frontiers in Genetics

ORIGINAL RESEARCH
published: 16 October 2019
doi: 10.3389/fgene.2019.00991

A Chromosome-Scale Reference Assembly of a Tibetan Loach, *Triplophysa siluroides*

Liandong Yang[1], Ying Wang[2], Tai Wang[3], Shengchang Duan[4], Yang Dong[4], Yanping Zhang[3] and Shunping He[1,5]**

[1] The Key Laboratory of Aquatic Biodiversity and Conservation of Chinese Academy of Sciences, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, China, [2] School of Life Sciences, Jianghan University, Wuhan, China, [3] Gansu Key Laboratory of Cold Water Fishes Germplasm Resources and Genetics Breeding, Gansu Fishers Research Institute, Lanzhou, China, [4] Nowbio Biotechnology Company, Kunming, China, [5] Center for Excellence in Animal Evolution and Genetics, Chinese Academy of Sciences, Kunming, China

OPEN ACCESS

Edited by:
Chuan-Le Xiao,
Sun Yat-sen University, China

Reviewed by:
Milind B. Ratnaparkhe,
ICAR Indian Institute of Soybean Research, India
Olga Vinnere Pettersson,
Science for Life Laboratory (SciLifeLab), Sweden
Jia-Tang Li,
Chengdu Institute of Biology (CAS), China
Qiang Qiu,
Northwestern Polytechnical University, China

***Correspondence:**
Yanping Zhang
aqhongqi@qq.com
Shunping He
clad@ihb.ac.cn

Specialty section:
This article was submitted to Genomic Assay Technology, a section of the journal Frontiers in Genetics

Received: 05 July 2019
Accepted: 18 September 2019
Published: 16 October 2019

Citation:
Yang L, Wang Y, Wang T, Duan S, Dong Y, Zhang Y and He S (2019) A Chromosome-Scale Reference Assembly of a Tibetan Loach, Triplophysa siluroides. Front. Genet. 10:991. doi: 10.3389/fgene.2019.00991

Cobitoidea is one of the two superfamilies in Cypriniformes; however, few genomes have been sequenced for Cobitoidea fishes. Here, we obtained a total of 252.90 Gb of short Illumina reads and 31.60 Gb of long PacBio Sequel reads, representing approximate genome coverage of 256× and 50×, respectively. The final assembled genome is about 583.47 Mb with contig N50 sizes of 2.87 Mb, which accounts for 91.44% of the estimated genome size of 638.07 Mb. Using Hi-C–based chromatin contact maps, 99.31% of the genome assembly was placed into 25 chromosomes, and the N50 is 22.3 Mb. The gene annotation completeness was evaluated by BUSCO, and 2,470 of the 2,586 conserved genes (95.5%) could be found in our assembly. Repetitive elements were calculated to reach 33.08% of the whole genome. Moreover, we identified 25,406 protein-coding genes, of which 92.59% have been functionally annotated. This genome assembly will be a valuable genomic resource to understand the biology of the Tibetan loaches and will also set a stage for comparative analysis of the classification, diversification, and adaptation of fishes in Cobitoidea.

Keywords: ***Triplophysa siluroides*, PacBio sequencing, genome assembly, evolution, adaptation**

INTRODUCTION

The fish superfamily Cobitoidea is one of the two superfamilies of the order Cypriniformes, which is the largest monophyletic group of freshwater fishes in the world (Nelson et al., 2016). The classification and relationship of Cobitoidea are still under debate based on morphological and few molecular markers, for example, which families constitute the Cobitoidea (Tang et al., 2006; Slechtova et al., 2007; Mayden et al., 2009; Kottelat, 2012; Nelson et al., 2016). Therefore, it is essential to investigate the relationship of Cobitoidea fishes at the genomic level. Compared with the many genome sequences released from fishes in Cyprinoidea, there are still few genomes that have been sequenced yet for Cobitoidea fishes, which hampers remarkably further comparative analyses of all Cobitoidea fishes. As ecologically and commercially important freshwater species, some fishes of the superfamily Cobitoidea play important roles in the commercial fisheries on China, such as oriental weatherloach (*Misgurnus anguillicaudatus*) (Chen et al., 2014) and giant stone loach (*Triplophysa siluroides*) (**Figure 1**) (Zhu, 1989; Chen et al., 2016).

