scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Assembling chromosome-level genomes of male and female *Chanodichthys mongolicus* using PacBio HiFi reads and Hi-C technologies

Qi Liu[1], Xiaopeng Wang[1]✉, Dinaer Yekefenhazi[1], Jingyu Wang[2], Keer Zhong[2], Ying Zhang[2], Huiyun Fu[3], Zhiyong Zhou[2], Jiangfeng Huang[2], Wanbo Li[1]✉ & Xiandong Xu[2]✉

***Chanodichthys mongolicus*, a carnivorous fish belonging to the Cyprinidae family (Erythroculter), is widely distributed in reservoirs and lakes across China. However, the lack of research on whole genome assembly has impeded advancements in genetic studies for this species. In this study, we employed PacBio sequencing and Hi-C technology to assemble high-quality genomes for both female and male *Chanodichthys mongolicus* at the chromosome level. The assembly results revealed a male genome size of 1.10 GB with a scaffold N50 of 43 Mb, while the female genome was 1.09 GB with a scaffold N50 of 42 Mb. Both assemblies consist of 24 chromosomes and demonstrate an average genome integrity of 98.5%, as assessed by BUSCO. We annotated the male genome using a combination of *ab initio* predictions, protein homology comparisons, and RNAseq data, resulting in the identification of 33,581 genes, of which 88.15% were predicted to have functional roles. These findings provide a valuable resource for future research on the genetic breeding and genome evolution of *Chanodichthys mongolicus*.**

Background & Summary

Chanodichthys mongolicus is a formidable carnivorous fish belonging to the Cyprinidae family, widely distributed throughout East Asia, particularly in the reservoirs and freshwater lakes of China[1]. This species typically inhabits the middle and upper levels of slow-moving bays and lakes, and is known for its quick movements and active behavior, and does not exhibit migratory habits[1]. *C. mongolicus* is highly prized for its tender and fresh meat, rich nutrients, and its dual use in medicine and food, offering significant economic value in China. However, in recent years, wild populations have sharply declined due to environmental degradation and overfishing, bringing them close to depletion[2]. Fortunately, domestication and artificial breeding efforts have been initiated for *C. mongolicus*[3–8], which not only helpful to the local economy but also helped preserve biodiversity. Currently, researches on *C. mongolicus* predominantly focused on its meat quality[9,10], seedling transport[11], and growth characteristics[12]. Studies on its genome have primarily addressed the mitochondrial genome[13,14] and population genetics[15–19]. Nevertheless, a comprehensive genome assembly for this species remains unavailable, hindering advances in molecular breeding and evolutionary studies.

Whole genome sequencing technology encompasses both next-generation sequencing (NGS) and third-generation sequencing. NGS allows for rapid and cost-effective whole genome sequencing, while single-molecule real-time (SMRT) sequencing, a type of third-generation technology, offers read lengths of up to 30–50 kb, significantly enhancing the integrity of genome assembly. Hi-C (high-throughput chromosomal conformation capture) facilitates the study of DNA interactions across the genome, enabling the creation of

[1]Key Laboratory of Healthy Mariculture for the East China Sea, Ministry of Agriculture and Rural Affairs, Jimei University, Xiamen, 361021, China. [2]Fisheries Research Institute of Jiangxi Province, Nanchang, 330039, China. [3]Jiangxi Center for Agricultural Technical Extension, Nanchang, 330046, China. ✉e-mail: xiaopengwang@jmu.edu.cn; li.wanbo@jmu.edu.cn; xuxiandong2004@163.com

