A Chromosome-level genome assembly of the striped catfish (Pangasianodon hypophthalmus) reveals molecular mechanisms for its high-fat trait

Zijian Gao[1], Xinxin You[2], Xinhui Zhang[3], Jieming Chen[2], Xu Tengfei[2], Yu Huang[2], Xueqiang Lin[2], Junmin Xu[2], Chao Bian[4], and Qiong Shi[2]

[1]University of the Chinese Academy of Sciences
[2]Affiliation not available
[3]BGI Academy of Sciences, BGI Marine
[4]BGI

March 30, 2022

Abstract

Striped catfish (Pangasianodon hypophthalmus), belonging to the Pangasiidae family, has become an economically important fish with wide cultivation in Southeast Asia. Owing to the high-fat trait, it is always considered as an oily fish. In our present study, a high-quality genome assembly of the striped catfish was generated by integration of short reads from an Illumina Hiseq Xten platform, long reads from a Nanopore platform, and Hi-C sequencing data. This assembled genome is 731.7 Mb in length, with a scaffold N50 of 29.5 Mb and anchoring on 30 chromosomes. A total 18,895 protein-coding genes were predicted, among them 98.46% were functionally annotated. Interestingly, we identified a tandem triplication of fatty acid binding protein 1 gene (fabp1; thereby named as fabp1-1, fabp1-2 and fabp1-3 respectively), which may be critical for molecular regulation of the high-fat trait in the striped catfish. Compared with Fabp1-1 (similar to the conserved Fabp1 in various vertebrate species), the R126T mutation may potentially affect the fatty acid binding capacity of the Fabp1 isotypes 2 and 3. In summary, we report a high-quality chromosome-level genome assembly of the striped catfish, which provides a valuable genetic resource for biomedical studies on the high-fat trait, and laying a foundation for practical aquaculture and molecular breeding of this international teleost species.

Introduction

Striped catfish (*Pangasianodon hypophthalmus* Sauvage, 1878) is a typical catfish in the Pangasiidae family. It is a migratory and benthic species with a natural origin in the Mekong River and the Chao Phraya River basins (Ali et al, 2013). Due to fast growth, high disease resistance and good adaptability to high density, it is suitable for international intensive production, although it is now predominantly cultured in Vietnam. Influenced by trade and market requirements, striped catfish was introduced to several other Asian countries, such as Singapore, Philippines, Malaysia, India and China (De Silva et al., 2006; Singh & Lakra, 2012). In view of burgeoning trade, the introduction of striped catfish may boost up the aquaculture production in some of these countries (De Silva et al., 2006; De Silva & Soto, 2009). Currently, this catfish species is widely cultured in many Asian countries as an important food source with a great economic significance.

In a previous study, a primary assembly of striped catfish, based on only Illumina short reads, was reported (Kim et al., 2018). In recent years, combination of short reads, long reads and high-resolution chromosome conformation capture (HIC) techniques has been widely employed to improve the quality of many sequenced

genomes (You et al., 2020). In this study, we applied this integrated strategy to obtain a chromosome-level genome assembly for the striped catfish with much higher integrity and continuity values.

In addition to fish fillet, fish oil is also one of the important value-added products with unsaturated fatty acids. Striped catfish has been considered as an oily fish along with other freshwater fishes from the Mekong sub-region, such as por fish (*Pangasius bocourti*), Mong fish (*Pangasius conchophilus*), and Mekong giant catfish (*Pangasianodon gigas* ; Hemung et al., 2010). The crude fat from head and flab parts of striped catfish sums up to be 58.83%, and unsaturated fatty acids in the crude fat account for 60.28% (Hemung et al., 2010). Striped catfish therefore may be a considerable source of unsaturated fatty acids for human consumption. However, for fish fillet, the high fat content may have a negative impact on the economic value of striped fishes (Van Sang et al., 2012). Hence, body fat of striped fish has become an important trait for any practical breeding program.

By in-depth genomic analysis, we observed a tandem duplication of fatty acid binding protein 1 gene (*fabp1*) in the striped catfish genome. In fact, *fabp1* is known to be critical for fatty acid uptake and intracellular transport, and it also plays an important role in regulation of lipid metabolism and cellular signaling pathways (Furuhashi & Hotamisligil, 2008). Since *fabp1* was proved to be relevant to fatty acid storage, the tandem duplication of *fabp1* may be related to the high-fat trait of striped catfish. It may provide a new perspective for molecular breeding of striped catfish with an appropriate body fat content.

Methods

Sample collection, library construction and sequencing

A mature female striped catfish, cultivated in Wenchang City, Hainan Province, China, was selected for whole genome sequencing. Genomic DNAs were extracted from pooled muscle tissues using a genomic DNA isolation kit (Qiagen, Hilden, Germany). For the routine genome shotgun sequencing, a 500-bp insert library was constructed, followed by paired-end sequencing on an Illumina Hiseq Xten platform (Illumina Inc., San Diego, CA, USA). For the long-read sequencing, a Nanopore library with an insert-size of 20 kb was prepared using the extracted genomic DNAs, and subsequently sequenced on a Nanopore platform (Nanopore Technologies, Oxford, Head Office, UK). The Hi-C library, constructed with the same DNA pool, was sequenced on the same Illumina Hiseq Xten platform.

In order to assist genome annotation, total RNAs were extracted from various tissues, including muscle, gill, skin, liver and kidney. Individual cDNA library was constructed for transcriptome sequencing (RNA-seq) on the same Illumina Hiseq Xten platform.

This project was reviewed and approved by the Institutional Review Board on Bioethics and Biosafety of BGI, China (No. FT 18134).

Genome size estimation based on the routine k-mer method

Short reads generated from the whole genome shotgun sequencing were filtered using SOAPnuke v1.0 (Chen et al., 2018). Clean short reads were then used for genome size estimation based on a 17-mer distribution according to the following formula: Genome Size=Kmer_num/Kmer_depth, where kmer_num is the total number of k-mers and Kmer_depth represents the sequencing depth of each unique k-mer with the highest frequency.