FIGURE 1 | Photograph of the Tibetan loach, *Triplophysa siluroides*.

The giant stone loach belongs to the family Nemacheilidae (Cypriniformes) and is an endemic species restricted to the upper reaches of the Yellow River drainage in China (Ding, 1994). It is the biggest member of balitorid loachs in the world attaining about 0.5 m in total length and weight of about 1.5 kg (Zhu, 1989) and was previously an important economic fish in its distribution regions (Zhu, 1989; Chen et al., 2016). However, the natural population of the giant stone loach has reduced sharply in recent years because of heavy fishing and habitat destruction, which makes it a vulnerable species in the China Red Data Book of Endangered Animals (Wang et al., 1998; Wang and Xie, 2004; He et al., 2008). Thus, it is in urgent need to perform genetic analysis on the giant stone loach to protect their natural resources, especially at the genomic level. However, so far, only transcriptome, mitochondrial genome, and SNP data have been reported for the giant stone loach (Wang et al., 2015; Chen et al., 2016; Chen et al., 2018). It is thus essential to sequence the genome of the giant stone loach, which will help protect this species, identify functional genes controlling important economic traits, reveal the genetic basis of adaptation to the extreme environments of the Tibetan Plateau, and, most importantly, provide a reference genome for the Cobitoidea fishes.

In this work, we integrated genomic sequencing data from Illumina short reads and PacBio long reads to generate a reference genome for *T. siluroides*. The completeness and continuity of this chromosome level genome were comparable with other model fish species, which will definitely provide valuable genomic resources for studies for the evolution and adaptation of Cobitoidea fishes.

MATERIALS AND METHODS

Tissue Sampling and Ethics Statement

Tissue for genome sequencing of *T. siluroides* (NCBI taxonomy ID: 422203) was sampled from a single individual collected from the Yellow River at Gansu Province in China (33°25′N, 102°17′E). Muscle was collected and frozen in liquid nitrogen. All animal experimental procedures were approved by the ethics committee of Institute of Hydrobiology, Chinese Academy of Sciences.

Library Construction and Sequencing

Genomic DNA was extracted from the muscle tissue using Qiagen GenomicTip100 (Qiagen, Hilden, Germany). For Illumina sequencing, we constructed a total of seven libraries with four short-insert libraries (170, 220, 320, and 600 bp) and three long-insert libraries (2, 5, and 10 kb) using the standard protocol provided by Illumina (San Diego, CA, USA). Paired-end sequencing was performed using the Illumina HiSeq 2000 platform for each library.

For the long insert size library, we sequenced it on a PacBio Sequel instrument with Sequel SMRT cells 1M v2 (Pacific Biosciences, Menlo Park, CA, USA) with one movie of 600 min at the Genome Center of Nextomics (Wuhan, China). In brief, approximately 5 μg of DNA was used to construct one single-molecule real-time (SMRT) library with an insert size of 20 kb. The library was sequenced in five SMRT DNA sequencing cells.

Genome Size Estimation and Genome Assembly

We estimated the genome size based on the 17-mer depth frequency distribution method (Liu et al., 2013) with the following formula: genome size = k-mer_number/k-mer_depth (k-mer_number is the total number of k-mer from the sequencing data, and k-mer_depth is the peak frequency that was higher than any other frequencies).

Hybrid assembly of Illumina short reads and PacBio Sequel long reads was performed using the programs Platanus (Kajitani et al., 2014) and DBG2OLC (Ye et al., 2016). In short, the high-quality paired-end reads were used to construct accurate de Bruijin graph contigs using the program Platanus (Kajitani et al., 2014). Then, the program DBG2OLC (Ye et al., 2016) was used to map short contigs to PacBio Sequel long reads and generate a hybrid assembly. We further corrected the mixed assembly results by Pilon (Walker et al., 2014), with default parameters. Finally, the program SSPACE (Boetzer et al., 2011) was used to scaffold the hybrid assembly by incorporating mate pair reads.

