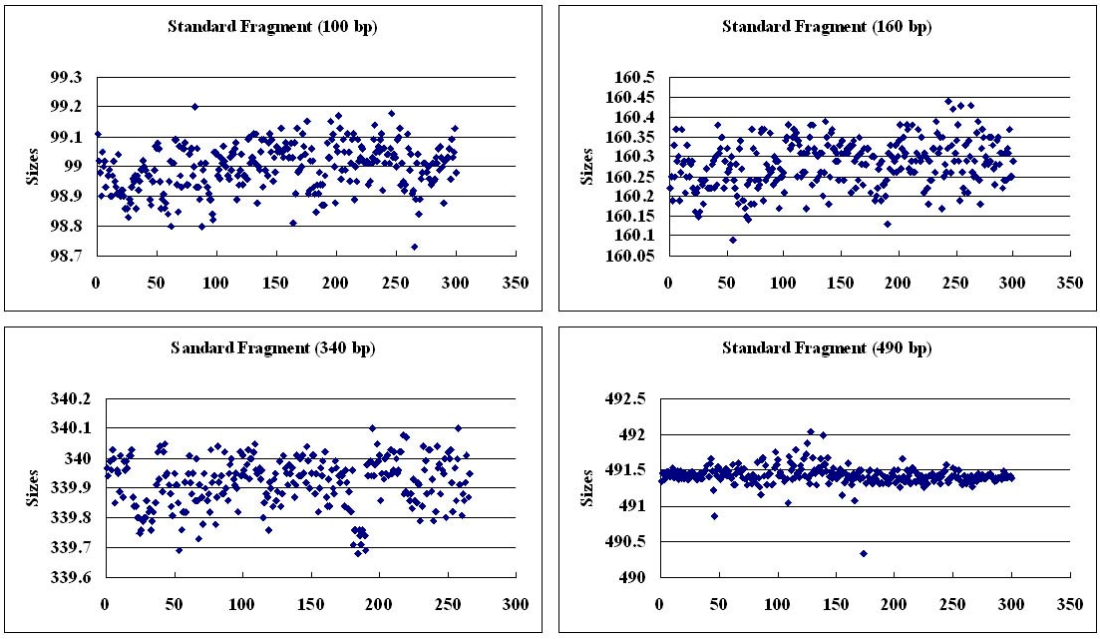
The cutoff value is the threshold of the Sulston score, the probability that fingerprint bands match by coincidence. Lowering the cutoff value (e.g. $1e^{-12}$ to $1e^{-15}$) would increase the stringency and therefore increase the likelihood that reported overlapping BAC clones are truly overlapping. However, setting an appropriate stringency of Sulston score is always challenging; too low a cutoff would lead to splitting of true contigs into multiple contigs or singletons, whereas too high a cutoff would lead to chimeric contigs. During the assembly of the catfish physical map, a series of cutoff values ranging from $1e^{-12}$ to $1e^{-40}$ were tested. The resulting number of contigs, Q-contigs (questionable-contigs), singletons, and number of Q-clones (questionable-clones) were considered. At high stringencies ($1e^{-25}$ to $1e^{-40}$), the number of singletons increased drastically (Figure 4), causing many contigs to collapse. As expected, a lower number of contigs were assembled using lower stringencies: only 1,798 and 2,460 contigs resulted using a cutoff value of $1e^{-12}$ and $1e^{-15}$, respectively; but a large number of clones were in the category of Q-clones, 31.0% and 13.7%, respectively. Clearly, these stringencies ($1e^{-12}$, $1e^{-15}$) were too low as almost 1/3 of the contigs contain Q-clones, and the

Figure 3. The size distributions of (A) vector fragments and (B) size standard fragments from GS500-LIZ in 300 randomly selected fingerprinting samples.

(A)

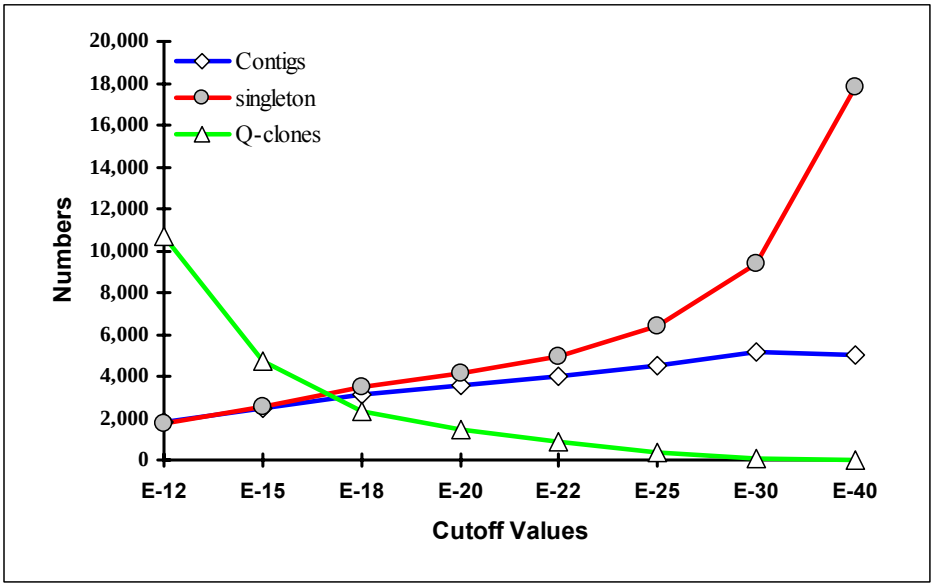


(B)



percentage of Q-clones was too high. It was noted from the plot of the numbers of contigs, singletons, and Q-clones versus the assembly stringencies that these values cross over each other at approximately $1e^{-17}$ where the number of contigs, the number of singletons, and the number of Q-clones were all reasonably low (Figure 4). This provided a starting point for determining the proper cutoff value. At the stringency of $1e^{-18}$ and $1e^{-20}$, all indicators were similar, suggesting these cutoff values provide relatively stable assemblies. However, the number of Q-contigs was significantly more with $1e^{-18}$ than $1e^{-20}$ (Figure 4). Higher levels of stringency, such as $1e^{-22}$ and $1e^{-25}$, would increase the reliability of the assembly, but also increase the chances of splitting true contigs. Further experiments using overgo hybridizations (see below) demonstrated that the cutoff values of $1e^{-22}$ and $1e^{-25}$ were too stringent leading to the breaking of many *bona fide* contigs. I therefore chose $1e^{-20}$ as the cutoff value for the assembly of the physical map.

Figure 4. Plot of the number of contigs, singletons, and Q-clones over the stringencies used for the assemblies.



Contig assembly

Contigs were assembled from the fingerprint data using the computer program FPC version 8.5 (http://www.agcol.arizona.edu/software/fpc/). A total of 3,307 contigs were assembled with the valid fingerprints of 34,580 BAC clones using FPC with a cutoff value of $1e^{-20}$ and a tolerance level of 4, followed by end-to-end merging and end-to-single merging at progressively lower stringencies until $1e^{-12}$. A total of 30,582 clones were placed into the 3,307 contigs, leaving the remaining 3,998 BAC clones as singletons. The assembly can be accessed through a web-based physical map viewer, WebFPC, at http://titan.biotec.uiuc.edu/WebAGCoL/AU02-20/WebFPC/ (Figure 5).

The contig size (clones per contig) distribution is shown in Figure 6. The top half of the contigs (1,654 contigs) contained 83% (25,398 clones) of the total assembled clones (Table 1). The top 580 contigs (17.5%) contained 50% of BACs in the contigs. The largest contig contained 364 BAC clones, while the smallest contig contained 2 BAC clones. The contigs contained an average of 9.25 clones each, and had an average estimated length of 292 kb per contig. The contig examples are demonstrated in Figure 7.