Genome assembly

Before assembly, LoRDEC (Salmela & Rivals, 2014) was applied for calibration of the long sequencing reads along with the clean short reads. Subsequently, minimap2 (Li, 2018) was employed to assemble the corrected long reads to contigs. Genome polishing was preformed two rounds with different strategies. In brief, for the first round, the contigs were polished via Racon v1.3.1 (Vaser et al., 2017) based on uncorrected Nanopore long reads; for the second round, the clean short reads were used to polish the contigs with pilon (Walker et al., 2014). After heterozygosity reducing with Redundans (Pryszcz & Gabaldón, 2016), we finally obtained a draft genome for the striped catfish.

Chromosome construction and genome assessment

Pseudo-chromosomes were constructed on the basis of the assembled draft genome and the reads generated from the Hi-C library. After quality control of Hi-C data, clean reads were mapped to the assembled draft genome via Bowtie2 (Langdon, 2015) with default parameters. Mapped reads were further clustered using Juicer (Durand et al., 2016), followed by ordering and orientation performed with 3d-dna (Dudchenko et al., 2017). Finally, the assembled genome was anchored to 30 chromosomes.

To assess the quality of our final chromosome-level genome assembly, contig N50 and scaffold N50 values were calculated for comparison with those of other Siluriformes species. With the popular actinopterygii_-odb9 database, Benchmarking Universal Single-Copy Orthologs (BUSCO) was employed to evaluate the completeness of the assembled striped catfish genome (Simão et al., 2015).

Genome annotations and gene predictions

Prediction of repeat elements was based on *de novo* and homology methods. RepeatModeler (Smit et al., 2014) and LTR-FINDER (Xu & Wang, 2007) were primarily used for the *de novo* prediction, generating a repeat library respectively. Both libraries were combined and then aligned to the assembled genome with RepeatMasker (Tarailo-Graovac & Chen, 2009). Meanwhile, the homology prediction was performed using RepeatMasker and RepeatProteinMask (Tarailo-Graovac & Chen, 2009), based on the known repeat library (Repbase; Jurka et al., 2005). In addition, tandem repeats were detected with Tandem Repeat Finder (Benson, 1999). Finally, by integrating these predicted data, we obtained nonredundant repeat elements.

We employed two approaches, homology and transcriptome-based, to predict protein-coding genes. For the homology-based prediction, protein sequences of 11 representative vertebrate species, including zebrafish (*Danio rerio*), channel catfish (*Ictalurus punctatus*), Atlantic cod (*Gadus morhua*), three-spined stickleback (*Gasterosteus aculeatus*), Australian ghost shark (*Callorhinchus milii*), spotted gar (*Lepisosteus oculatus*), Nile tilapia (*Oreochromis niloticus*), medaka (*Oryzias latipes*), Japanese pufferfish (*Takifugu rubripes*), green spotted puffer (*Tetraodon nigroviridis*) and human (*Homo sapiens*), were downloaded from Ensembl (Flicek et al., 2013) for mapping to our assembled genome with TBLASTn (Gertz et al., 2006). Subsequently, GeneWise (Birney et al., 2004) was used to predict gene structures of the achieved alignments. For the transcriptome-based prediction, we applied Cufflinks (Ghosh & Chan, 2016) and Hisat (Kim et al., 2015) to predict gene structures with generated transcriptome data. Finally, these predicted results were integrated using MAKER (Campbell et al., 2014) to obtain a final consistent gene set.

To perform functional annotation, we employed BLASTp (Altschul et al., 1990) to align the predicted protein sequences against four public databases, including SwissProt (Boeckmann et al., 2003), TrEMBL (Boeckmann et al., 2003), KEGG (Kanehisa & Goto, 2000) and InterPro (Hunter et al., 2009). These results were retrieved using Gene Ontology (GO; Consortium, 2004) terms.

Gene family analysis

To construct gene families form protein-coding genes of striped catfish, We downloaded coding sequences (CDS) of 7 representative vertebrate species, including human, mouse (*Mus musculus*), zebrafish, yellow catfish (*Tachysurus fulvidraco*), channel catfish, black bullhead (*Ameiurus melas*), giant devil catfish (*Bagarius yarrelli*) from Genbank (Benson et al., 2010). After multiple sequence alignment with predicted CDS of striped catfish and other species using BLASTp (Altschul et al., 1990) (e-value [?] 1e-5), gene families were clustered with OrthMCL (Li et al., 2003). In order to reveal the phylogenetic position of striped catfish, we employed ClustalW (Thompson et al., 2003) to align the CDS sequences of single-copy ortholog gene families.

After obtaining the conserved regions via Gblocks (Castresana, 2002), the aligned CDS of all single-copy genes were connected as a supergene. With phase1 sites extracted from the supergene, a phylogenetic tree was constructed using PhyML with the maximum likelihood method (Guindon et al., 2009). Subsequently, the MCMCTREE model in PAML (Yang, 1997) was employed to estimate the divergence times, with assistance of fossil calibration from TIMETREE (http://www.timetree.org/). Molecular clocks include 85-97 million

years ago (Mya) between human and mouse, as well as 130-174 Mya between zebrafish and stripped catfish (*P. hypophthalmus*). Moreover, via CAFE (De Bie et al., 2006), we identified expanded and contracted gene families based on clustered gene families and the achieved phylogenetic tree.

Results

Summary of the Genomeassembly and annotation data

Our genome survey was based on a total of 44.23-Gb short reads generated from the Illumina platform. The peak depth was determined at 54, and the total 17-mer number was 42,252,144,258 (see Fig. 1). Therefore, the genome size of striped catfish was calculated to be about 782 Mb. In addition, the heterozygosity rate and repeat rate were estimated to be 0.26% and 43.39%, respectively (Fig. 1).

A total of 63.07-Gb long reads generated from the Nanopore platform were assembled into a 742.6-Mb draft genome, with a contig N50 of 3.5 Mb and the GC content of 38.89% (Table 1). Additionally, A total of 682 million raw reads with total length of about 90 Gb generated from the Hi-C library were applied to identify contacts among contigs, of which 304 million pairs of reads (89.3% of raw reads) were mapped to the assembled genome. The mapped reads were then used to assemble contigs into scaffolds, resulting in a 731.7-Mb genome assembly with 30 chromosomes and a scaffold N50 of 29.5 Mb (Fig. 2a, Table1). The chromosome-level genome assembly of striped catfish was mapped as a Circos atlas, with genome-wide distributions of gene density, genomic GC content, and the internal syntenic blocks (Fig. 2b).