Genome Scaffolding With Chromatin Contact Maps

The processes of crosslinking, lysis, chromatin digestion, biotin marking, proximity ligations, crosslinking reversal, and DNA purification steps were used in previous studies (Dudchenko et al., 2017). Briefly, the fresh fish muscle sample was treated with 1% formaldehyde for 10 min at room temperature to perform cross-linking. The reaction was then quenched by adding 2.5 M glycine to 0.2 M for about 5 min. Nuclei were further digested with 100 units of DpnII and marked with biotin-14-dCTP (Invitrogen) and then ligated by T4 DNA ligase. After incubating overnight to reverse cross-links, the ligated DNA was then sheared to 300- to 600-bp fragments. The DNA fragments were further blunt-end repaired and A-tailed, followed by purification through biotin-streptavidin–mediated pull-down. Finally, the Hi-C libraries were quantified and sequenced on the Illumina HiSeq X Ten platform (San Diego, CA, USA) with 150 paired-end mode. The sequencing reads were mapped to the hybrid genome assembly with BWA (Li and Durbin, 2010),

and uniquely mapped read pairs were retained. Contigs from the hybrid genome assembly were clustered, ordered, and oriented using Proximo (Burton et al., 2013).

Assessment of Genome Completeness

The completeness of our *de novo* genome assembly was evaluated using benchmarking universal single-copy orthologs (BUSCO, v3) (Simao et al., 2015), which quantitatively assesses genome completeness using evolutionarily highly conserved 2,586 single-copy vertebrate genes. We also assessed the percentage of reads covered in our genome assembly by mapping the high-quality Illumina reads for short insert size libraries onto the *de novo* genome assembly using bwa (Li and Durbin, 2010) with default parameters.

Repeat Annotation

We analyzed the repetitive sequences in *T. siluroides* genome with a combination of *de novo* and homology-based methods. First, we constructed a *de novo* repeat library using the RepeatModeller (v. 1.05) (Tarailo-Graovac and Chen, 2009) and LTR FINDER (Xu and Wang, 2007) with default parameters. Then, we mapped our assembled genome sequences against the constructed *de novo* repeat libraries and the RepBase (v. 21.01) (Jurka et al., 2005) to detect the novel and known transposable elements using the RepeatMasker (v. 4.06) (Tarailo-Graovac and Chen, 2009). Meantime, we employed the Tandem Repeat Finder (v. 4.04) (Benson, 1999) to predict the tandem repeats. Finally, we used the RepeatProteinMask software (v. 4.0.6) (Tarailo-Graovac and Chen, 2009) to annotate transposable element relevant proteins in our genome assembly.

Gene Annotation

To annotate the structures and functions of putative genes in *T. siluroides* genome assembly, we used both *ab initio* prediction and homology-based prediction methods. For *ab initio* prediction, we used Augustus (Stanke et al., 2006), GenScan (Burge and Karlin, 1997), and glimmerHMM (Majoros et al., 2004) programs to analyze the repeat-masked *T. siluroides* genome assembly. For homology-based prediction, homologous protein sequences of cave fish (*Astyanax mexicanus*) (McGaugh et al., 2014), zebrafish (*Danio rerio*, GRCz10) (Howe et al., 2013), medaka (*Oryzias latipes*) (Kasahara et al., 2007), and Japanese puffer (*Fugu rubripes*) (Aparicio et al., 2002) were obtained from Ensembl (release 89) (Cunningham et al., 2015) and aligned to the repeat-masked *T. siluroides* genome using TblastN (version 2.2.26) with an *E* value cutoff of 1e−5. Then, the aligned sequences and corresponding query protein were filtered and passed to Genewise (version 2.4.1) (Birney et al., 2004) to predict the potential gene structures on all alignments. Finally, the above two gene sets were integrated to yield a comprehensive and nonredundant gene set using EVidenceModeler (EVM, version 1.1.1) (Haas et al., 2008).

Then, gene functional annotations were performed by aligning translated gene coding sequences to known databases, including SwissProt and TrEMBL, Gene Ontology (GO), InterProScan, and Kyoto Encyclopedia of Genes and Genomes (KEGG), using BLASTP (version 2.2.26) with an *E* value of 1e−5.

In addition, we also identified noncoding RNA genes in the *T. siluroides* genome. We used blast to search rRNA against the rRNA database and tRNAscan-SE (Lowe and Eddy, 1997) to search tRNA in the genome sequences. We also used blast to search miRNA and snRNA genes *via* the Rfam database (Gardner et al., 2011).