There were a total of 570,766 consensus bands distributed in the 3,307 contigs, representing approximately 0.96 Gb (965,279 kb) linear length of DNA according to the average band size of 1.69 kb (the BACs have an average insert size of 161 kb and produced an average of 95.2 bands per BAC, and therefore, each band represented an average segment size of 1.69 kb), equivalent to genome size of channel catfish (Tiersch and Goudie 1993). On average, each BAC in the contigs contributed 18.7 unique consensus bands to the assembly, or approximately 31.6 kb to the linear length of the contig assembly.

Figure 5. WebFPC display of the assembly using a cutoff value of 1e-20 and a tolerance of 0.4.

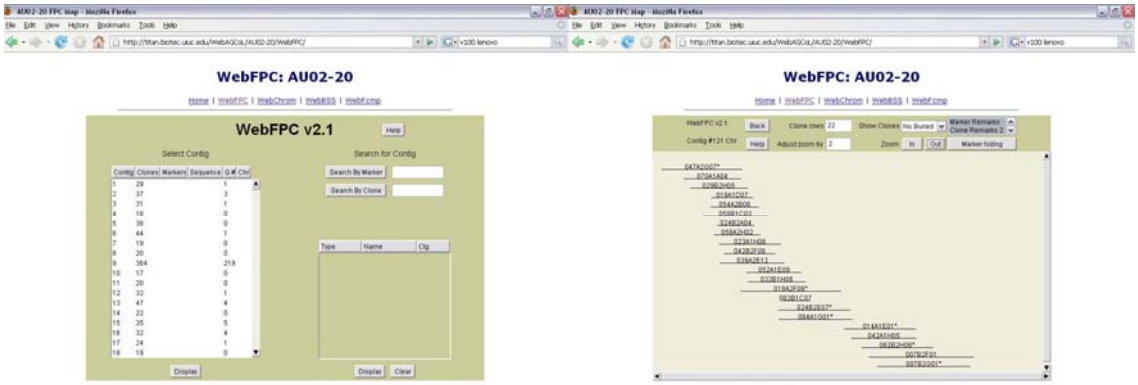


Figure 6. Distribution of BAC clones in contigs of various sizes.

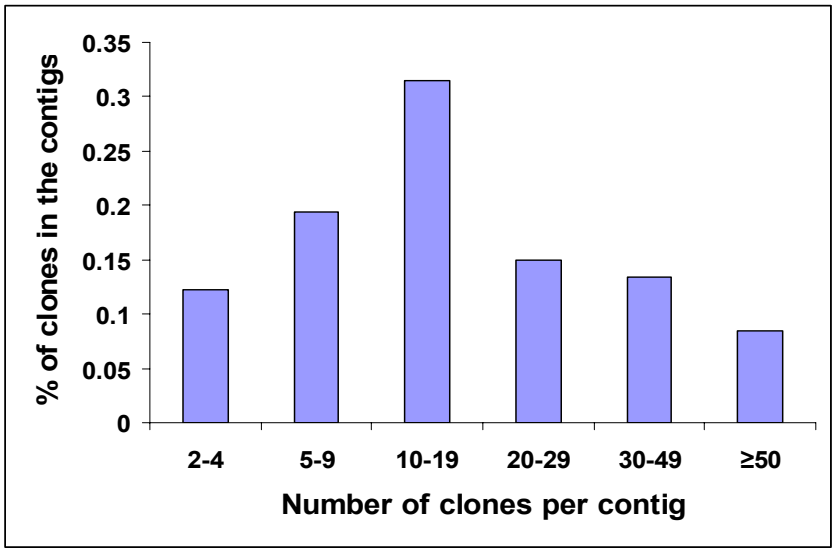
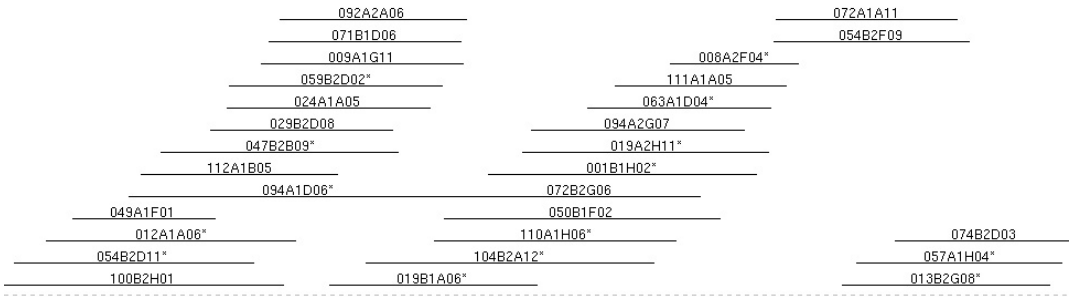
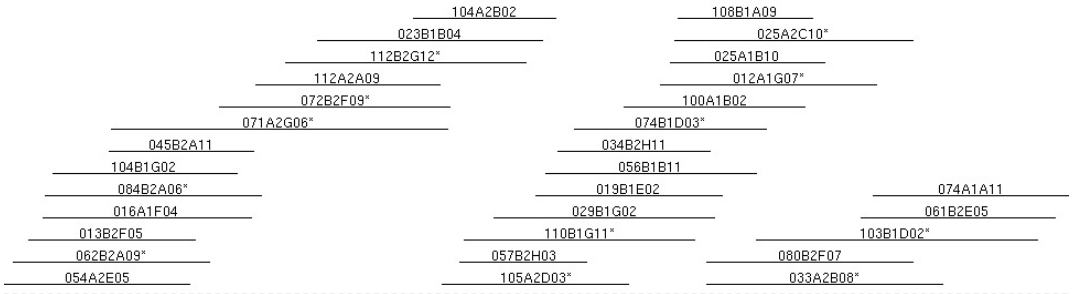


Figure 7. Example BAC contigs assembled in FPC using cutoff value of $1e^{-20}$ and tolerance of 0.4 bp.

(A) BAC Contig # 74



(B) BAC Contig # 276



Q contigs and dQ process

There were a total of 1,494 questionable clones (4.3%) distributed in 517 Q-contigs. However, the distribution of the Q-clones was uneven. The vast majority of contigs (84.4%) were free of Q-clones; 301 contigs had only one Q-clone each; 96 and 44 contigs had 2 and 3 Q-clones, respectively, while the vast majority of Q-clones were placed into a small number of contigs (Table 2). This indicated that it is likely that some highly repetitive elements could have been involved in the contigs containing many Q-clones, and they need to be assembled at a higher stringency. The dQer in the FPC program was

used to eliminate the Q-clones in the Q contigs with more than 5 Q-clones. The dQer automatically reran the assembly algorithm with a lower cutoff of $1e^{-21}$, $1e^{-22}$ and $1e^{-23}$, split Q contigs, and then assembled the generated contigs and singletons with existing contigs. Finally, dQer generated 523 Q-contigs, in which 15 contigs still have more than 5 Q-clones. The final contig and singleton numbers were 3,366 and 4,104, respectively. There were only 55 contigs impacted in the dQer reassembly, which suggested high stability of the contig assembly.

Table 2. Distribution of Q-clones in contigs assembled using a cutoff value of $1e^{-20}$. Note that 84.4% of the contigs are free of Q-clones, and the most Q-clones were involved in a small number of contigs.

