# Chromosome-level genome assembly of the Chinese Longsnout catfish *Leiocassis longirostris*

Wenping He[1,2†], Jian Zhou[3†], Zhe Li[1,2], Tingsen Jing[1,2], Chunhua Li[4], Yuejing Yang[1,5], Mengbin Xiang[1,5], Chaowei Zhou[1,2], Guangjun Lv[1,2], Hongyan Xu[1,2], Hui Luo[1,2*], Hua Ye[1,2*]

[1]Key Laboratory of Freshwater Fish Reproduction and Development (Ministry of Education), College of Fisheries, Southwest University, Chongqing, 402460, China

[2] Key Laboratory of Aquatic Science of Chongqing 400175, China

[3] Fisheries Institute, Sichuan Academy of Agricultural Sciences, Chengdu 611731, China

[4] BGI-Qingdao, BGI-Shenzhen, Qingdao 266555, China

[5] Fisheries Research Institute in Wanzhou Chongqing, Chongqing, 404000, China

H. Luo (🖂), e-mail: luohui2629@126.com

H. Ye (🖂), e-mail: yhlh2000@126.com

[†] Wenping He, and Jian Zhou contributed equally to this work

## Abstract

The Chinese Longsnout catfish *Leiocassis longirostris* (*L. longirostris*) is one of the most economically important freshwater catfish in China. It is a valuable model for studies on sexual dimorphism, comparative and conservation biology since its wild resources have declined sharply. However, there is lack of high-quality chromosome-level genome information for comparative genomic analysis and genome evolutionary studies. Therefore, we constructed the first high-quality chromosome-level reference genome for *L. longirostris* using a combined strategy of BGI-SEQ500, Nanopore, and Hi-C technologies. The assembled genome of *L. longirostris* contained a total length of 703.19 Mb with 389 contigs, and an N50 size of 4.29 Mb. Using the Hi-C data, we finally successfully generated 82 chromosome-level scaffolds anchored onto 26 chromosomes with a total length of 685.53 Mb (97.44% of the total genome sequences), ranging in size from 17.36 Mb to 43.97 Mb. A total of 23,708 protein-coding genes were identified in the *L. longirostris* genome, and up to 97.73% of *L. longirostris* genes were functionally annotated. In addition, the genome contained 239.11 Mb (33.99% in the total genome) repetitive sequences and 6,303 non-coding RNAs. The phylogenetic analysis indicated that the divergence time between *L. longirostris* and their closest relative species *Pelteobagrus fulvidraco* was approximately 26.6 million years. Collinearity analyses showed 26 chromosomes of *L. longirostris* displayed high homology with the corresponding scaffold (≥3M) of *P.*

*fulvidraco* and the corresponding chromosomes of the *Ictalurus punctatus*. The high-quality reference genome of *L. longirostris* was assembled for the first time and will pave a way for genome-scale selective breeding, genome comparisons and evolution investigations.

**Keywords:** *Leiocassis longirostris*; Chromosome-level genome assembly; Nanopore sequencing; Hi-C; Comparative genomics

## 1 Introduction

The Siluriformes (catfish) is one of the most diverse order in fish, with roughly 4100 species, which account for nearly 12% of all fish species (Z. Liu et al., 2016). Catfish is the major aquaculture species in the world especially in the United States, China and Vietnam (Kocher & Kole, 2008). Additionally, catfish can serve as a model for comparative genome studies because they are evolutionarily closer to a common fish ancestor than most bony fish in phylogenetic analysis (Moyle, Cech, & Cummings, 2004).

The Chinese Longsnout catfish (*Leiocassis longirostris* Günther), also named Jiangtuan, belonging to the family Bagridae which consists of more than 220 species (Ferraris, 2007), order Siluriformes, is a semi-migratory freshwater species, and an indigenous commercially important fish species commonly distributed in the Huaihe River, Liaohe River, Minjiang River, Yangtze River and Pearl River in China, and western regions of the Korean Peninsula (Shen et al., 2014; Wang, Zhou, Ye, Wei, & Wu, 2006; Zhu et al., 2005). In recent years, the wild resources of *L. longirostris* are rapidly decreased due to over-fishing, water pollution, and other human disturbances, such as building hydropower stations (Liang, Guo, Luo, Li, & Zou, 2016; Luo, Jiang, Liu, Zhan, & Xia, 2000; Wang et al., 2006; Xiao & Yang, 2009). So, it is urgent to conduct studies on the conservation biology of *L. longirostris*.

With the increasing demands of consumption owing to its nutritional value, and wonderful flavor, the commercial value of *L. longirostris* is becoming higher. The rapid expansion and intensification of *L. longirostris* aquaculture have led to tremendous challenges, including germplasm degeneration and poor diseases resistance, which have seriously limited the sustainable development of the industry of this species. *L. longirostris* exhibits markedly sex dimorphism in growth, that the males grow much faster than females, therefore, constructing reference genome of *L. longirostris* will facilitate developing genetic breeding programs and the sex control technique, thus benefit the aquaculture industry with increasing the yield of fishery.