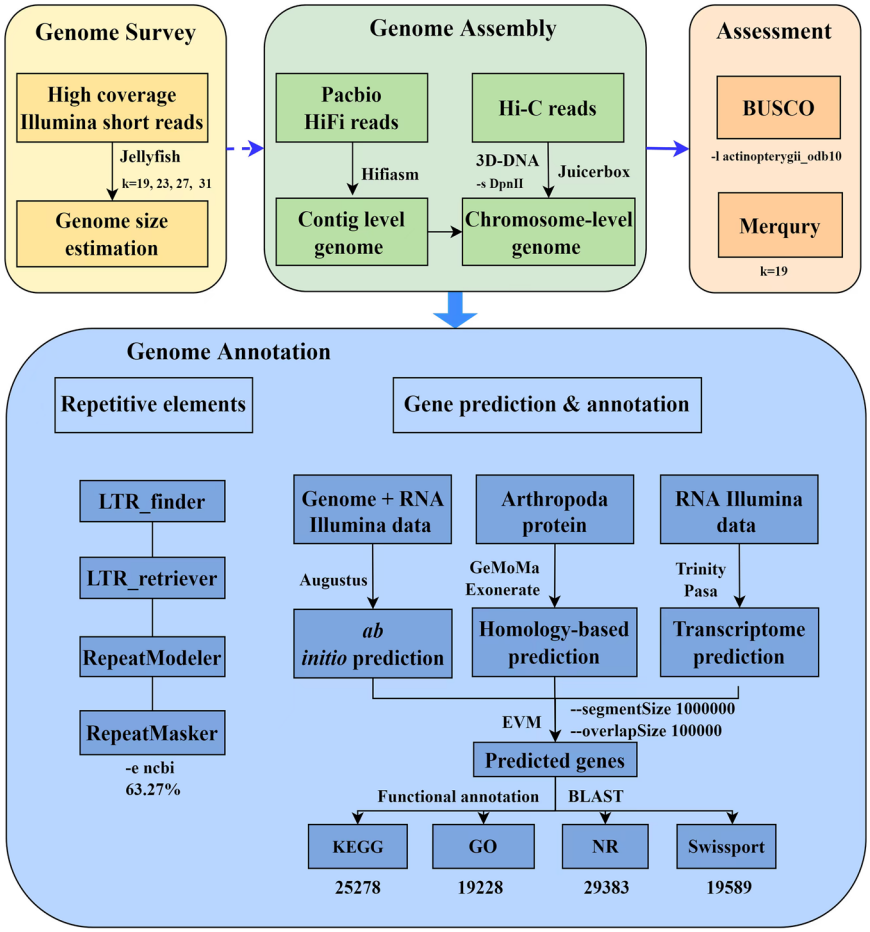

Fig. 1 Flowchart of chromosome-level genome assembly of *Chanodichthys mongolicus*.

high-resolution three-dimensional chromatin structure maps. The HiFi + Hi-C approach improves the fidelity, continuity, and completeness of genome assembly by leveraging high-fidelity long reads and chromatin interaction data, minimizing errors and biases, and enhancing structural resolution, especially in complex and large genomes[20].

In recent years, this strategy has been applied to sequence the genomes of various fish species, including *Acrossocheilus fasciatus*[21], *Spinibarbus caldwelli*[22], *Elopichthys bambusa*[23], *Plagiognathops microlepis*[24] and *Epinephelus awoara*[25]. These genomes contribute to studies in aquaculture, germplasm resource protection, and environmental adaptability. In this study, we assembled high-quality chromosome-level genomes for both female and male *Chanodichthys mongolicus* using PacBio HiFi reads and Hi-C technologies (see Fig. 1). The final assemblies yielded lengths of approximately 1,100 Mb for the male and 1,093 Mb for the female, with scaffold N50 values of 43 Mb and 42 Mb, respectively. The resulting assemblies comprised 24 chromosomes, covering about 99.25% of the genome size after Hi-C integration.

This represents the first genome assembly of *Chanodichthys mongolicus*, providing a valuable genetic foundation for future research in genetic breeding and evolutionary analysis for this species.

Methods

Sampling and genome sequencing. We collected muscle and fin tissues from a male (M) and a female (F) *Chanodichthys mongolicus*, preserving them in liquid nitrogen and a 75% alcohol solution, respectively. Genomic DNA was extracted from the fin tissue using the Fast Pure Cell/Tissue DNA Isolation Mini Kit (Vazyme, Nanjing, China) according to the manufacturer's instructions. The short-read sequencing was then performed on an Illumina NovaSeq platform, following the manufacturer's protocols. This generated 25.4 Gb data for the male and 30.6 Gb for the female, with 150 bp paired-end reads (Table 1).

Genomic DNA was isolated from muscle using the phenol/chloroform method for long-read sequencing, we then prepared SMRTbell libraries using high-quality DNA (Novogene, Beijing, China). This high-quality DNA was digested using the restriction endonuclease DpnII to produce DNA fragments with sticky ends. Following end repair and phosphorylation, the DNA fragments were appended with SMRTbell junctions, which include PacBio adapters and primer binding sites. DNA ligase then sealed the complementary junction halves, forming the closed-loop SMRTbell DNA molecule. After verifying the quality of the library, it was sequenced on the PacBio Sequel-Revio platform, utilizing the Sequel Binding Kit 3.0 (Pacific Biosciences, Menlo Park, CA, USA). The long-read sequencing produced 30.3 Gb of HiFi reads for the male and 35.9 Gb for the female (Table 1).

Type	Sample	Platform	Data (Gb)(M)	Data (Gb)(F)
HiFi	Muscle	Sequel-Revio	30.3	35.9
Hi-C	Muscle	Illumina NovaSeq	88.6	104.6
DNAseq	Fin	Illumina NovaSeq	25.4	30.6
RNAseq	Pooled	Illumina NovaSeq	36.2	36.2

Table 1. Statistics of the sequencing data for *C. mongolicus* genome assembly.

Hi-C sequencing. The Hi-C library is constructed using muscle tissue according to standard protocols and has undergone some modifications. Initially, 4% formaldehyde was used for sample fixation to enable DNA-protein crosslinking. Following cell lysis, DNA was digested with Mbol I restriction enzyme to generate fragments with sticky ends. These fragments were modified with sequencing junctions through end repair, A-tailing, and ligation using T4 DNA ligase. The junction-labeled DNA fragments were then purified by isolating biotinylated Hi-C samples with streptavidin-coated magnetic beads. The Hi-C sequencing library was sequenced on the Illumina NovaSeq platform. In the end, sequencing yielded 88.6 Gb and 104.6 G data for M and F, respectively (Table 1).