Phylogenetic Analysis

Protein sequences of 11 ray-finned fishes (D. rerio, Gasterosteus aculeatus, Lepisosteus oculatus, Oreochromis niloticus, O. latipes, Takifugu rubripes, Xiphophorus maculatus, A. mexicanus, Gadus morhua, Poecilia formosa, and Tetraodon nigroviridis) were downloaded from the Ensembl database (Release 90), and the protein sequences of Hippocampus comes (Lin et al., 2016) and Boleophthalmus pectinirostris (You et al., 2014) were acquired from the authors. The longest coding sequence was chosen to represent each gene. We first performed all-against-all comparison of all proteins using BLASTP (version 2.2.26) with a cutoff of E value <1e−5 to both genes and then clustered the genes into gene families using solar and hcluster_sg in TreeFam (Li et al., 2006). Subsequently, we extracted the one-to-one orthologous genes from the aforementioned 14 species. The protein sequences of these orthologous genes were aligned using MUSCLE (Edgar, 2004) with the default parameters. We then converted the protein alignments to their corresponding coding sequences (CDSs) using an in-house perl script. All these aligned nucleotide sequences were then concatenated into a supergene. Next, the 4D sites (fourfold degenerate sites) were extracted from the supergenes to construct a phylogenetic tree using RAxML (Stamatakis, 2014) with the GTR+G+I model.

RESULTS AND DISCUSSION

In total, we generated about 252.90 Gb of raw Illumina reads, including 43.96, 40.33, 43.64, 42.74, 27.84, 27.81, and 26.58 Gb of reads from the 170-, 220-, 320-, 600-, 2k-, 5k-, and 10-kb libraries, respectively (**Table S1**). We also generated about 31.60 Gb of raw PacBio long data with an average read length of 10,563 bp (**Table S2**). After removal of low-quality and redundant reads, 163.37 Gb of clean Illumina reads and 31.26 Gb of clean PacBio reads were obtained for genome assembly (**Tables S1** and **S2**). The genome size estimated by k-mer analysis was approximately 638 Mb, with the main peak at a depth of 183× (**Figure S1** and **Table S3**). The small peak at a depth of 92 indicated that the genome heterozygosity of *T. siluroides* was low (0.29%).

We assembled the genome with the hybrid method of Illumina short reads and PacBio Sequel long reads using the programs Platanus (Kajitani et al., 2014) and DBG2OLC (Ye et al., 2016). The final *de novo* assembly for the *T. siluroides* has a total length of 583.47 Mb, representing 91.44% of the estimated genome size, with contig N50 length of 2.87 Mb and the longest contig length 14.65 Mb (**Table 1**), which makes it one of the most high-quality genome assemblies currently available.

To construct a chromosome-scale reference genome assembly of the Tibetan loach, chromatin contact maps were produced by Frasergen Information Co. Ltd. (Wuhan, China)

TABLE 1 | Summary of genome assembly of *T. siluroides*.

Terms	Size (bp)	Number
N90	452,639	251
N80	881,594	163
N70	1,446,248	112
N60	2,005,727	77
N50	2,872,994	53
Max length	14,649,642	—
Total length	583,471,586	—
Total number	—	1,004
Total number (≥1 kb)	—	1,004
Total number (≥5 kb)		1,001

(**Figure 2**). We sequenced a total of 65.39 Gb of HiSeq data and obtained 27.7 Gb valid data (43.95%) that could be used to anchor the contigs into chromosomes. The contig clustering allowed the placement of 856 contigs into 25 scaffolds (chromosomes) with lengths ranging from 16.67 to 34.49 Mb (**Table S4**). While only 81.99% of the contigs were anchored to chromosomes, this corresponds to 99.31% of the total length of primary hybrid genome assembly. This genome scaffolding step improved substantially the primary assembly contiguity, raising the N50 approximately 8.3-fold from 2.7 Mb to 22.3 Mb (**Table 2**).

The completeness of our *de novo* genome assembly was assessed by BUSCO, which showed that 98.4% of the 2,586 highly