By repeat annotation, we predicted a total of 274-Mb repeat sequences, covering 36.9% of the total assembled genome (Table S1). Among them, 172.8 Mb of DNA repeat elements, 63.1 Mb of long interspersed nuclear elements (LINE), 48.8 Mb of long terminal repeats (LTR) and 5.4 Mb of short interspersed nuclear elements (SINE) were identified (Table S2). After integrating the results from both homology and transcriptome-based annotations, we predicted 18,895 protein-coding genes in the stripped catfish genome (Table S3). Finally, using four public databases, we predicted that 98.46% of the predicted protein-coding genes (a total of 18,604) are functionally annotated (Table S4).

Genome quality evaluation

The final chromosome-level genome assembly of striped catfish, 731.7 Mb in length, accounted for 98.5% of the assembled genome. To evaluate the genome quality, we compared our genome assembly to those previously published genomes of striped catfish (GenBank assembly accession: GCA_003671635.1; Kim et al., 2018) and other Siluriformes species. On account of the integrative strategy using shotgun sequencing and long-read Nanopore sequencing as well as the Hi-C data, our chromosome-level genome assembly of striped catfish with higher values of contig and scaffold N50 than most of previously published genome assemblies. Especially, among all published Siluriformes genomes, both contig and scaffold N50 values of the assembled genome for striped catfish in the present study reached a comparatively higher level (Table 2), indicating a considerable continuity of our striped catfish genome assembly. Subsequently, BUSCO (Simao et al., 2015) was used to estimate the completeness of our assembled genome, with the popular actinopterygii_-odb9 database. Among the totally searched 4,584 BUSCO groups, 4,279 (93.3%) BUSCO core genes were completely identified (Table S5).

Gene family and gene clustering

These CDS, predicted from assembled genomes of striped catfish and other 7 species, were applied for gene families clustering. Eventually, protein-coding genes of striped catfish were clustered into 13,391 gene families (containing 18,220 genes), among them 1,581 were identified as single-copy orthologous gene families (Table S6). With these single-copy orthologous gene families, we constructed a phylogenetic tree based on the maximum likelihood method, predicting that the divergence of striped catfish from others occurred 54.9 Mya (Fig. 3).

In addition, we found 414 expanded gene families and 4,212 contracted gene families in striped catfish (Fig. 3). By KEGG enrichment analysis, those genes in the expanded gene families were clustered into several

critical metabolic pathways, indicating a greater degree of expansion in terms of association with calcium signaling pathway, salivary secretion, apelin signaling pathway and oxytocin signaling pathway (Table S7). Subsequently, we examined the expanded genes with a relation to lipid metabolism, in which we observed an interesting expansion of fatty acid binding protein1 gene (*fabp1*) in the striped catfish genome.

Tandem triplication of the *fabp1* gene

Fatty acid binding proteins (Fabps) are important members of the intracellular lipid binding protein family (iLBP; Bass, 1988). Fabps can reversibly bind intracellular hydrophobic ligands (including the long-chain fatty acids, eicosanoids, bile salts and peroxisome proliferators) and participate in their transfer processes (Storch & Thumser, 2000). They are distributed in many tissues, such as intestine, liver, and heart muscle. From the gene family analysis, we observed a tandem triplication event of the*fabp1* gene in the striped catfish genome. Three copies were located on the chromosome 10 (Fig. 4), and thereby named as*fabp1-1* , *-2* and-*3* . The mRNA transcriptions of *fabp1-1* and *-3* were detectable at a high level by RNA-Seq (Table 3). The deduced protein sequences from *fabp1-2* and *-3* genes are much similar, but there are many mutation variances in the Fabp1-1 (Fig. 4).

In order to examine the phylogenetic relationships of these *fabp1*genes in various teleost species, we constructed a phylogenetic tree using PhyML (the left section in Fig. 5). A synteny region of 9 Siluriformes species for the *fabp1* gene was identified, in which we observed the remarkable expansion with 3 copies of *fabp1* in the stripped catfish (top in the right section of Fig. 5). Compared with other Siluriformes species, mutations were identified in the Fabp1 isotypes 2 and 3 of striped catfish, such as M117T, F124L, and R126T (Fig. 4).

Potential functional effects of these amino acid substitutions were evaluated by using PolyPhen-2 (Adzhubei et al., 2013) and PROVEAN (Choi & Chan, 2015; Table 4). Interestingly, the R126T in Fabp1-2 and -3 was predicted as "damaging" by both HumDiv and HumVar methods in PolyPhen-2 and "deleterious" by PROVEAN, suggesting that this mutation may affect both proteins' functions. Based on homology, the Arg126 of Fabp1 in striped catfish corresponds to the Arg122 of human Fabp1, which has been proved to act as two fatty acid binding sites in relevant studies (Thompson et al., 1997). We therefore propose that the R126T mutation may potentially affect the fatty acid binding capacity of the Fabp1 isotypes 2 and 3.

Peroxisome proliferator activated receptors (PPARs), classified into three subtypes (α, β or δ, and γ), are nuclear receptors with functional regulation of reproduction, development, and metabolism (Braissant et al., 1996). They are proved to participate in the regulation process of several genes including *fabp1* by heterodimerizing with the retinoid X receptor (RXR) and binding to a PPAR response element (PPRE) in the promoters (Hsu et al., 1998). The consensus sequence for the vertebrate PPREs is defined as 5'-CAAAACAGGTCANAGGTCA-3'. Based on various isotypes of PPAR, including PPARα and PPARγ, it seems that PPRE binds to different PPARs (Juge-Aubry et al., 1997). A PPRE with high sequence identity in the 5' flanking region (5'FR; 5'-CAAAAC-3') shows a higher binding activity with PPARα compared to PPARγ. On the contrary, a PPRE with low sequence identity in the 5'FR and high sequence identity in the direct repeat element (DR1; 5'-AGGTCANAGGTCA-3') may exhibit PPARγ-selective (Juge-Aubry et al., 1997).