In this report, we constructed the first high-quality chromosome-level reference genome for *L. longirostris* using a combined strategy of BGI-SEQ500, Nanopore and high-throughput chromosome conformation capture (Hi-C) technologies. In addition, a genome-wide phylogenetic analysis and gene family analysis among 11 teleost species had been performed. The findings of this work will be useful for genome-scale selective breeding of *L. longirostris*, as well as offering chromosome information for genome comparisons and evolution investigations among important aquaculture species.

## 2 Materials and Methods

### 2.1 Sample collection and DNA / RNA extraction.

One healthy adult female *L. longirostris* (Figure 1) collected from a farm of Sichuan Academy of Agricultural Sciences in Meishan, Sichuan Province, China, was used for genome sequencing. The muscle was collected for DNA extraction after treatment with the anaesthetic tricaine MS-222. Then the liver tissues of 15 *L. longirostris* collected from the same farm were harvested for RNA extraction. All samples used for DNA extraction were kept at -80°C, while the tissues used for RNA extraction were immediately frozen in liquid nitrogen for 2 hours and then stored at −80°C.

Genomic DNA was isolated from muscle using the standard chloroform-isoamyl alcohol extraction procedures (Orkin, 1990). DNA quality and quantity were measured using NanoDrop™ One UV-Vis spectrophotometer (Thermo Fisher Scientific, USA) and Qubit® 3.0 Fluorometer (Invitrogen, USA), respectively. This study followed the guidelines of the Committee of Laboratory Animal Experimentation at Southwest University.

The livers were used for RNA extraction using TRIzol reagent (Invitrogen, USA), and then treated with DNase I (Invitrogen, USA) to remove genomic DNA (Ye et al., 2018). RNA concentration and integrity were measured using Qubit® RNA Assay Kit in Qubit® 2.0 Flurometer (Life Technologies, CA, USA) and RNA Nano 6000 Assay Kit of the Agilent Bioanalyzer 2100 system (Agilent Technologies, CA, USA), respectively. Three RNA sequencing libraries with an insert size of 250-300 bp were prepared using NEBNext® Ultra™ RNA Library Prep Kit for Illumina® (NEB, USA) following the manufacturer's protocol, and then sequenced on an Illumina Hiseq X

Ten platform (Illumina Inc., San Diego, CA, USA) as 150 bp paired-end reads.

### 2.2 Library construction and sequencing.

Genomic DNA from the one healthy adult female *L. longirostris* muscle was used for sequencing on the BGI-SEQ500 and Nanopore platform. DNA library with 200-400 bp insert size was constructed following the manufacturer's instructions as the description in the previous study (Huang et al., 2017). Then, the library was sequenced according to the BGISEQ-500 protocol (Huang et al., 2017). For the Nanopore sequencing, the DNA after recovering, terminal repairing, and quantifying was used to prepare a library with Ligation Sequencing Kit (Oxford, SQK-LSK109) according to manufacturer's instructions. This library was sequenced with the Nanopore GridION X5 sequencer on a flow cell. For the construction of Hi-C library, 1 g muscle tissue was used to prepare a library according to the previous studies (Rao et al., 2014). The library was then sequenced on a BGISEQ-500 sequencer using 100 bp pair end sequencing.

### 2.3 *de novo* assembly of the *L. longirostris* genome.

In the genome survey, the raw reads of the *L. longirostris* from BGI-SEQ500 platform were filtered using SOAPnuke software (Y. Chen et al., 2017), and then BLAST was applied for the evaluation of sample contamination (Altschul, Gish, Miller, Myers, & Lipman, 1990). The adapter sequences were removed from the reads, and paired reads with more than 10% ambiguous or low-quality (Phred Score < 5) bases were also discarded. GenomeScope was used for *Kmer*-based analysis of *L. longirostris* genome (Vurture et al., 2017). A *Kmer* frequency distribution of *L. longirostris* was obtained using *Kmer* size of 17.

Long reads generated from the Nanopore sequencing platform were used for the de novo genome assembly using Canu software according three steps of error-correction, repairing, and assembling (Koren et al., 2017). Then, contigs were polished three times using BGI-SEQ500 short reads by Pilon (Walker et al., 2014). Finally, we used the Purge Haplotigs pipeline to produce an improved, deduplicated assembly (Roach, Schmidt, & Borneman, 2018). The completeness of assembled genome for *L. longirostris* was validated by Benchmarking sets of Universal Single-Copy Orthologs (BUSCO) analysis using BUSCO v3.0 with the actinopterygii_odb9 database (PMID: 26059717 DOI: 10.1093/bioinformatics/btv351).

For chromosome-level assembly of *L. longirostris*, the Hi-C reads from the library sequenced on a BGISEQ-500 sequencer using 100 bp pair end sequencing were first filtered by HIC-Pro (Servant et al., 2015). Then, Juicer (version 1.5) (Durand, Shamim, et al., 2016) and 3D-DNA (3D de novo assembly) (Dudchenko et al., 2017) were used to map the Hi-C reads to the assembled contig sequences. The interaction frequencies among each contig were calculated from the sequencing data, and the contig contact matrix of *L. longirostris* genome was mapped using Juicebox (Durand, Robinson, et al., 2016).