Number of contigs	Q-clones/contig	Percentage of all contigs
2790	0	84.4%
301	1	9.1%
96	2	2.9%
44	3	1.3%
19	4	0.6%
16	5	0.5%
41	>5	1.2%

Assessment of the physical map reliability

Several approaches were used to assess the quality of the contig assembly. First, I checked if the BAC clones containing known genes were actually assembled into the same contigs. Overgo probes were designed for known genes using available cDNA sequences. Overgo hybridization was used to screen the high-density CHORI-212 channel catfish BAC library filters. All the positive clones were identified. The positive

clones of each gene were tabulated to test the validity of contigs assembled using FPC with cutoff of $1e^{-20}$ and tolerance of 4 (Table 3). In general, two types of situations were observed. In the first case, all the fingerprinted positive clones fell within a single contig, confirming the reliability of the assembly; these included the genes for bactericidal permeability-increasing protein (BPI), interleukin-1beta (IL-1β), chemokine CXCL8, and CXCL12. In the second case, the fingerprinted positive clones were split into two contigs or a major contig plus singletons, suggesting either the stringency of the assembly was too high, or these genes are duplicated in the catfish genome; these included genes for liver-expressed antimicrobial peptide 2 (LEAP-2), hepcidin, CXCL10, CXCL14, TLR20, and TLR21.

Second, map reliability was assessed by checking randomly selected contigs to determine if all the BAC clones truly belong to the contigs. If all the BAC clones truly belong to the contig, then use of a few probes should allow hybridization of all clones in the contig, thereby validating the contig. As shown in Figure 8, use of three probes allowed validation of contig 1046 as the three probes collectively hybridized to all BAC clones. Similarly, a total of five randomly selected contigs were all validated this way (Table 4), providing strong evidence for the high reliability of the physical map.

Third, after initial assembly at the cutoff of $1e^{-20}$, trial assemblies were conducted at higher stringencies such as $1e^{-22}$ and $1e^{-25}$. As the stringency was increased, some contigs were split into two or more contigs. Based on the position of the breaking points, overgo probes were designed from BAC end sequences from the clone covering the breaking points. In all 10 instances examined, the overgo probes hybridized to both splitting contigs, confirming that the contigs split at the cutoff values of $1e^{-22}$ or $1e^{-25}$ were indeed

bona fide contigs, and they were split simply because the overlapping regions were not long enough, and the stringency was too high (Table 5). This set of hybridization experiments not only provided strong evidence for the high reliability of the contig assembly, but also provided a strong experimental basis for the selection of assembly cutoff values.

Figure 8. An example of contig validation using overgo hybridization. Three probes were used (dark color with different shapes: ✸ , ★ and ●). Hybridizations using these probes allowed detection of all positive clones to each probe (with same shaped symbols, but faded fillings). The three probes collectively hybridized to all clones with many hybridizing to multiple clones allowing confirmation of the contig.

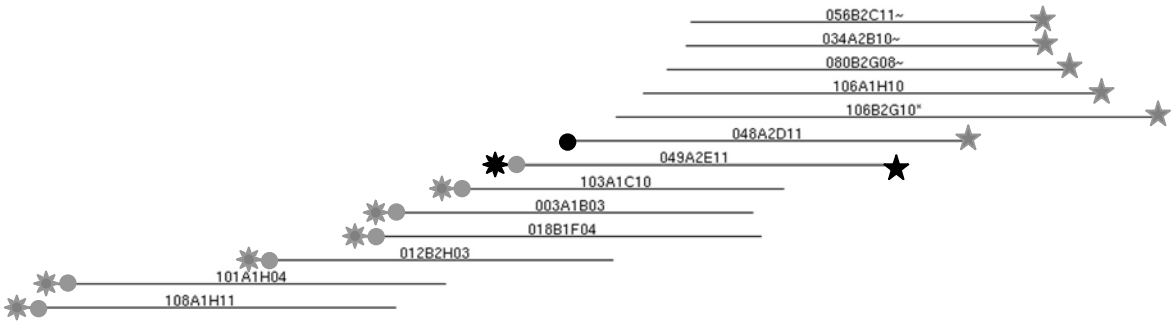


Table 3. Assembly of BAC clones positive for selected gene probes. *Note that hybridization was conducted using the high density filters containing all BAC clones in the CHORI 212 library, whereas only a proportion of the CHORI BAC library was fingerprinted for the construction of the physical map. Bold clones indicate the successfully fingerprinted clones. Italic clones indicate clones where the fingerprints did not pass the quality cutoff. Clones in normal Latin font were not fingerprinted.

Genes	Positive clones*	Contigs	Clones in the contig	# in the same contig
BPI	**005B2E02, 013A1F03, 017A1B07, 030B1E12, 077A1G10, 104A2A02, 107B1H12,** 126B1D02, 174B2D01	212	005B2E02, 013A1F03, 017A1B07, 030B1E12, 077A1G10, 104A2A02, 107B1H12	100%
LEAP-2	*007B1D12,* *009A2G03,* **023A1G01, 027B1B07, 033B1E02, 035A2F08, 063B2E04, 063B2F04,** *066B2D04,* **076A2D04, 080A1B04, 087B2B06, 101B1C12**, 113B2C06, 113A2D02, 134A1A05, 138A2G04, 154A2A05, 160A2H10, 171A2F06, 172B2A09, 178B1E12, 178A1H11	1162 2525	023A1G01, 033B1E02, 063B2E04, 063B2F04, 076A2D04, , 087B2B06, 101B1C12 027B1B07, 035A2F08, 080A1B04	Split into 2 contigs
IL-1β	*017B1E01,* **028A2A02, 039B1H01, 039B1H11, 042B2H09, 051A1A07, 072A1A09,** *089B1F09,* *093A1B01,* *094A1H03,* **102B2C12,** 175B2F11, 187A1H06	699	028A2A02, 039B1H01, 039B1H11, 042B2H09, 051A1A07, 072A1A09, 102B2C12	100%
Hepcidin	**010B2C11, 047B1B09, 049B2G01, 090B2H07, 095A1G05, 102B1F01**, 132A2D08, 176A2C06, 188B2A07	948 1788	010B2C11, 049B2G01, 095A1G05, 102B1F01 047B1B09, 090B2H07	Split into 2 contigs
TLR20	**005A1G05, 047A1G09, 050B1C09, 070A1H02, 093A2B06, 105B2D06,** 136A2B09	1754	005A1G05, 050B1C09, 070A1H02, 105B2D06	4 of the 6 clones in this contig
TLR21	**006A2D09, 009B1A01,** *025A1E09,* **075A1G05, 086A1H05,** *092A2D05,* **104A1C08, 108B1B01,** 160B2F10, 161B1F08, 164A1E11, 176B2H06, 192B2H03	1703	006A2D09, 009B1A01, 086A1H05, 104A1C08, 108B1B01	5 of the 6 clones in this contig
CXCL8 -like	**001B1A07,** *025A2D04,* **053B2F01, 056B1C06, 066A2A05,** 189A1D12	1734	001B1A07, 053B2F01, 056B1C06, 066A2A05	100%
CXCL10	**010B1E05,** *016A2F06,* **024A1D07, 036B2D06, 037A1C06,** *050A1C12,* **056B1C06, 060A1A04, 063A2E12, 107A1B01,** 176B1G02, 180B2B03, 182A1B12	972 1734	024A1D07, 060A1A04, 063A2E12, 107A1B01 036B2D06, 037A1C06, 056B1C06,	Split into 2 contigs
CXCL12	**019B2H05, 024A1G02, 031A2H09, 034B1C09,** *037A2C01,* **038B1B07, 052A1G07, 059B1E12, 068B2C11, 081B1F10, 090A1G01, 096B1D08,** *099B1B07,* **111A2D11,** 130A1E05, 135B1C01, 136A1B04, 139A2G09, 142A2D12, 158B2E11, 160B1B06, 170B2G05, 184A1F01, 185B2H12, 189B2E12, 191B1G09	766	019B2H05, 024A1G02, 031A2H09, 034B1C09, 038B1B07, 052A1G07, 059B1E12, 068B2C11, 081B1F10, 090A1G01, 096B1D08, 111A2D11	100%
CXCL14	**010A2E01, 020B1C02, 026B1B01, 035A2F03, 039B1G09, 040B2C05, 053A1G05, 060A1F01, 079B2H12,** 118B1H10, 123A2E12, 143A1D10, 144B1E02, 164B2C01	981	010A2E01, 020B1C02, 026B1B01, 035A2F03, 039B1G09, 040B2C05, 053A1G05, 060A1F01	8 of the 9 clones in this contig

Table 4. Assessment of map reliability using overgo probes designed from BAC end sequences. A number of probes were designed from BAC end sequences such that all clones in the contigs were positive (also see Fig. 3). Note that some clones are positive to multiple probes leading the larger number of positives than the total number of clones in the contig, thereby colleterality was established.