To find potential PPREs in the *fabp1* promoter of striped catfish, we first extracted sequences of 3,000 bp from 5' upstream of the transcription start sites (TSS) for the 3 isotypes. JASPAR 2020 (Fornes et al., 2020) was then applied for the prediction using two PPREs matrixes (PPARα-selective, Matrix ID: MA1148.1; PPARγ-selective, Matrix ID: MA0065.1), with the relative profile score threshold of 70%. After extraction of putative PPREs located in the sense strand, they were aligned to the consensus sequence of vertebrate PPRE using BLASTn (Altschul et al., 1990). Subsequently, PPREs with 7 or more consecutive nucleotide matches were selected, and then mapped to the promoter sequences (Fig. 6). For the 5 predicted PPREs, those regions highly homologous to the DR1 of the consensus PPRE were identified, but no region with high sequence identity to the 5'FR of the consensus PPRE was detected. It seems that all these putative PPREs within the *fabp1* promoters are PPARγ-selective in striped catfish.

Discussion

Belonging to the fatty acid binding protein gene family, *fabp1* mainly expressed in liver and intestine with critical functions in intracellular transport and fatty acid uptake (Wang et al., 2015). In intestine, Fabp1 binds to endoplasmic reticulum, participates in budding of the pre-chylomicron, and transfers the pre-chylomicron transport vesicle to Golgi (Cifarelli & Abumrad, 2011). Revolving in the fatty acid dependent activation of PPAR, Fabp1 also plays an important role in regulation of lipid metabolism and cellular signaling pathways (Georgiadi & Kersten, 2012). In addition, Fabp1 is proved to be related to obesity (Atshaves et al., 2010), and overexpression of *fabp1* may cause diseases, such as nonalcoholic fatty liver disease (NAFLD), that are relevant to excessive accumulation of fat (Pi et al., 2019). In conclusion, the tandem triplication of *fabp1* gene in the striped catfish genome may make a great difference on fatty acid trafficking and storage in this fish species.

PPAR signaling pathway is an important regulatory mechanism of *fabp1* . With different PPAR-selectivity, various promoter regulatory elements (PPRE) may lead to differential control of duplicated *fabp1* genes (Laprairie et al., 2016; Taylor & Raes, 2004). Compared to the fact that PPARβ/δ is ubiquitously expressed, the expression of PPARα and PPARγ varies in different tissues. PPARα is highly expressed in the liver (Kim et al., 2015); PPARγ is divided into two isoforms (PPARγ1 and PPARγ2), in which PPARγ1 is expressed in many tissues, whereas PPARγ2 is exclusively expressed in adipose tissue (Desvergne & Wahli, 1999). In addition, PPARα and PPARγ were proved to selectively bind to the PPREs in the liver and the adipose tissue, respectively (Shimizu et al., 2004). A recent study (Venkatachalam et al., 2017) reports that retention of *iLBP* genes in the zebrafish genome may be caused by subfunctionalization of PPREs in the *iLBP* promoters. In our present study, we detected PPAR-selectivity for the PPREs of *fabp1-1* , *-2* and *-3* in the striped catfish genome by silico analyses, and all putative PPREs were predicted to be PPARγ-selective, suggesting similar regulation of the three *fabp1* isotypes based on the putative PPAR regulatory mechanism. In addition, subfunctionalization of PPRE was not detected in the promoters of various *fabp1* isotypes, implying that there may be some other factors for the retention of *fabp1* genes in the striped catfish genome.

According to the duplication-degeneration-complementation (DDC) model, there are three possible consequences of duplicated genes. First, with the accumulation of mutations, one or more copies of the duplicated genes will be lost or become non-functional (Force et al., 1999). Second, some of the duplicates may acquire a novel function (Lynch & Conery, 2000). Third, the subdivision of functions between the duplicated genes will occur (Taylor & Raes, 2004). From our RNA-Seq data (Table 3), the mRNA levels of *fabp1-1* and *-3* were detectable while *fabp1-2* was not. It is possible that *fabp1-2* lost the original function, although this proposal is still waiting for more investigations. Although both *fabp1-1* and *-3* were predicted to be PPARγ-selective, the R126T mutation of Fabp1-3 may potentially affect the fatty acid binding capacity of the protein, suggest that subfunctionalization may be differential between *fabp1-1* and *-3* .

With great economic significance, striped catfish has become an important aquaculture species for worldwide consumption. In the process of breeding and trading, the high-fat trait of striped catfish has both advantages and disadvantages. On one hand, the good ability of fat storage may enhance adaptability of striped catfish to a harsh condition with insufficient food sources, and these striped catfishes with a high fat content are good resources for fish-oil extraction (De Silva & Soto, 2009). On the other hand, during the preparation of fish products for current consumer markets, visceral mass is usually discarded as waste instead of being utilized, once they are commonly rich in fat (De Silva et al., 2006). The increasement of fat content in striped catfish fillet, therefore, may affect its economic value (Islami et al., 2014). Our detailed investigation on *fabp1* genes may support further molecular breeding to select individuals with low fat content.

Fatty-acid trafficking and storage in cells are complicated and dynamic, with a combined regulation by multiple genes. Meanwhiles, the fatty acid storage is also influenced by environment, especially temperature can affect lipid content and fatty acid composition of fish oil (Hemung et al., 2010). Since the high-fat trait of striped catfish is related to various factors, in-depth researches are necessary.

Conclusions

In summary, we constructed the high-quality chromosome-level genome assembly for striped catfish. The phylogenetic position of striped catfish was revealed by our phylogenomic analysis. Through a genome-wide study, we observed the tandem triplication of *fabp1* gene on the chromosome 10, which may contribute to the high-fat trait in striped catfish. Comparing the deduced protein sequences of the three *fabp1* genes in striped catfish, we proposed the potential functional effects of several mutations (especially the R126T) and predicted the PPAR-selectivity of PPREs. This high-quality genome assembly will be an improved genetic resource for further biomedical studies of striped catfish. It will also be a good reference for genetic improvement of this economically important fish. Moreover, the detailed investigations on *fabp1* genes in striped catfish may provide new perspectives on molecular breeding for screening individuals with many special traits.