### 2.4 Gene prediction and functional annotation.

For the annotation of repetitive sequences, we used RepeatModeler v1.0.10 (RepeatModeler, RRID:SCR 015027), which employs two complementary computational methods (RECONv1.08 and RepeatScout v1.0.5 (RepeatScout, RRID:SCR 014653)) for identifying repeat element boundaries and family relationships from sequence data. Subsequently, the outputs from RepeatModeler and the RepBase library were combined and used for further characterization of transposable elements (TEs), many of which are not repetitive, and other repeats by homology-based methods, including identification with Repeat-Masker (v4.0.7, rmblast-2.2.28) (RepeatMasker, RRID:SCR 012954).

The combined strategy of homology-based, *de novo*, and transcriptome-based method was used for genes structure prediction. The protein sequences of 9 fish species including *Danio rerio*, *Gasterosteus aculeatus*, *Ictalurus Punctatus*, *Larimichthys crocea*, *Oreochromis niloticus*, *Oryzias latipes*, *Pangasianodon hypophthalmus*, *Tachysurus fulvidraco*, and *Takifugu rubripes*, were downloaded from the Ensembl database and mapped onto the assembled *L. longirostris* genome using BLASTN (Altschul et al., 1990). And then, GeneWise (version 2.2.0) (Birney, Clamp, & Durbin, 2004) were used to homologous annotation. For *de novo* prediction, Augustus (Stanke & Waack, 2003) was used to predict the structure of genes. In addition, the RNA-seq data were aligned to the assembled *L. longirostris* genome to predict gene coding regions by PASA (DOI: 10.1093/nar/gkg770). We used Swissprot (Boeckmann et al., 2003), Kyoto Encyclopedia of Gene and Genomes (KEGG) (Kanehisa & Goto, 2000), TrEMBL (Boeckmann et al., 2003), Interpro (Zdobnov &

Apweiler, 2001), and Gene Ontology (GO) (Ashburner et al., 2000) to re-annotate protein-coding genes.

For non-coding RNAs, miRNAs and snRNAs were predicted by the INFERNAL tool using Rfam database (Kalvari et al., 2018). The tRNAs and rRNAs were identified by tRNAscan-SE v1.3.1 (tRNAscan-SE, RRID:SCR_010835) software (Lowe & Eddy, 1997) and RNAmmer v1.2 (Lagesen et al., 2007), respectively.

### 2.5 Gene family and phylogenetic analysis

To identify gene families, protein sequences from the longest transcripts of each gene from *L. longirostris* and other 10 fish species, including *D. rerio, Astyanax mexicanus*, *G. aculeatus*, *Glyptosternum maculatum*, *I. punctatus*, *Lepisosteus oculatus*, *O. niloticus*, *O. latipes*, *Pelteobagrus fulvidraco*, and *T. rubripes*, were aligned to each other using BLASTP (Altschul et al., 1990) with e-value threshold of 1e$^{-5}$. And then, OrthoMCL (L. Li, Stoeckert, & Roos, 2003) were used to construct gene families.

To investigate the evolutional relationship of *L. longirostris* with the other above mentioned 10 fish species, the common single-copy genes were used for phylogenetic reconstruction by MUSCLE (Edgar, 2004). Then, RAxML (Stamatakis, 2014) was employed to construct the phylogenetic tree. MCMCTREE (PAML software package) (Yang, 2007) were used to estimate the divergence time based on the “Correlated molecular clock” and “HKY85” model. Collinearity analyses of chromosomes or scaffold between *L. longirostris* and *P. fulvidraco* and between *L. longirostris* and *I.*

*punctatus* were performed using MCScan software55 (Tang et al., 2008).

# 3 Results and Discussion

## 3.1 Genome Assembly

BGI-SEQ500 sequencing data, Nanopore sequencing and Hi-C reads were used for estimating genome size, contig assembly, and the chromosome assembly, respectively. As a result, we obtained a total of 64.11 Gb clean data for the genome survey, 43.23 Gb long reads with the average sequencing coverage of 61.48 X for the following genome assembly, and 243.13 Gb raw Hi-C data (Table S1). Using *Kmer* size of 17, a *Kmer* frequency distribution for *L. longirostris* was obtained (Figure S1). As a result, the genome size of *L. longirostris* was estimated as 688.99 Mb with the heterozygosity, repeat content, and GC content of 0.35%, 42.53%, and 38.43%, respectively. After contig assembly using long reads generated from the Nanopore sequencing platform, we obtained the assembled genome of *L. longirostris* containing a total length of 703.19 Mb with 389 contigs, and an N50 size of 4.29 Mb, which is the middle genome size in sequenced catfish genomes (Table 1; Table S2). The overall GC-content of 39.67% in *L. longirostris* genome was slightly higher than that of the walking catfish (*Clarias batrachus*) (N. Li et al., 2018) and *Cyprinus carpio* but much lower than those of most of the teleost genomes (Xu et al., 2014). The completeness of assembled genome for *L. longirostris* was validated by Benchmarking sets of Universal Single-Copy Orthologs (BUSCO) analysis using BUSCO v3.0 with the actinopterygii_odb9 database. As a result, 4,293 (93.6 %) of the 4,584 BUSCO genes were completely identified in the genome with 4,109 (89.6%) single-copy and 184 (4.0%) duplicated genes (Table 2), suggesting a great success of the de novo genome assembly in *L. longirostris*.