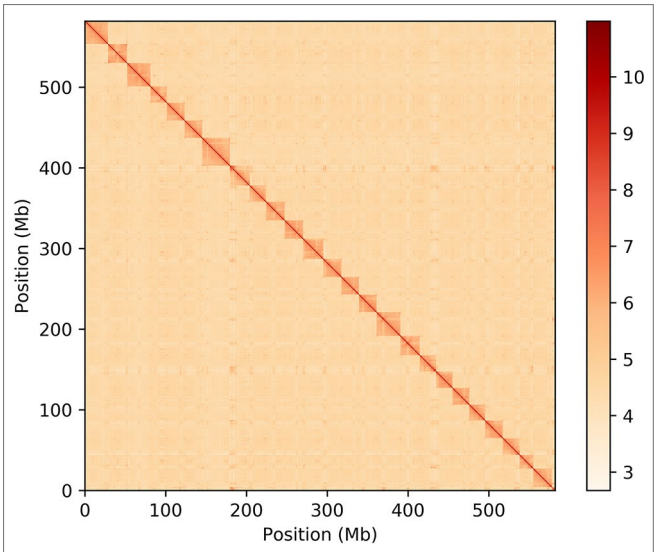

FIGURE 2 | Hi-C interaction heat map showing the clustering of the primary hybrid contigs into 25 sets of chromosome-scale superscaffolds.

TABLE 2 | Chromosome metrics before and after Hi-C scaffolding.

Terms	Contig (original)	Scaffold (Hi-C)
Number	1,004	856
N50	2,872,994	22,312,937
Total length	582,350,959	578,738,912

conserved single-copy genes can be detected in the *T. siluroides* genome, with 95.5% and 2.9% identified as complete and fragmented, respectively (**Table S5**). We also found that 98.36% of the high-quality Illumina reads can be mapped onto the *de novo* genome assembly (**Table S6**). These results suggested that the quality of our *de novo* assembled genome was high for both completeness and base level accuracy.

A total of 193 Mb of nonredundant repetitive sequences are identified in *T. siluroides* genome, which account for 33.08% of the whole genome. The percentage of repetitive sequences is similar to other fish species (Chalopin et al., 2015). The most predominant repeat is the DNA transposons, which account for 12.58% (73.38 Mb in total) of the genome (**Table S7**, **Table S8**, **Figure S2**, and **Figure S3**). The fraction of DNA transposons in *T. siluroides* genome is in good agreement with those in other fish species, which indicated that the fraction of DNA transposons in fish genomes (10%) is significantly higher than those in mammals (3%) (Chalopin et al., 2015).

After the characterization of repetitive sequences in the *T. siluroides* genome assembly, gene annotation was performed by using both *ab initio* prediction and homology-based prediction methods. In total, 25,406 protein-coding genes were identified (**Table S9**, **Figure S4**, and **Figure S5**). Approximately 92.59% of the predicted genes were successfully annotated using five protein databases: InterPro (83.40%), GO (67.67%), KEGG (69.67%), Swiss-Prot (85.84%), and TrEMBL (92.38%) (**Table S10**). Finally, we identified noncoding RNA genes in the *T. siluroides* genome and found that a total of 6,822 microRNAs (miRNA), 6,513 transfer RNA (tRNA), 8,053 ribosomal RNA (rRNA), and 12,655 snRNA genes could be detected in the *T. siluroides* genome (**Table S11**).

We further obtained the gene families for 14 fish species and then classified these gene families for a subset of five species (*D. rerio*, *H. comes*, *T. rubripes*, *X. maculatus*, and *T. siluroides*) (**Figure 3A**). In brief, the 25,406 protein-coding genes in *T. siluroides* comprised 2,104 single-copy orthologs, 14,222 multiple-copy orthologs, 1,003 unique paralogs, 6,122 other orthologs, and 1,955 unclustered genes (**Figure S6**). Furthermore, 9,225 gene families were identified in the *T. siluroides* genome, and 300 of these were found to be unique in *T. siluroides* genome (**Table S12**). We found that the *T. siluroides* species-specific gene families were mainly enriched in the following GO categories, including immune response, energy metabolism, and hormone activity, implying that species-specific genes may play important roles in *T. siluroides* adaptation to the extreme environments on the Tibetan Plateau. Based on the TreeFam gene clusters and MUSCLE multiple alignment, 1,087 one-to-one orthologs were identified from the 14 fish genomes. Phylogenetic analysis from these orthologs supported the placement of *T. siluroides* close to zebrafish (**Figure 3B**).