Contig at $1e^{-20}$	Number of BACs	Number of probes	Total number of positive clones	Contig assembly completely validated
1046	13	3	22	Yes
1558	11	2	12	Yes
673	16	2	18	Yes
586	13	4	29	Yes
284	17	2	21	Yes

Table 5. Validation of contigs through overgo hybridizations and collateral inferring. Only one overgo probe was used to cover the break points of randomly selected contigs when one single contig assembled at a lower stringency ($1e^{-20}$) was split into more than one contigs at a higher stringency ($1e^{-22}$ or $1e^{-25}$). Note that the contigs were assigned different numbers in each assembly. Numbers in the parenthesis are the number of clones in the contigs. For instance, Contig 1246(12) was one contig containing 12 BACs in the assembly using a cut off value of $1e^{-20}$, this contig remained as a single contig in the assembly using a cut off value of $1e^{-22}$ (though with a different contig number, contig 1405 now); however, this contig was split into two contigs plus a singleton in the assembly using a cut off value of $1e^{-25}$. An overgo probe designed from the singleton BAC clone hybridized to some clones contained within contig 1477 and contig 3765 as well as to itself, providing evidence that the contig was split because of the high assembly stringency.

Contigs $1e^{-20}$	Contigs $1e^{-22}$	Contigs $1e^{-25}$	Probe location	Positive contigs
Contig1246(12)	Contig1405(12)	Contig1477(8) Contig3765(3) 1 singleton	Singleton	Contig1477 Contig3765 Singleton
Contig135(42)	Contig138(42)	Contig148(19) Contig492(22) 1 singleton	Contig492	Contig148 Contig492
Contig199(51)	Contig261(30) Contig192(17) Contig3000(4)	Contig3413(4) Contig271(30) Contig175(17)	Contig271	Contig3413 Contig271
Contig276(53)	Contig275(41) Contig793(12)	Contig322(41) Contig974(12)	Contig322	Contig974 Contig322
Contig121(41)	Contig118(36) Contig2079(4) 1 singleton	Contig2470(4) Contig107(33) Contig 3373(3) 1 singleton	Contig107	Contig2470 Contig107
Contig1046(13)	Contig1088(13)	Contig1104(7) Contig2624(6)	Contig1104	Contig2624 Contig1104
Contig1997(10)	Contig1435(10)	Contig2972(7) Contig4477(3)	Contig2972	Contig4477 Contig2972
Contig958(21)	Contig2616(3) Contig980(16) Contig3724(2)	Contig1116(10) Contig2064(6) Contig2909(3) Contig4363(2)	Contig1116	Contig2909 Contig2064 Contig1116
Contig 366(24)	Contig 364(24)	Contig1915(4) Contig339(20)	Contig1915	Contig339 Contig1915
Contig579(21)	Contig591(14) Contig2340(7)	Contig642(14) Contig2853(3) Contig3277(4)	Contig642	Contig2853 Contig3277 Contig642

DISCUSSION

This work produced a high quality BAC-based physical map of the diploid catfish genome with 3,307 contigs spanning approximately 0.965 Gb, equivalent to the size of the catfish genome. The generation of this physical map filled a critical gap in catfish genome research. This BAC-based physical map should provide a material and information basis for comparative mapping (Leeb et al. 2006; Romanov et al. 2006) and large scale genome analysis of the catfish genome using existing genome sequence

information from several model fish species. The availability of the contig information and BAC end associated polymorphic markers also provide opportunity for integration of the physical map with genetic linkage maps. Polymorphic microsatellite markers are being generated from the BAC end sequences, mapping of which to the genetic linkage map would allow integration of the genetic and physical maps. With the physical map, chromosomal regional markers can be developed from targeted genomic regions for fine mapping of candidate genes associated with performance traits important to aquaculture, laying grounds for eventual positional cloning of economically important genes (Fernando et al. 2007). This physical map will also allow generation of a minimal tiling path in preparation for whole genome sequencing.

A plot analysis of the fingerprinted clones versus the number of contigs (Figure 2) indicated that the fingerprinted BACs (5.6X genome coverage) should provide a reasonable coverage of the whole genome of catfish. Although the calculated consensus band (CB) map distance was the same size as the catfish genome, the actual physical map length could be longer as the 3,998 singletons were not included in the map length. In addition, there would be gaps among the contigs as well. However, such gaps and map distances represented by non-contig singletons could be offset by undetected overlaps among the contigs. The CB map estimation is very similar to that estimated from a gynogen catfish (Quiniou et al. 2007) where the physical map size was estimated to be 0.93 Gb. Considering the lack of a Y-chromosome in the gynogen genome, the slightly larger estimation here using a normal diploid male would be expected. The number of consensus bands used for the assembly was also very similar (516,956 in the paper of Quiniou et al. 2007 and 570,766 here).

The contig numbers initially increased with the number of clones being fingerprinted. However, the number of contigs reached a plateau around 25,000 clones. Due to budget limitations, fingerprinting of additional BAC clones was not possible at this time. Additional fingerprinting in the future could potentially fill some gaps allowing contigs to be merged, thereby reducing the total number of contigs. However, the use of a complementary BAC library using a different restriction enzyme may be more effective in gap filling as some genomic regions would have been left out during library construction using restriction digestion.

One key issue for the assembly of a physical map is the selection of a proper cutoff value. With agarose gels, a Sulston cutoff value of 3×10^{-12} was used for the human genome (3×10^{9} bp) for automated assembly (McPherson et al. 2001), and a less stringent score of 1×10^{-9} was used for the smaller *Arabidopsis* genome (Marra et al. 1997). For the construction of the tilapia BAC contigs, a cutoff value of 10^{-8} was used (Katagiri et al. 2005), and for the Atlantic salmon Ng et al. used a cutoff value of 10^{-16} for initial contig assembly, then the contigs were merged with a cutoff value of 10^{-10} (Ng et al. 2005). A more stringent cutoff value is required for the assembly of fingerprints produced using high information content fingerprinting (HICF) (Nelson et al. 2005) as compared to the assemblies of fingerprints produced by agarose gels. However, the use of too stringent a cutoff value could lead to split of many *bona fide* contigs. In a recent assembly of BAC contigs using HICF, a cutoff value of $1e^{-40}$ was used for the contig assembly (Quiniou et al. 2007). However, the BAC library, CCBL1, was constructed using DNA from a gynogenetic female whose genome is homozygous, whereby sequence polymorphism is minimal, if any. It is obvious that the level of sequence polymorphism, reflected in the

DNA from a single diploid organism as sequence differences between the two sets of homologous chromosomes, could greatly affect the choice of cutoff values. The greater the sequence divergence, the less stringent the cutoff value should be. In addition, the genome size should be considered for the selection of the cutoff value. In this regard, the catfish genome is approximately 1/3 of the size of the Atlantic salmon genome, and is similar to the size of the tilapia genome. With all existing information, I believe that the assembly using $1e^{-20}$ should provide a reasonably conservative assembly that can easily be updated when more genome data of channel catfish becomes available. The overgo hybridization experiments strongly support the use of $1e^{-20}$ as the proper cutoff value for the assembly of the catfish physical map. At this cutoff, the percentage of Q-clones was reasonably low (4.3%) (for comparison, 7.3% Q-clones in the paper of Quiniou et al. 2007).