A total of 126.35 Gb clean Hi-C reads were obtained, and finally, we successfully generated 82 chromosome-level scaffolds containing 473 contigs anchored onto 26 chromosomes with a total length of 685.53 Mb (97.44% of the total genome sequences). The number of chromosomes scaffold in this study was consistent with the result of previous karyotype analyses for *L. longirostris* by Hong et al. 1984 (2n=52) (Hong & Zhou, 1984). The lengths of chromosomes ranged from 17.36 Mb to 43.97 Mb. The scaffold and contig N50 of the chromosome assembly was 28.03 and 3.09 Mb, respectively (Table S3). The heatmap of chromosome crosstalk indicated that the *L. longirostris* genome assembly was complete and robust (Figure 2).

## 3.2 Genome Annotation

Using a combination of homology based on the Repbase and *de novo* methods, a total of 239.11 Mb (33.99% in the total genome) were identified as repetitive elements, in which DNA transposons (146.40Mb, 20.81%) were most abundant repeat type in the genome (Table 3). The proportion of repetitive elements of *L. longirostris* is similar to the *Glyptosternon maculatum* genome (33.96%) (H. Liu et al., 2018), and higher than those of most teleost genomes, such as 5.7% in *Tetraodon nigroviridis* (Van de Peer, 2004), 7.1% in *T. rubripes* (Aparicio et al., 2002), 13.48% in *G. aculeatus* (Jones et al., 2012), 26.13% in *L. crocea* (B. Chen et al., 2019), 30.68% in *Oryzias latipes*

(Kasahara et al., 2007), 31.3% in *C. carpio* (Xu et al., 2014), 32.56% in *I. punctatus* (X. Chen et al., 2016), but lower than that for the *Bagarius yarrelli* (35.26%) (Jiang et al., 2019), *Onychostoma macrolepis* (46.23%) (Sun et al., 2020), and *D. rerio* genome (52.2%) (Howe et al., 2013).

We combined the results from the three approaches (homology-based, *de novo*, and transcriptome-based method) used for genes structure prediction, and found that a total of 23,708 protein-coding genes were identified in the *L. longirostris* genome (Table 4). Compared with other published catfish genomes, the number of genes in *L. longirostris* is similar to that in *P. fulvidraco* (Gong et al., 2018; Zhang et al., 2018), *G. maculatum* (H. Liu et al., 2018), *I. Punctatus* (X. Chen et al., 2016; Z. Liu et al., 2016), and *C. batrachus* (N. Li et al., 2018), but more than that in *B. yarrelli* (Jiang et al., 2019), and less than that in *P. hypophthalmus* (Kim et al., 2018) (Table 1). The comparison between *L. longirostris* and other 3 catfish species in gene length, coding DNA sequence (CDS), intron length, exon number, and exon length distribution, were showed in Figure S2. A total of 23,170 protein-coding genes were annotated with at least one of the databases, and up to 97.73% of *L. longirostris* genes were functionally annotated (Table S4). After non-coding RNAs analysis, 422 miRNA, 2,118 tRNA, 1,838 rRNA, and 1,925 snRNA, were annotated in *L. longirostris* genome (Table S5). The BUSCO analysis was performed to test for completeness of the assembled genome using single-copy gene database of Actinopterygii. As a result, about 92.4% complete BUSCOs were found, and with 5.6% duplicates, 4.0% fragmented, and 3.6% missing from the reference gene set. The results indicated that we obtained a good assembly genome for *L. longirostris*.

### 3.3 Comparative Genomic Analyses

To estimate species-specific and shared genes in the *L. longirostris* compared with other 10 fish species, we used OrthoMCL (L. Li et al., 2003) to define the orthologous genes. The results showed that a total of 19,438 gene families were identified among the 11 species, in which 3,585 single-cope ortholog families from the 11 species were identified, and 68 families were specific to *L. longirostris* (Table S6). Furthermore, a total of 11,729 gene families were shared by the 4 catfish species, and 301 gene families were specific to *L. longirostris* (Figure 3).

Then, a genome-wide phylogenetic tree was constructed based on the identified single-cope ortholog genes. The phylogenetic analysis indicated that the same family Bagridae species *L. longirostris* and *P. fulvidraco* were clustered into one branch, and *L. longirostris* is close to clades of the *P. fulvidraco*, *G. maculatum*, and *I. punctatus*, which belong to the Siluriformes order. This result was similar to the phylogenetic analysis based on mitochondrial genome by Liu et al. (Y. Liu et al., 2019). The divergence time was estimated using 7 calibration points between *L. longirostris* and their closest relative species *P. fulvidraco* was approximate 26.6 million years (Figure 4). In addition, the phylogenetic analysis estimated that the *I. punctatus* around 82.2 million years ago from their common ancestor *P. fulvidraco*, which consistent with the result of 81.9 million years analysis by Gong et al. (Gong et al., 2018).