CONCLUSION

We report the high-quality whole genome sequencing, assembly, and annotation of the Tibetan loach (*T. siluroides*). The high-quality genome assembly will provide a valuable resource for

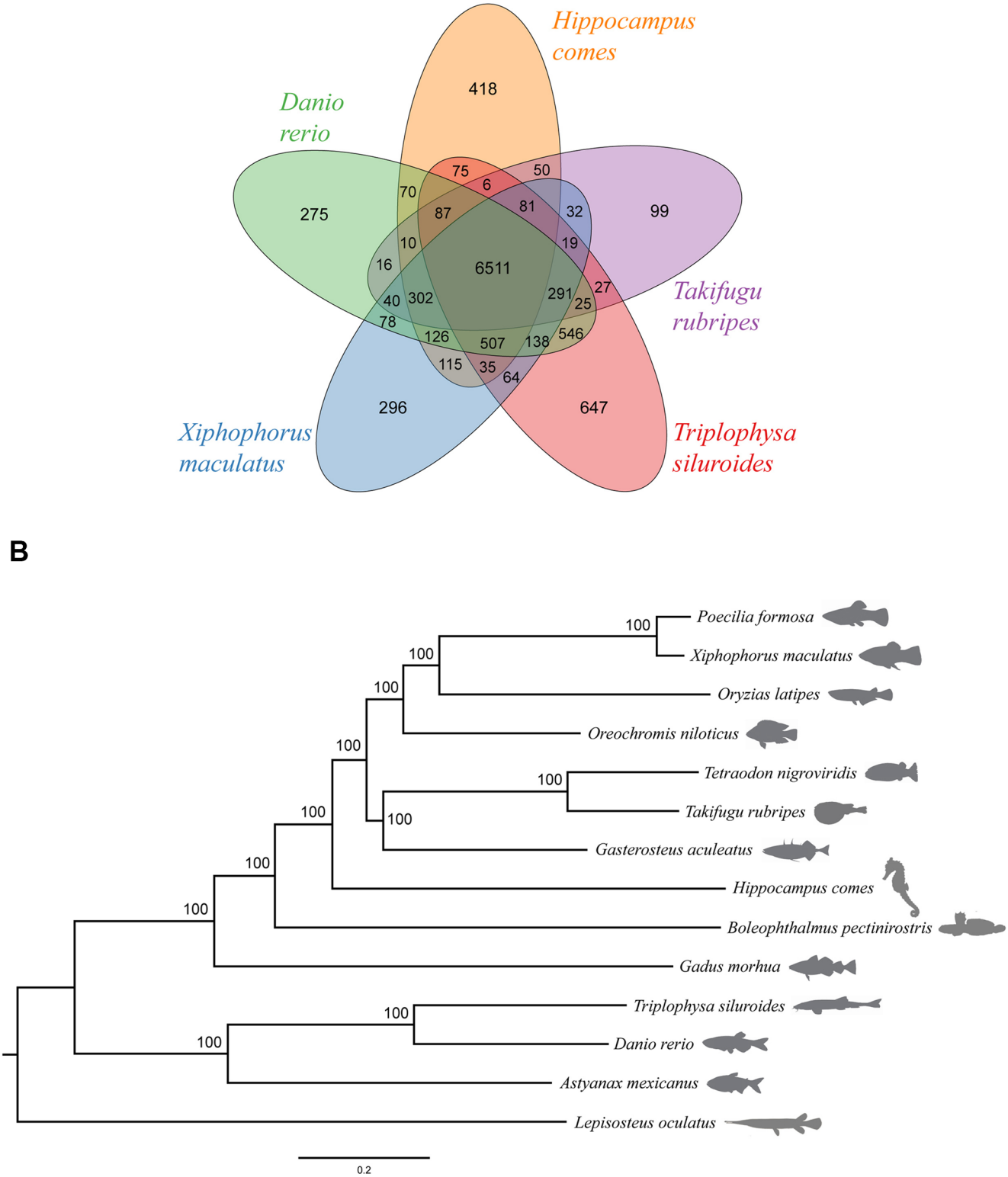

FIGURE 3 | Genome evolution. **(A)** A Venn diagram of the orthologous gene families shared among five ray-finned fish genomes (*Danio rerio*, *Hippocampus comes*, *Takifugu rubripes*, *Triplophysa siluroides*, *Xiphophorus maculatus*). **(B)** Phylogeny of the 14 ray-finned fishes with the spotted gar as the outgroup. The bootstrap support value for the topology is 100.

studying the genetic mechanisms of adaptation to the Tibetan Plateau in fishes.

DATA AVAILABILITY STATEMENT

The raw sequencing reads of all libraries have been deposited in the SRA database (SRP198880).