Acknowledgements

This study is supported by Natural Science Foundation for Fundamental Research in Shenzhen (No. JCYJ20190812105801661), Shenzhen Science and Technology Innovation Program for International Cooperation (no. GJHZ20190819152407214), and Grant Plan for Demonstration City Project for Marine Economic Development in Shenzhen (No. 86).

Reference

Adzhubei, I., Jordan, D. M., & Sunyaev, S. R. (2013). Predicting functional effect of human missense mutations using PolyPhen-2. *Current Protocols in Human Genetics* , *76* (1), 7–20.

Ali, H., Haque, M. M., & Belton, B. (2013). Striped catfish (Pangasianodon hypophthalmus, Sauvage, 1878) aquaculture in Bangladesh: an overview. *Aquaculture Research* , *44* (6), 950–965.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* , *215* (3), 403–410.

Atshaves, B. P., Martin, G. G., Hostetler, H. A., McIntosh, A. L., Kier, A. B., & Schroeder, F. (2010). Liver fatty acid-binding protein and obesity. *The Journal of Nutritional Biochemistry* , *21* (11), 1015–1032.

Bass, N. M. (1988). The cellular fatty acid binding proteins: aspects of structure, regulation, and function. In *International review of cytology* (Vol. 111, pp. 143–184). Elsevier.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2010). GenBank. *Nucleic Acids Research* ,*39* (suppl_1), D32–D37.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* , *27* (2), 573–580.

Birney, E., Clamp, M., & Durbin, R. (2004). GeneWise and genomewise. *Genome Research* , *14* (5), 988–995.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., ... Phan, I. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* , *31* (1), 365–370.

Braissant, O., Foufelle, F., Scotto, C., Dauça, M., & Wahli, W. (1996). Differential expression of peroxisome proliferator-activated receptors (PPARs): tissue distribution of PPAR-alpha,-beta, and-gamma in the adult rat. *Endocrinology* , *137* (1), 354–366.

Campbell, M. S., Holt, C., Moore, B., & Yandell, M. (2014). Genome annotation and curation using MAKER and MAKER-P. *Current Protocols in Bioinformatics* , *48* (1), 4–11.

Castresana, J. (2002). GBLOCKS: selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Version 0.91 b. Copyrighted by J. Castresana, EMBL* .

Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., ... Li, Z. (2018). SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput

sequencing data. *Gigascience* , *7* (1), gix120.

Choi, Y., & Chan, A. P. (2015). PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels.*Bioinformatics* , *31* (16), 2745–2747.

Cifarelli, V., & Abumrad, N. A. (2011). Intestinal CD36 and other key proteins of lipid utilization: role in absorption and gut homeostasis.*Comprehensive Physiology* , *8* (2), 493–507.

Consortium, G. O. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* ,*32* (suppl_1), D258–D261.

De Bie, T., Cristianini, N., Demuth, J. P., & Hahn, M. W. (2006). CAFE: a computational tool for the study of gene family evolution.*Bioinformatics* , *22* (10), 1269–1271.

De Silva, S. S., Nguyen, T. T. T., Abery, N. W., & Amarasinghe, U. S. (2006). An evaluation of the role and impacts of alien finfish in Asian inland aquaculture. *Aquaculture Research* , *37* (1), 1–17.

De Silva, S. S., & Soto, D. (2009). Climate change and aquaculture: potential impacts, adaptation and mitigation. *Climate Change Implications for Fisheries and Aquaculture: Overview of Current Scientific Knowledge. FAO Fisheries and Aquaculture Technical Paper* ,*530* , 151–212.

Desvergne, B., & Wahli, W. (1999). Peroxisome proliferator-activated receptors: nuclear control of metabolism. *Endocrine Reviews* ,*20* (5), 649–688.

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., ... Aiden, A. P. (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds.*Science* , *356* (6333), 92–95.

Durand, N. C., Shamim, M. S., Machol, I., Rao, S. S. P., Huntley, M. H., Lander, E. S., & Aiden, E. L. (2016). Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems* , *3* (1), 95–98.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., ... Fitzgerald, S. (2013). Ensembl. *Nucleic Acids Res* ,*41* (2013), D48-55.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y., & Postlethwait, J. (1999). Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* , *151* (4), 1531–1545.

Fornes, O., Castro-Mondragon, J. A., Khan, A., Van der Lee, R., Zhang, X., Richmond, P. A., ... Baranašić, D. (2020). JASPAR 2020: update of the open-access database of transcription factor binding profiles.*Nucleic Acids Research* , *48* (D1), D87–D92.

Furuhashi, M., & Hotamisligil, G. S. (2008). Fatty acid-binding proteins: role in metabolic diseases and potential as drug targets.*Nature Reviews Drug Discovery* , *7* (6), 489–503.

Georgiadi, A., & Kersten, S. (2012). Mechanisms of gene regulation by fatty acids. *Advances in Nutrition* , *3* (2), 127–134.

Gertz, E. M., Yu, Y.-K., Agarwala, R., Schäffer, A. A., & Altschul, S. F. (2006). Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biology* ,*4* (1), 1–14.

Ghosh, S., & Chan, C.-K. K. (2016). Analysis of RNA-Seq data using TopHat and Cufflinks. In *Plant Bioinformatics* (pp. 339–361). Springer.

Guindon, S., Dufayard, J. F., Hordijk, W., Lefort, V., & Gascuel, O. (2009). PhyML: fast and accurate phylogeny reconstruction by maximum likelihood. *Infection Genetics and Evolution* , *9* (3), 384–385. ELSEVIER SCIENCE BV PO BOX 211, 1000 AE AMSTERDAM, NETHERLANDS.

Hemung, B., Visetsunthorn, A., & Pariwat, S. (2010). Chemical properties and fatty acid profile of lipids extracted from freshwater fish species. *Food Innovation Asia Conference 2010 Poster Presentation Proceedings*

, 669–675.

Hsu, M.-H., Palmer, C. N. A., Song, W., Griffin, K. J., & Johnson, E. F. (1998). A carboxyl-terminal extension of the zinc finger domain contributes to the specificity and polarity of peroxisome proliferator-activated receptor DNA binding. *Journal of Biological Chemistry* , *273* (43), 27988–27997.

Hsu, M. H. (n.d.). Novel Sequence Determinants in Peroxisome Proliferator Signaling. *Journal of Biological Chemistry* ,*270* (27).