Collinearity analyses of chromosomes or scaffold between *L. longirostris* and *P.*

*fulvidraco* and between *L. longirostris* and *I. punctatus* were performed using MCScan software55 (Tang et al., 2008). As a result, all 26 pseudo-chromosomes of *L. longirostris* displayed high homology with the corresponding scaffold (≥3M) of *P. fulvidraco* and the corresponding chromosomes of the *I. punctatus* (Figure 5), suggesting a good assembly genome for *L. longirostris*.

## 4 Conclusion

In the present study, the first high-quality chromosome sequences for *L. longirostris* were constructed using a combined strategy of BGI-SEQ500, Nanopore and Hi-C technologies. Likewise, the contig N50 of 4.29 Mb was assembled with a high quality of the reference genome (703.19 Mb) in *L. longirostris.* Also, 473 contigs were successfully scaffolded into 26 chromosomes with a scaffold N50 length of 28.03 Mb through Hi-C technologies. In addition, 23,708 protein-coding genes were predicted and annotated in the assembled *L. longirostris* genome. Intriguingly, the phylogenetic analysis indicated that the same family Bagridae species *L. longirostris* and *P. fulvidraco* were clustered into one branch, and the divergence time between these two fish species was approximately 26.6 million years. The findings of this study will facilitate developing genome-scale selective breeding of *L. longirostris*, as well as offering chromosome information for genome comparisons and evolution investigations among important aquaculture species.

**Conflict interests**

The authors declare that they have no conflict interests.

**Acknowledgements**

This research was supported by China Agriculture Research System (CARS-46), the Fundamental Research Funds for the Central Universities (XDJK2017B008, XDJK2017C035, XDJK2019C025, 5360300098), Natural Science Foundation of Chongqing (cstc2020jcyj-msxmX0438) and Natural Science Foundation of China (32071651).

**Author contributions**

W. H., H. L., J. Z., and H. Y. designed the experiments; W. H., H. L., J. Z., Z. L., T. J., C. L., Y. Y., M. X., and C. Z. performed the experiments and/or analysed data; W. H., G. L., W. H., H. X., and H. Y. wrote the paper. All authors reviewed the manuscript.

# References

Altschul, S., Gish, W., Miller, W., Myers, E., & Lipman, D. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology*, *215*(3), 403–410. doi: 10.1016/S0022-2836(05)80360-2

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J., Dehal, P., … Brenner, S. (2002). Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science*, *297*(5585), 1301–1310. doi: 10.1126/science.1072104

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., … Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. doi: 10.1038/75556

Birney, E., Clamp, M., & Durbin, R. (2004). GeneWise and genomewise. *Genome Research*, *14*(5), 988–995. doi: 10.1101/gr.1865504

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., … Schneider, M. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research*, *31*(1), 365–370. doi: 10.1093/nar/gkg095

Chen, B., Zhou, Z., Ke, Q., Wu, Y., Bai, H., Pu, F., & Xu, P. (2019). The sequencing and de novo assembly of the *Larimichthys crocea* genome using PacBio and Hi-C technologies. *Scientific Data*, *6*, 188. doi: 10.1038/s41597-019-0194-3

Chen, X., Zhong, L., Bian, C., Xu, P., Qiu, Y., You, X., … Bian, W. (2016). High-quality genome assembly of channel catfish, *Ictalurus punctatus*. *Gigascience*, *5*, 39. doi: 10.1186/s13742-016-0142-5

Chen, Y., Chen, Y., Shi, C., Huang, Z., Zhang, Y., Li, S., … Chen, Q. (2017). SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput sequencing data. *Gigascience*, *7*(1). doi: 10.1093/gigascience/gix120

Dudchenko, O., Batra, S. S., Omer, A. D., Nyquist, S. K., Hoeger, M., Durand, N. C., … Aiden, E. L. (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science*, *356*(6333), 92–95. doi: 10.1126/science.aal3327

Durand, N. C., Robinson, J. T., Shamim, M. S., Machol, I., Mesirov, J. P., Lander, E. S., & Aiden, E. L. (2016). Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Systems*, *3*(1), 99–101. doi: 10.1016/j.cels.2015.07.012

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. doi: 10.1016/j.cels.2016.07.002

Ferraris, C. J. (2007). Checklist of catfishes, recent and fossil (Osteichthyes: Siluriformes), and catalogue of siluriform primary types. *Zootaxa*, *1418*(1), 1–628. doi: 10.11646/zootaxa.1418.1.1

Gong, G., Dan, C., Xiao, S., Guo, W., Huang, P., Xiong, Y., … Mei, J. (2018). Chromosomal-level assembly of yellow catfish genome using third-generation DNA sequencing and Hi-C analysis. *Gigascience*, *7*(11). doi: 10.1093/gigascience/giy120

Hong, Y., & Zhou, T. (1984). Karyotypes of nine species of Chinese Catfish (Bagridae). *Zoological Research (China)*, *5*(3), 21–26.

Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., … Stemple, D. L. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, *496*(7446), 498–503. doi: 10.1038/nature12111

Huang, J., Liang, X., Xuan, Y., Geng, C., Li, Y., Lu, H., … Gao, S. (2017). A reference human genome dataset of the BGISEQ-500 sequencer. *Gigascience*, *6*(5). doi: 10.1093/gigascience/gix024

Jiang, W., Lv, Y., Cheng, L., Yang, K., Bian, C., Wang, X., … Shi, Q. (2019). Whole-Genome Sequencing of the Giant Devil Catfish, *Bagarius yarrelli*. *Genome Biology and Evolution*, *11*(8), 2071–2077. doi: 10.1093/gbe/evz143

Jones, F. C., Grabherr, M. G., Chan, Y. F., Russell, P., Mauceli, E., Johnson, J., … Kingsley, D. M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, *484*(7392), 55–61. doi: 10.1038/nature10944

Kalvari, I., Argasinska, J., Quinones-Olvera, N., Nawrocki, E. P., Rivas, E., Eddy, S. R., … Petrov, A. I. (2018). Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Research*, *46*(D1), D335–D342. doi: 10.1093/nar/gkx1038

Kanehisa, M., & Goto, S. (2000). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, *28*(1), 27–30. doi: 10.1093/nar/28.1.27

Kasahara, M., Naruse, K., Sasaki, S., Nakatani, Y., Qu, W., Ahsan, B., … Kohara, Y. (2007). The medaka draft genome and insights into vertebrate genome evolution. *Nature*, *447*(7145), 714–719. doi: 10.1038/nature05846

Kim, O. T. P., Nguyen, P. T., Shoguchi, E., Hisata, K., Vo, T. T. B., Inoue, J., … Satoh, N. (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*, 733. doi: 10.1186/s12864-018-5079-x

Kocher, T. D., & Kole, C. (2008). Genome Mapping and Genomics in Fishes and Aquatic Animals. In *Genome Mapping and Genomics in Fishes and Aquatic Animals* (Vol. 2). Heidelberg,Germany: Springer-Verlag Berlin.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. doi: 10.1101/071282

Lagesen, K., Hallin, P., Rodland, E. A., Staerfeldt, H.-H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, *35*(9), 3100–3108. doi: 10.1093/nar/gkm160

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, *13*(9), 2178–2189. doi: 10.1101/gr.1224503

Li, N., Bao, L., Zhou, T., Yuan, Z., Liu, S., Dunham, R., … Liu, Z. (2018). Genome sequence of walking catfish (*Clarias batrachus*) provides insights into terrestrial adaptation. *Bmc Genomics*, *19*(1), 952. doi: 10.1186/s12864-018-5355-9

Liang, H., Guo, S., Luo, X., Li, Z., & Zou, G. (2016). Molecular diagnostic markers of *Tachysurus fulvidraco* and *Leiocassis longirostris* and their hybrids. *Springerplus*, *5*(1), 2115–2121. doi: 10.1186/s40064-016-3766-0

Liu, H., Liu, Q., Chen, Z., Liu, Y., Zhou, C., Liang, Q., … Mou, Z. (2018). Draft genome of *Glyptosternon maculatum*, an endemic fish from Tibet Plateau. *Gigascience*, *7*(9). doi: 10.1093/gigascience/giy104

Liu, Y., Wu, P.-D., Zhang, D.-Z., Zhang, H.-B., Tang, B.-P., Liu, Q.-N., & Dai, L.-S. (2019). Mitochondrial genome of the yellow catfish *Pelteobagrus fulvidraco* and insights into Bagridae phylogenetics. *Genomics*, *111*(6), 1258–1265. doi: 10.1016/j.ygeno.2018.08.005

Liu, Z., Liu, S., Yao, J., Bao, L., Zhang, J., Li, Y., … Waldbieser, G. C. (2016). The channel catfish genome sequence provides insights into the evolution of scale formation in teleosts. *Nature Communications*, *7*, 11757. doi: 10.1038/ncomms11757

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, *25*(5), 955–964. doi: 10.1093/nar/25.5.955

Luo, M., Jiang, L., Liu, Y., Zhan, G. A., & Xia, S. (2000). COMPARATIVE STUDY ON LSOENZYMES IN CLEIOCASSIS LONGIROSTRIS. *Chinese Journal of Applied & Environmental Biology*, *6*(5), 447–451. doi: 10.3321/j.issn:1006-687X.2000.05.012

Moyle, P. B., Cech, J. J., & Cummings, B. (2004). *Fishes: An Introduction to Ichthyology*. Pearson Prentice Hall. doi: 10.2307/1308732

Orkin, S. (1990). Molecular-Cloning - a Laboratory Manual, 2nd Edition - Sambrook,j, Fritsch,ef, Maniatis,t. *Nature*, *343*(6259), 604–605. doi: 10.1038/343604a0

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., … Aiden, E. L. (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, *159*(7), 1665–1680. doi: 10.1016/j.cell.2014.11.021