ETHICS STATEMENT

The animal study was reviewed and approved by The ethics committee of Institute of Hydrobiology, Chinese Academy of Sciences.

AUTHOR CONTRIBUTIONS

SH and LY designed the study. YZ coordinated the study. YW and TW collected the sample. SD and YD performed the bioinformatics analysis. LY and YW analyzed the results and wrote the manuscript with inputs from the other authors. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

This work was supported by grants from the National Natural Science Foundation of China (31972866 and 31601858) and Institute of Hydrobiology, Chinese Academy of Sciences (Y55Z09 and Y85E03) to LY, from the Strategic Priority Research Program (XDB13020100 and XDB060101) and the National Natural Science Foundation of China (91731301) to SH, and from the National Natural Science Foundation of China (31702016) to YW. We thank Nowbio Biotechnology, Inc. (Kunming, China) and Frasergen Information Co. Ltd. (Wuhan, China) for technical support.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00991/full#supplementary-material.

REFERENCES

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J. M., Dehal, P., et al. (2002). Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* 297, 1301–1310. doi: 10.1126/science.1072104

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27, 573–580. doi: 10.1093/nar/27.2.573

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. *Genome Res.* 14, 988–995. doi: 10.1101/gr.1865504

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., and Pirovano, W. (2011). Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27, 578–579. doi: 10.1093/bioinformatics/btq683

Burge, C., and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* 268, 78–94. doi: 10.1006/jmbi.1997.0951

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R. L., Kitzman, J. O., and Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* 31, 1119-+. doi: 10.1038/nbt.2727

Chalopin, D., Naville, M., Plard, F., Galiana, D., and Volff, J. N. (2015). Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biol. Evol.* 7, 567–580. doi: 10.1093/gbe/evv005

Chen, I. S., Liu, G. D., and Prokofiev, A. M. (2016). The complete mitochondrial genome of giant stone loach Triplophysa siluroides (Cypriniformes: Balitoridae). *Mitochondrial DNA Part A* 27, 998–1000. doi: 10.3109/19401736.2014.926523

Chen, S., Zhang, G., Shao, C., Huang, Q., Liu, G., Zhang, P., et al. (2014). Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. *Nat. Genet.* 46, 253–260. doi: 10.1038/ng.2890

Chen, Y., Gong, Q., Liu, Y., Song, M., Du, J., and Lai, J. (2018). Isolation and characterization of 50 SNP markers in Triplophysa siluroides. *Conserv. Genet. Resour.* 1–4. doi: 10.1007/s12686-018-1059-3

Cunningham, F., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., et al. (2015). Ensembl 2015. *Nucleic Acids Res.* 43, D662–D669. doi: 10.1093/nar/gku1010

Ding, R. (1994). *The fishes of Sichuan*. Chengdu: Sichuan Publishing House of Science and Technology.

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., et al. (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* 356, 92–95. doi: 10.1126/science.aal3327

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Gardner, P. P., Daub, J., Tate, J., Moore, B. L., Osuch, I. H., Griffiths-Jones, S., et al. (2011). Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res.* 39, D141–D145. doi: 10.1093/nar/gkq1129

Haas, B. J., Salzberg, S. L., Zhu, W., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., et al. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. *Genome Biol.* 9, R7. doi: 10.1186/gb-2008-9-1-r7

He, C. L., Zhang, X. Y., Hou, F. X., Zhang, X. F., and Song, Z. B. (2008). Threatened fishes of the world: Triplophysa siluroides (Herzenstein 1888) (Balitoridae). *Environ. Biol. Fishes* 83, 305–305. doi: 10.1007/s10641-008-9339-5

Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., et al. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature* 496, 498–503. doi: 10.1038/nature12111

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J., (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* 110, 462–467. doi: 10.1159/000084979

Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., et al. (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* 24, 1384–1395. doi: 10.1101/gr.170720.113

Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., et al. (2007). The medaka draft genome and insights into vertebrate genome evolution. *Nature* 447, 714–719. doi: 10.1038/nature05846

Kottelat, M. (2012). Conspectus Cobitidum: an inventory of the loaches of the world (Teleostei: Cypriniformes: Cobitoidei). *Raffles Bull. Zool.*, 1–175.