The quality of the physical map was assessed by both hybridizations of selected genes, or validation of randomly selected contigs using overgo hybridizations with probes designed from BAC end sequences. The vast majority of contigs (84.4%) were free of Q-clones. Even in the contigs with Q-clones, these may not be a result of fingerprinting analysis. They could represent truly questionable clones caused by several means. First, teleost fish genomes are well known for their whole genome duplication events (Brunet et al. 2006; Jaillon et al. 2004; Woods et al. 2005). In addition to the whole genome duplications, teleost fish also exhibit a high level of tandem and segmental gene duplications (Peatman et al. 2006; Peatman and Liu 2006; Peatman and Liu 2007). Such genome duplications would certainly add complexity to physical genome analysis including the possibility of producing Q-clones. Secondly, the use of four sets of

restriction endonucleases, while providing great advantages, also increased the sensitivity for the detection of polymorphism. Some Q-clones could truly represent polymorphic genomic regions derived from homologous chromosomes of the diploid catfish.

A large number of Q-clones generally result from one or several false positive overlaps. Once a clone in the contig overlaps with another clone in another contig, the whole contig will be brought to that contig. However, in this case, FPC may not assign the appropriate linear order to each clone on the consensus band (CB) map. So the clones coming from different contigs stack on top of each other. The dQer can automatically increase the cutoff stringencies and split the Q contigs. However, there are still many contigs harboring Q-clones. Most possibly, the Q contigs could be caused by duplicated genome regions or repetitive elements in the genome. Repetitive sequences occupy a significant fraction of the catfish genome (Liu et al. 1998; Nandi et al. 2006; Xu et al. 2006). In previous analysis of repeat sequences from 11.4 million base pairs of the channel catfish BAC end sequences, approximately 11% sequences were masked by Repeatmasker software using zebrafish and *Takifugu* repeat databases (Xu et al. 2006). Some of the Q-clones could have been attributed to such repetitive sequences.

MATERIALS AND METHODS

BAC library and BAC fingerprinting

The CHORI-212 BAC library contains a total of 72,067 recombinant clones with average insert size of 161 kb, representing approximate 10.6x coverage of channel catfish genome (Wang et al. 2007).

BAC clones were inoculated into 2.2-ml 96-well culturing blocks with each well containing 1.5 ml 2×YT medium and 12.5 μg/ml chloramphenicol from 384-well stocking plates using a 96-pin replicator (V&P Scientific, Inc., San Diego, CA). Each 384-well plate of BAC clones was inoculated into four 96-well culturing blocks. To assure clone tracking, the BAC clones were always taken using the 96-pin replicator with the first pin of the replicator aligned with position A01 of the 384-well plates as the set "A" samples (which takes A01, A03, A05…., C01, C03, C05…, … and O01, O03, O05…, followed with the first pin of the replicator aligned with position A02, B01, and B02 as the "B" set, "C" set, and "D" set of samples. The four sets of samples were later decoded to their original 384-well locations. The 96-well culture blocks were covered with air permeable seals (Excel Scientific, Wrightwood, CA) and incubated at 37° C for 24 hours on a HiGro shaker incubator (Gene Machines, Inc., San Carlos, CA, USA) at 450 rpm. The blocks were centrifuged at 2500g for 10 min in an Eppendorf 5810 bench top centrifuge to precipitate the 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 Qiagen R.E.A.L Prep 96 plasmid kit (Valencia, CA) according to the manufacturer's instructions. BAC DNA was collected in 96-well plates and stored in -20°C before use.

For fingerprinting, the four fluorescence labeled restriction fragments were first created using the SnapShot kit of the Applied Biosystems (Foster City, CA) as previously described (Luo et al. 2003). BAC DNA was digested by *Bam*HI, *Eco*RI, *Xba*I, *Xho*I, and *Hae*III restriction endonucleases (New England Biolabs, Ipswich, MA) at 37°C for 4 h simultaneously. The 6-bp cutter restriction endonucleases *Eco*RI (G'AATTC), *Xba*I

(T'CTAGA), *Bam*HI (G'GATCC), *Xho*I (C'TCGAG) generate 5'-protruding ends allowing differentially fluorescence labeled A, C, G, and T to be incorporated at the 3' ends of fingerprints while the 4-bp cutter *Hae*III cleaves the fragments to small segments making them suitable for analysis using an automated sequencer. All the procedures for restriction digestion reactions and labeling reactions followed the protocols of the SnapShot kit provided by the manufacturer. The labeled BAC fragments were precipitated and analyzed with the ABI GS500-LIZ internal size standard (Applied Biosystems) on ABI 3730XL DNA Analyzer (Applied Biosystems) in W.M. Keck Center at University of Illinois.

Data processing

The fragment sizes in each BAC fingerprint profile were collected by the ABI Data Collection program. The data off the ABI 3730XL Genetic Analyzer were processed by the computer software package GenoProfiler (http://wheat.pw.usda.gov/PhysicalMapping/) and FPminer (http://www.bioinforsoft.com). Briefly, the fragment size calling was conducted using an automatic algorithm in FPminer. Several quality checks were applied to the fingerprints: the empty well was removed; fingerprints with less than 25 fragments or greater than 250 fragments were removed; the background fragments were identified and removed using the FPminer embed algorithm, the off-scale fragments with peak height greater than 6000 were removed. The data were then transferred to Genoprofiler to remove the vector fragments and frequent fragments. Only the fragments between 50 to 500 bp were used for contig assembly in FPC assembly.

BAC contig assembly

The program FPC version 8.5 was used to assemble the BAC fingerprint data to contigs. The size tolerance value was determined by the mean size deviation of the vector fragments and the size standard fragments in GS500-LIZ internal size standard (Applied Biosystems, Foster City, CA). The 250-bp fragment was not used since this fragment migrates abnormally under denaturing conditions (ABI, personal communications). A total of 300 samples were selected randomly for each fragment. The mean size deviations with 95% confidence intervals were computed to determine the tolerance of FPC assembly.

Library screening and assessment of physical map quality

Overgo hybridization was conducted to screen the BAC library to assess the physical map quality. For the explanation of overgo, see Appendix in Chapter VI. Two assessment approaches were conducted. First, the overgo probes designed from genes were hybridized to the high-density filters of a channel catfish BAC library, which were purchased from BACPAC Resources at Children's Hospital of Oakland Research Institute (BACPAC Resources, CHORI, Oakland, CA). All the primer sequences for the probes are listed in Table 6. Second, overgo probes designed from BAC end sequences were used to hybridize to the BAC clones in the selected contigs. DNA (200 ng) of each BAC clone from the same contig was treated with procedures of dot blot analysis (Sambrook et al. 1989) and spotted on a piece of nylon membrane using a pipette and crosslinked to the membrane using UV radiation using a UV Stratalinker 2400 (Stratagene, La Jolla, CA). Overgo hybridization was conducted as previously described in Bao’s paper (Bao et al.