Roach, M. J., Schmidt, S. A., & Borneman, A. R. (2018). Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *Bmc Bioinformatics*, *19*(1), 460. doi: 10.1186/s12859-018-2485-7

Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C.-J., Vert, J.-P., … Barillot, E. (2015). HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology*, *16*(1), 259. doi: 10.1186/s13059-015-0831-x

Shen, T., He, X., Lei, M., Wang, J., Li, X., & Li, J. (2014). Cloning and structure of a histocompatibility class IIA gene (Lelo-DAA) in Chinese longsnout catfish (*Leiocassis longirostris*). *Genes & Genomics*, *36*(6), 745–753. doi: 10.1007/s13258-014-0208-7

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. doi: 10.1093/bioinformatics/btu033

Stanke, M., & Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, *19*, II215–II225. doi: 10.1093/bioinformatics/btg1080

Sun, L., Gao, T., Wang, F., Qin, Z., Yan, L., Tao, W., … Wang, D. (2020). Chromosome-level genome assembly of a cyprinid fish *Onychostoma macrolepis* by integration of nanopore sequencing, Bionano and Hi-C technology. *Molecular Ecology Resources*. doi: 10.1111/1755-0998.13190

Tang, H., Wang, X., Bowers, J. E., Ming, R., Alam, M., & Paterson, A. H. (2008). Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Research*, *18*(12), 1944–1954. doi: 10.1101/gr.080978.108

Van de Peer, Y. (2004). Tetraodon genome confirms Takifugu findings: most fish are ancient polyploids. *Genome Biology*, *5*(12), 250. doi: 10.1186/gb-2004-5-12-250

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J., & Schatz, M. C. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics*, *33*(14), 2202–2204. doi: 10.1093/bioinformatics/btx153

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. doi: 10.1371/journal.pone.0112963

Wang, Z., Zhou, J., Ye, Y., Wei, Q., & Wu, Q. (2006). Genetic structure and low-genetic diversity suggesting the necessity for conservation of the Chinese Longsnout catfish, *Leiocassis longirostris* (Pisces : Bagriidae). *Environmental Biology of Fishes*, *75*(4), 455–463. doi: 10.1007/s10641-006-0035-z

Xiao, M., & Yang, G. (2009). Isolation and characterization of 17 microsatellite loci for the Chinese longsnout catfish (*Leiocassis longirostris*). *Molecular Ecology Resources*, *9*(3), 1039–1041. doi: 10.1111/j.1755-0998.2009.02554.x

Xu, P., Zhang, X., Wang, X., Li, J., Liu, G., Kuang, Y., … Sun, X. (2014). Genome sequence and genetic diversity of the common carp, *Cyprinus carpio*. *Nature Genetics*, *46*(11), 1212–1219. doi: 10.1038/ng.3098

Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution*, *24*(8), 1586–1591. doi: 10.1093/molbev/msm088

Ye, H., Xiao, S., Wang, X., Wang, Z., Zhang, Z., Zhu, C., … Luo, H. (2018). Characterization of Spleen Transcriptome of *Schizothorax prenanti* during *Aeromonas hydrophila* Infection. *Marine Biotechnology*, *20*(2), 246–256. doi: 10.1007/s10126-018-9801-0

Zdobnov, E. M., & Apweiler, R. (2001). InterProScan - an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, *17*(9), 847–848. doi: 10.1093/bioinformatics/17.9.847

Zhang, S., Li, J., Qin, Q., Liu, W., Bian, C., Yi, Y., … Chen, X. (2018). Whole-Genome Sequencing of Chinese Yellow Catfish Provides a Valuable Genetic Resource for High-Throughput Identification of Toxin Genes. *Toxins*, *10*(12), 488. doi: 10.3390/toxins10120488

Zhu, X. M., Xie, S. Q., Lei, W., Cui, Y. B., Yang, Y. X., & Wootton, R. J. (2005). Compensatory growth in the Chinese longsnout catfish, *Leiocassis longirostris* following feed deprivation: Temporal patterns in growth, nutrient deposition, feed intake and body composition. *Aquaculture*, *248*(1–4), 307–314. doi: 10.1016/j.aquaculture.2005.03.006

Figure 1 A picture of the *L. longirostris* used for the genome sequencing

Figure 2 The Hi-C contact map of the *L. longirostris* genome, The color bar shows the contact density from red (high) to white (low)

Figure 3 Venn diagram of the orthologous genes

Figure 4 Phylogenetic tree of 11 fish genomes constructed using 3,585 single copy orthologous genes

Figure 5 Genome comparisons between *L. longirostris* and *P. fulvidraco* (A), and between *L. longirostris* and *I. punctatus* (B)