Li, H., Coghlan, A., Ruan, J., Coin, L. J., Heriche, J. K., Osmotherly, L., et al. (2006). TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res.* 34, D572–D580. doi: 10.1093/nar/gkj118

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* 26, 589–595. doi: 10.1093/bioinformatics/btp698

Lin, Q., Fan, S., Zhang, Y., Xu, M., Zhang, H., Yang, Y., et al. (2016). The seahorse genome and the evolution of its specialized morphology. *Nature* 540, 395–399. doi: 10.1038/nature20595

Liu, B., Shi, Y., Yuan, J., Hu, X., Zhang, H., Li, N., et al. (2013). Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *Quant. Biol.* 35, 62–67.

Lowe, T. M., and Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955–964. doi: 10.1093/nar/25.5.955

Majoros, W. H., Pertea, M., and Salzberg, S. L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* 20, 2878–2879. doi: 10.1093/bioinformatics/bth315

Mayden, R. L., Chen, W. J., Bart, H. L., Doosey, M. H., Simons, A. M., Tang, K. L., et al. (2009). Reconstructing the phylogenetic relationships of the earth's most diverse clade of freshwater fishes—order Cypriniformes (Actinopterygii: Ostariophysi): a case study using multiple nuclear loci and the mitochondrial genome. *Mol. Phylogenet. Evol.* 51, 500–514. doi: 10.1016/j.ympev.2008.12.015

McGaugh, S. E., Gross, J. B., Aken, B., Blin, M., Borowsky, R., Chalopin, D., et al. (2014). The cavefish genome reveals candidate genes for eye loss. *Nat. Commun.* 5, 5307. doi: 10.1038/ncomms6307

Nelson, J. S., Grande, T. C., and Wilson, M. V. H. (2016). *Fishes of the world*. 5th ed. (New Jersey: John Wiley & Sons, Inc.) doi: 10.1002/9781119174844

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. doi: 10.1093/bioinformatics/btv351

Slechtova, V., Bohlen, J., and Tan, H. H. (2007). Families of Cobitoidea (Teleostei; Cypriniformes) as revealed from nuclear genetic data and the position of the mysterious genera Barbucca, Psilorhynchus, Serpenticobitis and Vaillantella. *Mol. Phylogenet. Evol.* 44, 1358–1365. doi: 10.1016/j.ympev.2007.02.019

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 30, 1312–1313. doi: 10.1093/bioinformatics/btu033

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* 34, W435–W439. doi: 10.1093/nar/gkl200

Tang, Q., Liu, H., Mayden, R., and Xiong, B. (2006). Comparison of evolutionary rates in the mitochondrial DNA cytochrome b gene and control region and their implications for phylogeny of the Cobitoidea (Teleostei: Cypriniformes). *Mol. Phylogenet. Evol.* 39, 347–357. doi: 10.1016/j.ympev.2005.08.007

Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* Chapter 4, Unit 4, 10. doi: 10.1002/0471250953.bi0410s25

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9, e112963. doi: 10.1371/journal.pone.0112963

Wang, S., and Xie, Y. (2004). *China species red list (in Chinese)* Vol. 1. Beijing, China: Higher Education Press, p. 692.

Wang, S., Yue, P., and Chen, Y., (1998). *China red data book of endangered animals: Pisces*. Beijing: Science Press.

Wang, Y., Yang, L., Zhou, K., Zhang, Y., Song, Z., and He, S. (2015). Evidence for adaptation to the Tibetan plateau inferred from Tibetan loach transcriptomes. *Genome Biol. Evol.* 7, 2970–2982. doi: 10.1093/gbe/evv192

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* 35, W265–W268. doi: 10.1093/nar/gkm286

Ye, C., Hill, C. M., Wu, S., Ruan, J., and Ma, Z. S. (2016). DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Sci. Rep.* 6, 31900. doi: 10.1038/srep31900

You, X., Bian, C., Zan, Q., Xu, X., Liu, X., Chen, J., et al. (2014). Mudskipper genomes provide insights into the terrestrial adaptation of amphibious fishes. *Nat. Commun.* 5, 5594. doi: 10.1038/ncomms6594

Zhu, S. (1989). *The loaches of the subfamily Nemacheilinae in China (Cypriniformes: Cobitidae)*. (Nanjing: Jiangsu Science and Technology Publishing House).

Conflict of Interest: Authors SD and YD were employed by Nowbio Biotechnology Company, Kunming, Yunnan, China.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The reviewer QQ declared a past co-authorship with two of the authors YD, SH to the handling editor.