2005). Briefly, overgo primers were selected following a BLAST search against GenBank to screen out repeated sequences and then purchased from Sigma Genosys (Woodlands, TX). Two hundred nanograms of overgo primers each were labeled with 40 µl of a freshly prepared master mix composed of 14.0 mM Tris–HCl (pH 7.5), 5 mM MgCl2, 0.02 mM dGTP, 0.02 mM dTTP, 20 µCi [α-33P]dCTP, 20 µCi [α-33P]dATP (3000 Ci/mmol, Amersham, Piscataway, NJ), and 5 units of Klenow enzyme (Invitrogen). Labeling reactions were carried out at room temperature for 2 h. After removal of unincorporated nucleotides using a Sephadex G50 spin column, probes were denatured at 95 °C for 10 min and added to the hybridization tubes containing high-density BAC filters. Hybridization was performed at 54 °C for 18 h in hybridization solution (50 ml of 1% BSA, 1 mM EDTA at pH 8.0, 7% SDS, 0.5 mM sodium phosphate, pH 7.2). BAC filters were washed with 2× SSC at room temperature for 15 min, and exposed to X-ray film at −80 °C for 2 days.

Table 6. All primers, probes, and their sequences used in this study. Ova and Ovb are a set of two primers for overgo probes.

Probes	Probe target	Overgo primer Ova	Overgo primer Ovb
AU50480	Hepcidin	CTGCTGCAGGTTCTAATAACGGAC	TGAAAACTTGCATGTGGTCCGTTA
AU50493	LEAP-2	AGGAGATCAGAGGTCACTCAAGAG	TGTCATACGGGCCATTCTCTTGAG
AU50531	BPI	TATCAGCCTTCACCCTGAACTCAG	TTGTACACGAATCCGGCTGAGTTC
AU50591	CXCL12	TTGCTGAACCAGCACTTAACCTGC	GAGGCAAGCAAGGTTTGCAGGTTA
AU50592	CXCL14	CAAATGCAGATGCACCAGGAAAGG	GTATCGTATCTTGGGGCCTTTCCT
AU50620	CXCL10	GAGAATCTTCAGAGCATCGAGTGT	CCTCTTGCTCTTGAACACACTCGA
AU50621	CXCL8	CAGTAACTGCCTTCTGCTGCTTTG	AAGGCAAACACTGTGGCAAAGCAG
AU50814	IL-1beta	AATATTCAGTCCACGGAGTTCACC	TGAAAAGCTCCTGGTCGGTGAACT
AU50860	TLR20	TGGGACTGGTGTTCTTCATGCTGG	TGATGGAGCAGCACTACCAGCATG
AU50861	TLR21	GCTTGTTACACTTCGCCTGGACAA	ATCAGACAGAAGGTTGTTGTCCAG
AU51030	Contig121	CAGTATTGGTAGCTAGCCATTTTC	ACGACAGTAGCTTTGTGAAAATGG
AU51056	Contig135	GACTGGCTTTGAAACGTGGGAAGC	GGACTGGCTTTTGTCTGCTTCCCA
AU51063	Contig199	CGCGAGTTTTGTCTTTGAGTCATC	CCAGATCATGCTCATGGATGACTC
AU51079	Contig276	TTTATATTGTAGGTGTTACCTAGG	GTCAACTTCCAGTTGCCCTAGGTA
AU51083	Contig1246	GTGGGACGAACAGATTTTAAGTTT	AACGAACTGCCACCTTAAACTTAA
AU51111	Contig 366	ATTACCTTGTATATTGCAAATGGG	GGAATGACCAAAAAAACCCATTTG
AU51135	Contig958	CCAACCTTGCCTTATGCTTTTTTC	GAAAGAAAGAAAGAAAGAAAAAAG
AU51141	Contig284	GTGTGCCGATCTAAACGTATCAGG	CAGTAATGATGCTCTACCTGATAC
AU51142	Contig284	TTGTTTTCCTTGGCCAACCTGTTC	CTAACAACCAGACATGGAACAGGT
AU51144	Contig579	CAGTGTATTAGGGTGGAAGTGGTG	GTAATGAAGACCCTGCCACCACTT
AU51146	Contig586	ACCACATGAAAAGTGCTCTATAAA	CTACTGATGAACTTCTTTTATAGA
AU51147	Contig586	TCGGTTGGAAACCTGACAAAAATG	AGCATTTCGACCCTCTCATTTTTG
AU51148	Contig586	CTCCATGTAAGTTCAGACACACCG	TGTAAACACCGCAGTCCGGTGTGT
AU51149	Contig586	GAGAAGCACAGTCAATAAACGCTG	GCCATGCCATTATGAGCAGCGTTT
AU51150	Contig673	ACACAAAATTGATTTTCATGCAAG	AGCTCACGAAAGTATGCTTGCATG
AU51151	Contig673	GGCGTGGATCACAGATGCTATATG	CGCGTCGCGTGATATACATATAGC
AU51152	Contig1046	AAAATCATGTGGAAATCAATGATC	AACCATTATGCACATGGATCATTG
AU51153	Contig1046	AACACGATTGAATCATTTCACTTG	GAGAAATCTAGCTGAGCAAGTGAA
AU51154	Contig1046	TTACTAAAACATATATCAATATTC	AATCAGTCGGTCTCAGGAATATTG
AU51164	Contig1558	ACCCAATGTACTGTAAAATTTGTG	TCACTCTTAGACTCTACACAAATT
AU51165	Contig1558	TAGACGAGCTCGTAGTTGAGAGAG	CTCTCAAGAACTTAGTCTCTCTCA
AU51168	Contig1997	TTTGCACCTTGGGGTTGAAGTAGC	CAGCCACAGTCCTGCCGCTACTTC

ACKNOWLEDGEMENTS

This project was supported by a grant from USDA NRI Animal Genome Tools and Resources Program (award #2006-35616-16685). I am grateful for an equipment grant from the National Research Initiative Competitive Grant no. 2005-35206-15274 from the USDA Cooperative State Research, Education, and Extension Service. I thank Dr. Pieter de Jong at the Children's Hospital of the Oakland Research Institute for providing us with the CHORI-212 channel catfish BAC library. I am very grateful to Dr. Ryan Kim and Chris Wright of the University of Illinois for electrophoresis runs of the fingerprints on their ABI PRISM 3730 DNA sequencers.

REFERENCES

Bao, B., E. Peatman, P. Li, C. He, and Z. Liu. 2005. Catfish hepcidin gene is expressed in a wide range of tissues and exhibits tissue-specific upregulation after bacterial infection. *Dev Comp Immunol* **29:** 939-950.

Brunet, F.G., H.R. Crollius, M. Paris, J.M. Aury, P. Gibert, O. Jaillon, V. Laudet, and M. Robinson-Rechavi. 2006. Gene loss and evolutionary rates following whole-genome duplication in teleost fishes. *Mol Biol Evol* **23:** 1808-1816.

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

Fernando, S.C., F.Z. Najar, X. Guo, L. Zhou, Y. Fu, R.D. Geisert, B.A. Roe, and U. Desilva. 2007. Porcine kallikrein gene family: Genomic structure, mapping, and differential expression analysis. *Genomics*.

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

Jaillon, O., J.M. Aury, F. Brunet, J.L. Petit, N. Stange-Thomann, E. Mauceli, L. Bouneau, C. Fischer, C. Ozouf-Costaz, A. Bernot, S. Nicaud, D. Jaffe, S. Fisher, G. Lutfalla, C. Dossat, B. Segurens, C. Dasilva, M. Salanoubat, M. Levy, N. Boudet, S. Castellano, V. Anthouard, C. Jubin, V. Castelli, M. Katinka, B. Vacherie, C. Biemont, Z. Skalli, L. Cattolico, J. Poulain, V. De Berardinis, C. Cruaud, S. Duprat, P. Brottier, J.P. Coutanceau, J. Gouzy, G. Parra, G. Lardier, C. Chapple, K.J. McKernan, P. McEwan, S. Bosak, M. Kellis, J.N. Volff, R. Guigo, M.C. Zody,

J. Mesirov, K. Lindblad-Toh, B. Birren, C. Nusbaum, D. Kahn, M. Robinson-Rechavi, V. Laudet, V. Schachter, F. Quetier, W. Saurin, C. Scarpelli, P. Wincker, E.S. Lander, J. Weissenbach, and H. Roest Crollius. 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* **431:** 946-957.