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., ... Duquenne, L. (2009). InterPro: the integrative protein signature database. *Nucleic Acids Research* ,*37* (suppl_1), D211–D215.

Islami, S. N.-E., Reza, M. S., Mansur, M. A., Hossain, M. I., Shikha, F. H., & Kamal, M. (2014). Rigor index, fillet yield and proximate composition of cultured striped catfish (Pangasianodon hypophthalmus) for its suitability in processing industries in Bangladesh.*Journal of Fisheries* , *2* (3), 157–162.

Juge-Aubry, C., Pernin, A., Favez, T., Burger, A. G., Wahli, W., Meier, C. A., & Desvergne, B. (1997). DNA Binding Properties of Peroxisome Proliferator-activated Receptor Subtypes on Various Natural Peroxisome Proliferator Response Elements IMPORTANCE OF THE 5'-FLANKING REGION.*Journal of Biological Chemistry* , *272* (40), 25252–25259.

Jurka, J., Kapitonov, V. V, Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research* ,*110* (1–4), 462–467.

Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* , *28* (1), 27–30.

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* ,*12* (4), 357–360.

Kim, O. T. P., Nguyen, P. T., Shoguchi, E., Hisata, K., Vo, T. T. B., Inoue, J., ... Kanda, M. (2018). A draft genome of the striped catfish, Pangasianodon hypophthalmus, for comparative analysis of genes relevant to development and a resource for aquaculture improvement.*BMC Genomics* , *19* (1), 733.

Langdon, W. B. (2015). Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks.*BioData Mining* , *8* (1), 1.

Laprairie, R. B., Denovan-Wright, E. M., & Wright, J. M. (2016). Divergent evolution of cis-acting peroxisome proliferator-activated receptor elements that differentially control the tandemly duplicated fatty acid-binding protein genes, fabp1b. 1 and fabp1b. 2, in zebrafish.*Genome* , *59* (6), 403–412.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences.*Bioinformatics* , *34* (18), 3094–3100.

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* , *13* (9), 2178–2189.

Lynch, M., & Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* , *290* (5494), 1151–1155.

Pi, H., Liu, M., Xi, Y., Chen, M., Tian, L., Xie, J., ... Yu, Z. (2019). Long-term exercise prevents hepatic steatosis: a novel role of FABP1 in regulation of autophagy-lysosomal machinery. *The FASEB Journal* , *33* (11), 11870–11883.

Pryszcz, L. P., & Gabaldon, T. (2016). Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Research* ,*44* (12), e113–e113.

Salmela, L., & Rivals, E. (2014). LoRDEC: accurate and efficient long read error correction. *Bioinformatics* , *30* (24), 3506–3514.

Shimizu, M., Takeshita, A., Tsukamoto, T., Gonzalez, F. J., & Osumi, T. (2004). Tissue-selective, bidirectional regulation of PEX11α and perilipin genes through a common peroxisome proliferator response element. *Molecular and Cellular Biology* , *24* (3), 1313–1323.

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V, & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* ,*31* (19), 3210–3212.

Singh, A. K., & Lakra, W. S. (2012). Culture of Pangasianodon hypophthalmus into India: impacts and present scenario. *Pakistan Journal of Biological Sciences* , *15* (1), 19.

Smit, A. F. A., Hubley, R., & Green, P. (2014). RepeatModeler Open-1.0. 2008–2010. *Access Date Dec* .

Storch, J., & Thumser, A. E. A. (2000). The fatty acid transport function of fatty acid-binding proteins. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids* , *1486* (1), 28–44.

Tarailo-Graovac, M., & Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* , *25* (1), 4–10.

Taylor, J. S., & Raes, J. (2004). Duplication and divergence: the evolution of new genes and old ideas. *Annu. Rev. Genet.* ,*38* , 615–643.

Thompson, J. D., Gibson, T. J., & Higgins, D. G. (2003). Multiple sequence alignment using ClustalW and ClustalX. *Current Protocols in Bioinformatics* , (1), 2–3.

Thompson, J., Winter, N., Terwey, D., Bratt, J., & Banaszak, L. (1997). The Crystal Structure of the Liver Fatty Acid-binding Protein A COMPLEX WITH TWO BOUND OLEATES. *Journal of Biological Chemistry* ,*272* (11), 7140–7150.

Van Sang, N., Klemetsdal, G., Odegard, J., & Gjoen, H. M. (2012). Genetic parameters of economically important traits recorded at a given age in striped catfish (Pangasianodon hypophthalmus). *Aquaculture* , *344* , 82–89.

Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research* , *27* (5), 737–746.

Venkatachalam, A. B., Parmar, M. B., & Wright, J. M. (2017). Evolution of the duplicated intracellular lipid-binding protein genes of teleost fishes. *Molecular Genetics and Genomics* , *292* (4), 699–727.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., ... Young, S. K. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One* , *9* (11), e112963.

Wang, G., Bonkovsky, H. L., de Lemos, A., & Burczynski, F. J. (2015). Recent insights into the biological functions of liver fatty acid binding protein 1. *Journal of Lipid Research* , *56* (12), 2238–2247.

Xu, Z., & Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* , *35* (suppl_2), W265–W268.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* ,*13* (5), 555–556.

You, X., Shan, X., & Shi, Q. (2020). Research advances in the genomics and applications for molecular breeding of aquaculture animals. *Aquaculture* , 735357.

[dataset] Jieming Chen; Year: 2020; Pangasianodon hypophthalmus genome;CNGBdb (https://db.cngb.org/); ID: CNA0013719

Data Availability

Supporting datasets are included within this article and its Supplementary Information. The genome reads generated in this study have been deposited at CNGBdb (https://db.cngb.org/) under the accession number CNA0013719.

Author Contributions

Conceptualization, Q.S., J.X.; Sample providing, X.L.; Sample preparation, X.Z., J.C., T.X.; Analytical tool providing, C.B.; Data analysis and writing-original draft preparation, Z.G.; writing-review and editing, X.Y., Q.S., Y.H.; supervision, Q.S.; funding acquisition X.Y., Q.S. All authors have read and approved the published version of the manuscript.