## Tables

Table 1 Summary of sequenced catfish genomes

| Species | Family | Sequencing platform | Assembled genome size (Mb) | Identified genes | Scaffold N50 (Mb) | Contig N50 (kb) | References |
|---|---|---|---|---|---|---|---|
| Longsnout catfish, *Leiocassis longirostris* | Bagridae | BGI-SEQ500, Nanopore, Hi-C | 703.19 | 23,708 | 28.03 | 3090.00 | |
| Yellow catfish, *Pelteobagrus fulvidraco* | Bagridae | Illumina, PacBio, Hi-C | 732.80 | 24,552 | 25.80 | 1100.00 | Gong et al., 2018 |
| | | Illumina, PacBio | 714.00 | 21,562 | 3.65 | 970.00 | Zhang et al., 2018 |
| *Glyptosternon maculatum* | Sisoridae | PacBio, Illumina, 10X Genomics, BioNano | 662.34 | 22,066 | 20.90 | 993.67 | H. Liu et al., 2018 |
| Channel catfish, *Ietalurus punctatus* | Ictaluridae | Illumina | 845.40 | 21,556 | 7.25 | 48.50 | X. Chen et al., 2016 |
| | | Illumina, Pacbio | 783.00 | 26,661 | 7.73 | 77.20 | Z. Liu et al., 2016 |
| Giant Devil Catfish, *Bagarius yarrelli* | Sisoridae | Illumina, Pacbio | 571.00 | 19,027 | 3.10 | 1600.00 | Jiang et al., 2019 |
| Walking catfish, *Clarias batrachus* | Clariidae | Illumina | 821.00 | 22,914 | 0.36 | 19.00 | N. Li et al., 2018 |
| Striped catfish, *Pangasianodon hypophthalmus* | Pangasiidae | Illumina | 700.00 | 28,600 | 14.29 | 6.00 | Kim et al., 2018 |

Table 2 BUSCO analysis results of the *L. longirostris* genome

| | Proteins | Percentage (%) |
|---|---|---|
| Complete identified BUSCOs | 4,293 | 93.6 |
| Complete Single-Copy BUSCOs | 4,109 | 89.6 |
| Complete Duplicated BUSCOs | 184 | 4.0 |
| Fragmented BUSCOs | 90 | 2.0 |
| Missing BUSCOs | 201 | 4.4 |
| Total BUSCOs searched | 4,584 | 100 |

Table 3 Classification of repeat elements in the *L. longirostris* genome

| Type | Repbase TEs | | *De novo* TEs | | TE protiens | | Combined TEs | |
|---|---|---|---|---|---|---|---|---|
| | Length (Mb) | % in genome | Length (Mb) | % in genome | Length (Mb) | % in genome | Length (Mb) | % in genome |
| DNA | 80.23 | 11.41 | 107.28 | 15.25 | 948.20 | 0.13 | 146.40 | 20.81 |
| LINE | 30.71 | 4.37 | 64.31 | 9.14 | 19.22 | 2.73 | 82.19 | 11.68 |
| SINE | 15.34 | 2.18 | 6.28 | 0.89 | 0 | 0.00 | 20.52 | 2.92 |
| LTR | 19.17 | 2.73 | 68.24 | 9.7 | 8.42 | 1.2 | 78.65 | 11.18 |
| Others | 0.03 | 0.00 | 0 | 0.00 | 0 | 0.00 | 0.03 | 0.00 |
| Unknown | 0 | 0.00 | 1.96 | 0.28 | 0 | 0.00 | 1.96 | 0.28 |
| Total | 129.07 | 18.35 | 214.87 | 30.55 | 28.53 | 4.06 | 239.11 | 33.99 |

Table 4 Statistics of predicted protein-coding genes in the *L. longirostris* genome

| Method | | Gene numbers | Average gene length (bp) | Average CDS length (bp) | Average intron length (bp) | Average exon length (bp) | Average exon per gene |
|---|---|---|---|---|---|---|---|
| Homolog | *Danio rerio* | 22,607 | 24,906.24 | 1,586.43 | 2,991.09 | 180.35 | 8.80 |
| | *Gasterosteus aculeatus* | 19,515 | 14,591.05 | 1,522.77 | 1,622.74 | 168.20 | 9.05 |
| | *Ictalurus punctatus* | 23,916 | 19,607.43 | 1,698.77 | 2,068.38 | 175.89 | 9.66 |
| | *Larimichthys crocea* | 19,859 | 21,505.49 | 1,685.09 | 2,269.19 | 173.10 | 9.73 |
| | *Oreochromis niloticus* | 19,427 | 25,587.29 | 1,664.02 | 2,771.77 | 172.78 | 9.63 |
| | *Oryzias latipe* | 19,122 | 58,907.43 | 1,603.53 | 7,358.53 | 182.48 | 8.79 |
| | *Pangasianodon hypophthalmus* | 22,691 | 20,471.44 | 1,737.03 | 2,119.68 | 176.56 | 9.84 |
| | *Tachysurus fulvidraco* | 25,662 | 19,646.64 | 1,676.25 | 2,081.63 | 174.01 | 9.63 |
| | *Takifugu rubripes* | 18,205 | 16,156.31 | 1,440.29 | 1,964.65 | 169.64 | 8.49 |
| *De novo* | Augustus | 23,758 | 14,112.19 | 1,448.53 | 1,734.99 | 174.54 | 8.30 |
| Transcript | PASA | 51,771 | | | | | |
| Merge | Glean | 23,708 | 16,546.44 | 1,792.67 | 1,547.65 | 170.20 | 10.53 |