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

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

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

Katagiri, T., C. Kidd, E. Tomasino, J.T. Davis, C. Wishon, J.E. Stern, K.L. Carleton, A.E. Howe, and T.D. Kocher. 2005. A BAC-based physical map of the Nile tilapia genome. *BMC Genomics* **6:** 89.

Kocabas, A., R. Dunham, and Z.J. Liu. 2004. Alterations in gene expression in the brain of white catfish (*Ameirus catus*) in response to cold acclimation. *Marine Biotechnology* **6:** 431-438.

Kocabas, A.M., P. Li, D. Cao, A. Karsi, C. He, A. Patterson, Z. Ju, R.A. Dunham, and Z. Liu. 2002. Expression profile of the channel catfish spleen: analysis of genes

involved in immune functions. *Mar Biotechnol (NY)* **4:** 526-536.

Leeb, T., C. Vogl, B. Zhu, P.J. de Jong, M.M. Binns, B.P. Chowdhary, M. Scharfe, M. Jarek, G. Nordsiek, F. Schrader, and H. Blocker. 2006. A human-horse comparative map based on equine BAC end sequences. *Genomics* **87:** 772-776.

Li, P., E. Peatman, S. Wang, J. Feng, C. He, P. Baoprasertkul, P. Xu, H. Kucuktas, S. Nandi, B. Somridhivej, and J. Serapion. 2007. Towards the catfish transcriptome: development of molecular tools from 31,215 catfish ESTs. *BMC Genomics* in press.

Li, R.W. and G.C. Waldbieser. 2006. Genomic organisation and expression of the natural killer cell enhancing factor (NKEF) gene in channel catfish, *Ictalurus punctatus (Rafinesque)*. *Fish Shellfish Immunol* **20:** 72-82.

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

Liu, Z., P. Li, and R.A. Dunham. 1998. Characterization of an A/T-rich family of sequences from channel catfish (*Ictalurus punctatus*). *Mol Mar Biol Biotechnol* **7:** 232-239.

Luo, M.C., C. Thomas, F.M. You, J. Hsiao, S. Ouyang, C.R. Buell, M. Malandro, P.E. McGuire, O.D. Anderson, and J. Dvorak. 2003. High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis. *Genomics* **82:** 378-389.

Marra, M.A., T.A. Kucaba, N.L. Dietrich, E.D. Green, B. Brownstein, R.K. Wilson, K.M. McDonald, L.W. Hillier, J.D. McPherson, and R.H. Waterston. 1997. High

throughput fingerprint analysis of large-insert clones. *Genome Res* **7:** 1072-1084.

McPherson, J.D. M. Marra L. Hillier R.H. Waterston A. Chinwalla J. Wallis M. Sekhon K. Wylie E.R. Mardis R.K. Wilson R. Fulton T.A. Kucaba C. Wagner-McPherson W.B. Barbazuk S.G. Gregory S.J. Humphray L. French R.S. Evans G. Bethel A. Whittaker J.L. Holden O.T. McCann A. Dunham C. Soderlund C.E. Scott D.R. Bentley G. Schuler H.C. Chen W. Jang E.D. Green J.R. Idol V.V. Maduro K.T. Montgomery E. Lee A. Miller S. Emerling Kucherlapati R. Gibbs S. Scherer J.H. Gorrell E. Sodergren K. Clerc-Blankenburg P. Tabor S. Naylor D. Garcia P.J. de Jong J.J. Catanese N. Nowak K. Osoegawa S. Qin L. Rowen A. Madan M. Dors L. Hood B. Trask C. Friedman H. Massa V.G. Cheung I.R. Kirsch T. Reid R. Yonescu J. Weissenbach T. Bruls R. Heilig E. Branscomb A. Olsen N. Doggett J.F. Cheng T. Hawkins R.M. Myers J. Shang L. Ramirez J. Schmutz O. Velasquez K. Dixon N.E. Stone D.R. Cox D. Haussler W.J. Kent T. Furey S. Rogic S. Kennedy S. Jones A. Rosenthal G. Wen M. Schilhabel G. Gloeckner G. Nyakatura R. Siebert B. Schlegelberger J. Korenberg X.N. Chen A. Fujiyama M. Hattori A. Toyoda T. Yada H.S. Park Y. Sakaki N. Shimizu S. Asakawa K. Kawasaki T. Sasaki A. Shintani A. Shimizu K. Shibuya J. Kudoh S. Minoshima J. Ramser P. Seranski C. Hoff A. Poustka R. Reinhardt and H. Lehrach. 2001. A physical map of the human genome. *Nature* **409:** 934-941.

Nandi, S., E. Peatman, P. Xu, S. Wang, P. Li, and Z. Liu. 2006. Repeat structure of the catfish genome: a genomic and transcriptomic assessment of Tc1-like transposon elements in channel catfish (*Ictalurus punctatus*). *Genetica*.in press

Nelson, W.M., A.K. Bharti, E. Butler, F. Wei, G. Fuks, H. Kim, R.A. Wing, J. Messing,

and C. Soderlund. 2005. Whole-genome validation of high-information-content fingerprinting. *Plant Physiol* **139:** 27-38.

Ng, S.H., C.G. Artieri, I.E. Bosdet, R. Chiu, R.G. Danzmann, W.S. Davidson, M.M. Ferguson, C.D. Fjell, B. Hoyheim, S.J. Jones, P.J. de Jong, B.F. Koop, M.I. Krzywinski, K. Lubieniecki, M.A. Marra, L.A. Mitchell, C. Mathewson, K. Osoegawa, S.E. Parisotto, R.B. Phillips, M.L. Rise, K.R. von Schalburg, J.E. Schein, H. Shin, A. Siddiqui, J. Thorsen, N. Wye, G. Yang, and B. Zhu. 2005. A physical map of the genome of Atlantic salmon, *Salmo salar*. *Genomics* **86:** 396-404.

Peatman, E., B. Bao, X. Peng, P. Baoprasertkul, Y. Brady, and Z. Liu. 2006. Catfish CC chemokines: genomic clustering, duplications, and expression after bacterial infection with *Edwardsiella ictaluri*. *Mol Genet Genomics* **275:** 297-309.

Peatman, E., P. Baoprasertkul, J. Terhune, P. Xu, S. Nandi, H. Kucuktas, P. Li, G. Wang, B. Somridhivej, and A. Dunham. 2007. Expression analysis of the acute phase response in channel catfish (*Ictalurus punctatus*) after infection with a Gram negative bacterium. *Dev Comp Immunol* **in press**.

Peatman, E. and Z. Liu. 2006. CC chemokines in zebrafish: evidence for extensive intrachromosomal gene duplications. *Genomics* **88:** 381-385.

Peatman, E. and Z. Liu. 2007. Evolution of CC chemokine in teleost fish: a case study in gene duplication and implications for immune diversity. *Immunogenetics* **in press**.

Quiniou, S.M., T. Katagiri, N.W. Miller, M. Wilson, W.R. Wolters, and G.C. Waldbieser. 2003. Construction and characterization of a BAC library from a gynogenetic channel catfish *Ictalurus punctatus*. *Genet Sel Evol* **35:** 673-683.

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

Romanov, M.N., M. Koriabine, M. Nefedov, P.J. de Jong, and O.A. Ryder. 2006. Construction of a California condor BAC library and first-generation chicken-condor comparative physical map as an endangered species conservation genomics resource. *Genomics* **88:** 711-718.