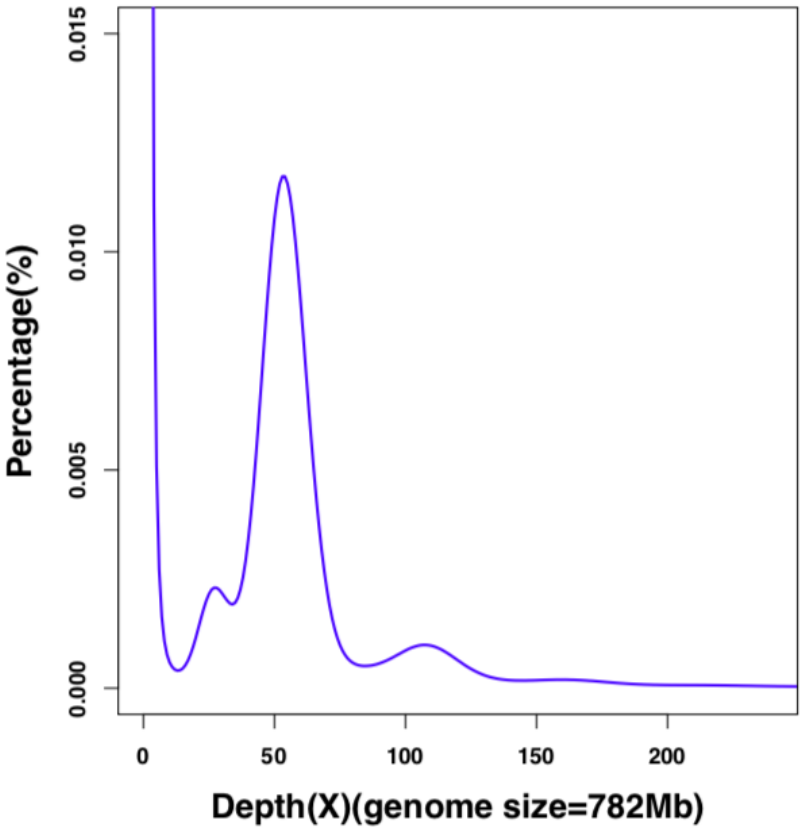

Fig. 1 Genome survey of the striped catfish. A 17-mer distribution of Illumina short reads was provided for genome size estimation. The x-axis represents the sequencing depth of each unique 17-mer, and the y-axis represents the percentage of unique 17-mers.

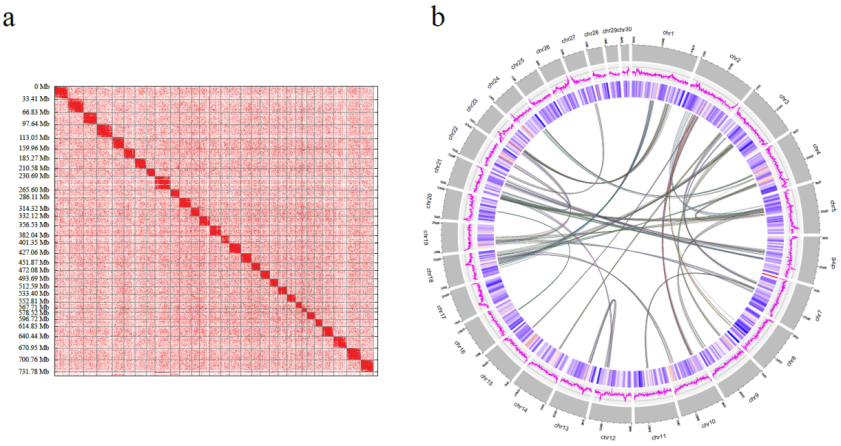

Fig. 2 A chromosome-level genome assembly of the striped catfish. (**a**)Contig contact matrix of the assembled genome. The color depth represents the density of Hi-C interactions. (**b**) A Circos atlas of the chromosomal genome for the striped catfish. From outside to inside rings include: (I) chromosomes (Mb in length), (II) distribution of gene density within 1-Mb non-overlapping windows, and (III) distribution of genomic GC content in 100-kb non-overlapping windows. The internal syntenic blocks are connected with lines.

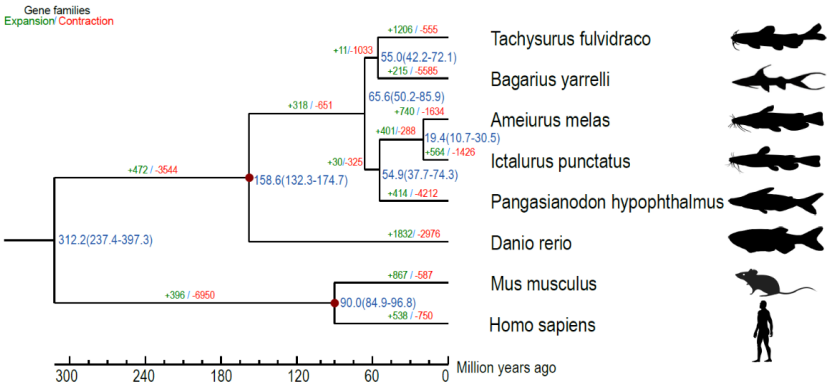

Fig. 3 A phylogenetic tree of eight representative vertebrates. Numbers on each branch represent the sum of expanded (green) and contracted (red) gene families for every clade. The number near each node indicates the divergence times, and the two red circles mark the calibration nodes.

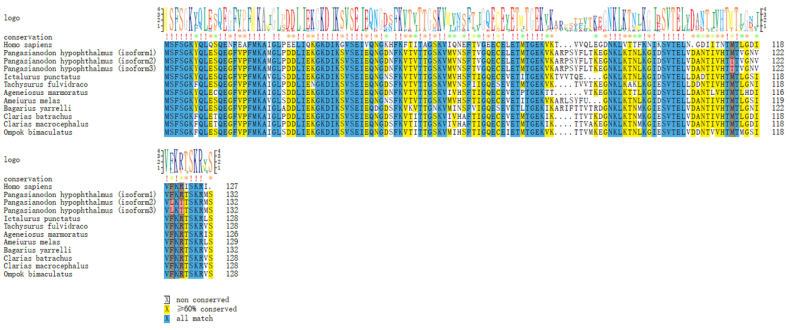

Fig. 4 Sequence alignment of Fabp1 proteins from various vertebrates. Red boxes mark the mutation sites in Fabp1-2 and -3 of the striped catfish.

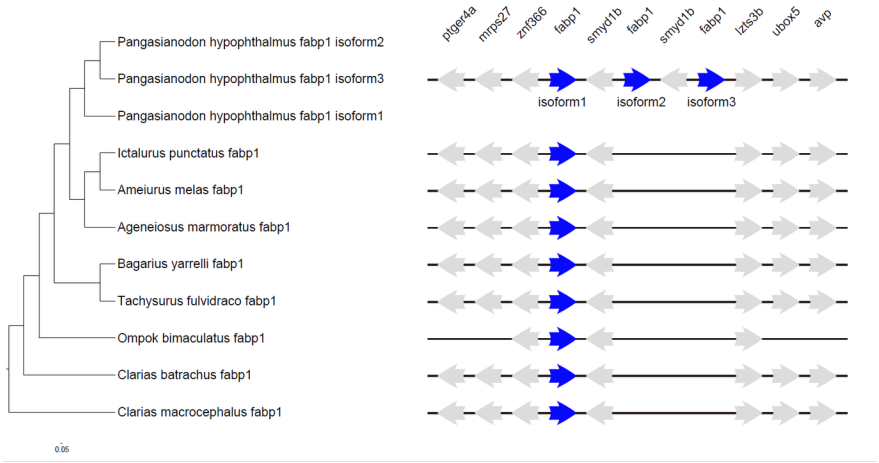

Fig. 5 Phylogenetic (left) and synteny (right) analyses of *fabp1* in various Siluriformes species. Blue arrows represent *fabp1* genes. Note the tandem triplication of *fabp1* gene (*fabp1-1* , *fabp1-2*and *fabp1-3*) in the striped catfish genome.

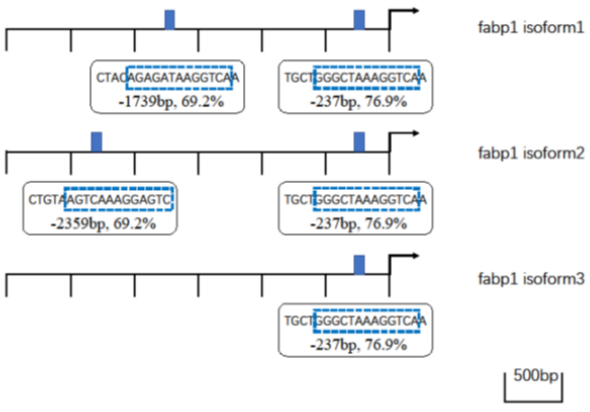

Fig. 6 Putative PPREs in the striped catfish *fabp1* promoters. Right-facing arrows represent the TSS. Rectangles mark the putative PPRE sequences, the position of putative PPRE relative to TSS, and sequence identity to the defined vertebrate PPRE consensus sequence. Blue dotted boxes indicate predicted DR1 in each putative PPRE.

Table 1 Summary of the genome assembly for the striped catfish

Genome assembly	**Parameter**
Contig N50 (Mb)	3.5
Contig number (>100 bp)	821
Scaffold N50 (Mb)	29.5
Scaffold number (>100 bp)	403
Total length (Mb)	742.6
Genome coverage (×)	147.7
Genome annotation	**Parameter**
Protein-coding gene number	18,895
Mean transcript length (bp)	22,381
Mean exons per gene	12.05
Mean exon length (bp)	167.91
Mean intron length (bp)	1,684

Table 2 Comparisons of the genome assemblies of various Siluriformes species

Species	**Scaffold N50**	**Contig N50**
Pangasianodon hypophthalmus (from the present study)	29,528,124	3,469,740
Pangasianodon hypophthalmus (GCF_003671635.1)	14,288,580	62,522
Ictalurus punctatus (GCA_004006655.2)	26,676,597	2,695,784
Ageneiosus marmoratus (GCA_003347165.1)	223,139	7,741
Ameiurus melas (GCA_012411365.1)	32,284,220	7,408,031
Bagarius yarrelli (GCA_005784505.1)	3,129,371	1,854,961
Clarias batrachus (GCA_003987875.1)	361,123	24,893
Clarias macrocephalus(GCA_011419295.1)	80,802	47,837
Ompok bimaculatus (GCA_009108245.1)	81,583	81,474
Tachysurus fulvidraco (GCA_003724035.1)	3,653,474	980,445

Table 3 Transcription of *fabp1-1* , *-2* and *-3* in the striped catfish

Gene id	**FPKM**	**Lower bound of the 95% confidence interval**	**Lower bound of the 95% confidence interv**
fabp1-1	22.6	18.9	26.4
fabp1-2	0	0	0
fabp1-3	4.8	3.1	6.5

Table 4 Prediction of functional effects for the amino acid substitutions in Fabp1 isotypes 2 and 3 of striped catfish

	PolyPhen-2 (HumDiv)	**PolyPhen-2 (HumDiv)**	**PolyPhen-2 (HumVar)**	**PolyPhen-2 (HumVar)**
	Score	**Prediction**	**Score**	**Prediction**

	PolyPhen-2 (HumDiv)	**PolyPhen-2 (HumDiv)**	**PolyPhen-2 (HumVar)**	**PolyPhen-2 (HumVar)**
F124L	0.008	BENIGN	0.009	BENIGN
M117T	0.884	DAMAGING	0.701	DAMAGING
R126T	0.999	DAMAGING	0.979	DAMAGING

Supplementary Tables

Table S1 Statistics of repeat elements in the striped catfish genome

(See the separate file)

Table S2 Classification of repeat elements in the striped catfish genome

(See the separate file)

Table S3 Protein-coding gene set of the striped catfish

(See the separate file)

Table S4 Function annotation of the striped catfish gene set

(See the separate file)

Table S5 BUSCO assessment of the assembled striped catfish genome

(See the separate file)

Table S6 Statistics of gene families among the examined eight species

(See the separate file)

Table S7 KEGG enrichment for genes in the expanded gene families of striped catfish

(See the separate file)