Sambrook, J., E.F. Frisch, and T. Maniatis. 1989. *Molecular cloning: a laboratory manual*. Cold Spring Harbor Laboratory Press.

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

Tiersch, T.R. and C.A. Goudie. 1993. Inheritance and variation of genome size in half-sib families of hybrid *ictalurid* catfishes. *J. Hered.* **84:** 122-125.

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

Wang, S., P. Xu, J. Thorsen, B. Zhu, P. de Jong, G. Waldbieser, and Z. Liu. 2007. Characterization of a BAC library from channel catfish Ictalurus punctatus: indications of high rates of evolution among teleost genomes. *Marine Biotechnology* in press.

Woods, I.G., C. Wilson, B. Friedlander, P. Chang, D.K. Reyes, R. Nix, P.D. Kelly, F. Chu, J.H. Postlethwait, and W.S. Talbot. 2005. The zebrafish gene map defines ancestral vertebrate chromosomes. *Genome Res* **15:** 1307-1314.

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

V. CONCLUSION

My dissertation research focused on physical characterization of the channel catfish genome using BAC end sequencing and BAC-based fingerprinting. The major results of my project are described in the following:

In the BAC end sequencing project:

1. A total of 12,672 BAC clones from CHORI212 channel catfish BAC library, representing 1.84X genome coverage, were sequenced from both ends of the inserts using universal primers T7 and SP6.

2. A total of 20,366 clean BAC end sequences were generated with an average size of 557 bp, in which 17,478 BAC end sequences were paired on BAC. The total base pairs generated in this project are 11,414,601 bp, equaling 1.27% of channel catfish genome on sequence length.

3. The overall status of repetitive elements in the catfish genome was assessed by repeatmasking analysis of BAC end sequences. The most abundant type of repeat in the catfish genome is DNA transposon, the vast majority consisting of *Tc1/mariner* transposon related sequences accounted for 4.12% of the base pairs of the BES. Retroelements were the second largest fraction of repetitive elements accounting for 3.13% of the base pairs of the BES. Simple repeats such as microsatellites accounted for

2.58% of the catfish BES base pairs with CA, AT, and CT type repeats accounting for 68% of all microsatellite types.

4. A total of 3,748 BES were found to contain one or more stretches of microsatellite sequences. Of these, 2,365 (63%) had sufficient flanking sequences on both sides for marker development.

5. The potential novel repeats in catfish genome were identified using BAC end sequences self-Blast analysis.

6. A total of 1,877 unique genes were virtually mapped to 2,351 BAC end sequences using BLASTX search against Non-redundant database.

7. Catfish BACs were anchored to the zebrafish and *Tetraodon* genome sequences by BLASTN search, revealing 16% and 8.2% significant hits ($E < e^{-5}$), respectively.

8. A total of 141 mate-paired BES were found to include genes on both sides of the BAC insert, from which 23 conserved syntenies were identified (~16.3%) among catfish, zebrafish, and *Tetraodon* genomes. Three syntenies were successfully extended by comparative analysis and sequencing within catfish BAC clones.

In the physical mapping project:

1. The fingerprint data were generated from ABI 3730 XL automatic genetic analyzer. A total of 40,416 BAC clones were fingerprinted using high information content fingerprinting (HICF) techniques and 34,580 of them (84.3% success rate, 5.6 genome coverage equivalents) were validated fingerprints with an average of 95.2 fragments.

2. The first version of channel catfish physical map was constructed: a total of 3,307 contigs containing 30,582 BAC clones were generated from FPC assembly using the tolerance of 0.4 bp and the cutoff value of $1e^{-20}$. The combined contig size for all contigs

represents 965,279 kb consensus length, approximately 1X genome size of the channel catfish.

3. The vast majority of contigs (84.4%) were free of Q-clones. A total of 1,494 questionable clones distributed in 517 Q-contigs. Most of the Q-clones were only involved in a small number of contigs.

4. The reliability of the contig assembly was validated by overgo hybridizations using the overgo probes from BAC end sequences or known genes. The overgo hybridizations also confirmed the cutoff value $1e^{-20}$ used in the FPC assembly was proper.

It is the first time that the large scale BAC end sequencing and BES analysis have been conducted and reported in the aquaculture species, which revealed extreme rich genomic information from the catfish genome and provided the solid base for the integration of linkage map and physical map in catfish. The physical map constructed based on channel catfish CHORI-212 BAC library is the first physical map for normal diploid catfish and is necessary for the future whole genome sequencing and importance traits associate gene identifications in catfish.

The future work should focus on second generation physical map and physical map and genetic map integration. Current 3,307 BAC contigs need to be merged to a smaller number of contigs. There are several ways working toward this goal. Obviously, increasing the number of fingerprinted BAC clones would be the most straightforward method. The new fingerprints will merge separated contigs and singletons until the contig number reach a saturated plateau level. Contig walking is another method for contig merge and extension. Probes are designed from the BAC end sequences on the end of the "seed" contig. Hybridization using the probes will identify the potential BAC clones or

contigs overlapping with the "seed" contig. Thus, the "seed" contig is extended or merged with other contigs. Alternatively, the linkage map can also be used as a reference for contig merging once the physical map and linkage map are integrated, even if a gap exists. Multiple contigs can be assigned or anchored linearly to one linkage group. Some contigs may share same genetic markers in the same neighborhood. These contigs can be merged in this case. If the contigs do not share the same genetic markers, their linear relationship and their distance on the linkage group still can be estimated, then use "contig walking" method to merge them. The more anchor points we build between the linkage map and the physical map, the more contigs can be merged.

Genetic markers, especially microsatellite markers, can be developed from catfish BAC end sequences. Microsatellites are abundant in the BAC end sequences, which had been described in the chapter III. Once the BAC-derived markers are mapped in the genetic maps using catfish resource families, they build the anchor points between genetic map and BAC-based physical map. Large number of the BAC-derived genetic markers is necessary for the integration of genetic map and physical map.

Publications of my dissertation research:

Xu, P., Wang, S., Liu, L., Thorsen, J., Liu, Z. A BAC-based physical map of the channel catfish genome. *Genomics*. 2007, in press.

Xu P., Wang S., Liu L., Peatman E., Somridhivej B., Thimmapuram J., Gong G., Liu Z. Channel catfish BAC-end sequences for marker development and assessment of syntenic conservation with other fish species. *Animal Genetics*. 2006, 37: 321-326

APPENDIX

What is an overgo?

The following explanation is cited from http://www.maizegdb.org/overgo.php with minor changes.

An overgo is a primer construct designed to pack a lot of label into a short sequence, so that it will serve as a sensitive probe for similar short sequences in genomic DNA. It is based on two partially OVERlapping oliGOnucleotide primers, each of which primes the other in a labeling reaction. For example, let's say you had two primers, one with the sequence TGGATCCTGTGCCTTTTGACATCG and another with the sequence GTCACCAAATCCCTCTCGATGTCA. The last eight characters of the first sequence are exactly the same as the first eight characters of the anti-sense sequence of the second primer. The DNA polymerase Klenow fragment and radioisotope labeled dATP and dCTP with unlabeled dGTP and dTTP are used in the probe labeling reaction to generate one radioactive labeled DNA probe for overgo hybridization: TGGATCCTGTGCCTTTTGACATCGAGAGGGATTTGGTGAC. Overgo probes can be both highly specific and highly radioactive, making them useful in probing BAC clones.

Two major advantages of overgos over standard "plasmid" probes are:

1. Hybridization can be carried out under standardized conditions due to standardized probe length and controlled GC content.

2. Overgo probe use makes it possible to select for specifically low or single copy portions of a longer sequence that may contain repetitive parts (which would otherwise result in high levels of background using other probe/hybridization techniques).