RNA-seq sequencing. Total RNA was extracted from muscle, liver, spleen, intestines, testis, and ovaries of male and female *C. mongolicus* using TRIzol Reagent (Invitrogen, MA, USA) according to the manufacturer's instructions. RNA samples from various tissues were pooled in equal amounts. mRNA enrichment was performed using oligo(dT)-coated magnetic beads. The enriched mRNA was then fragmented and reverse transcribed into first-strand cDNA using random hexamers. The second strand of cDNA was synthesized by adding buffer, dNTPs, RNase H, and DNA polymerase I. Following cDNA purification, the ends were repaired and A-tailing was performed before ligating sequencing adapters. Target size fragments were recovered through agarose gel electrophoresis, and PCR amplification was conducted to complete the preparation of the pooled RNA library. The constructed library was sequenced on the Illumina NovaSeq platform, producing 36.2 Gb of RNA-seq data (Table 1).

Genome size estimation. We estimated genome size using Jellyfish[26] (v2.3.0) counts. By setting 19, 23, 27, and 31 k-mer parameters and utilizing high-coverage short reads, we calculated the total number of k-mers. The genome size was determined by dividing the total k-mer count by the number of copies. The estimated genome sizes ranged from 1.05 to 1.11 Gb for the male (M) and from 1.07 to 1.10 Gb for the female (F) (Table 2).

Contig assembly. We assembled the contig-level of genome using the default parameters of Hifiasm[27] (v0.13.0-R307). Utilizing the HiFi data from the male and female, we obtained two assemblies of 1.097 Gb with N50 of 34.6 Mb for M, and 1.092 Gb with N50 of 34.4 Mb for F (Table 3). The assembly sizes are consistent with the k-mer estimates.

Chromosome-level genome assembly using Hi-C data. To achieve chromosome-level genome assembly, we incorporated Hi-C data into the genome construction process. The raw Hi-C data was first processed with HICUP[28] (v0.8.1) to generate valid Hi-C data for downstream analysis. This valid Hi-C data was then integrated into the scaffold assemblies using the default parameters of Hifiasm. We then utilized 3D-DNA[29] (v180419) to generate the chromosome-level genome, followed by manual corrections using Juicerbox[30] (v1.11.08). After manual correction, the genomes were reassembled using the run-asm-pipeline-post-review.sh script of 3D-DNA. The final results, labeled with "FINAL," represent the completed chromosome-level genomes.

We obtained a male and a female genome, each consisting of 24 chromosomes, with a Hi-C data loading rate of approximately 99.25% (Fig. 2). The male genome measures 1,092 Mb and consists of 58 scaffolds, while the female genome measures 1,093 Mb and consists of 98 scaffolds (Table 3).

Assessment of the genome assemblies. The Benchmarking Universal Single-Copy Orthologues (BUSCO)[31] (v5.1.2) was empolyed to assesse genomic completeness. A total of 3,640 orthologue genes in the assemblies of *Chanodichthys mongolicus* were identified via a single-copy direct homologous gene database (actinopteryGIi_ODb10, https://busco-data.ezlab.org). In the male assembly, 3,585 complete genes (C; 98.50% of the genome) were detected, including 3,517 complete and single-copy (S; 96.60% of C). The rest were duplicates (D; 1.90%), fragmented (F; 0.50%), and missing genes (M; 1.00%). For the female assembly, we identified 3,586 complete genes (C; 98.50% of the genome), with 3,537 complete and single-copy genes (S; 97.20% of C). The percentages of D, F, and M were 1.30%, 0.60%, and 0.90%, respectively (Table 4).

Additionally, we evaluated the accuracy and completeness of the assembled genomes using Merqury[32]. High-coverage Illumina short reads were employed to produce optimal k-mers, yielding consensus quality values (QV) and completeness metrics for the assemblies. The QV for the male assembly was 42.89, with a completeness of 94.94%. For the female assembly, the QV was 44.25, with a completeness of 94.75%.

Repeat annotation. We employed LTR_FINDER[33] (v1.0.7), LTR_retriever[34] (v2.9.0), RepeatModeler[35] (v1.0.8) and RepeatMasker[36] (v4.0.5) to generate *ab initio* repetitive sequences for the male genome. We initially detected full-length LTR retrotransposons in the *C. mongolicus* genome using LTR_FINDER with default parameters. Based on the output, we utilized LTR_retriever to identify LTR retrotransposons and generate a non-redundant LTR-RT library. The BuildDatabase function in RepeatModeler was then used to create a database

Sample	M				F			
K-mer (bp)	19	23	27	31	19	23	27	31
Total Nod	10,001	9,999	9,998	9,992	10,001	10,001	10,001	9,999
Total K-mers	20,483,991,649	19,916,028,877	19,358,673,523	18,723,043,939	24,333,552,542	23,676,218,765	22,972,216,802	22,215,022,542
Peak	19	18	18	17	23	22	21	20
Estimated size (Mb)	1,078	1,106	1,075	1,101	1,058	1,076	1,094	1,111
Single copy	694,238,169	753,965,234	749,189,566	800,684,119	663,790,161	737,826,691	788,930,523	824,561,294
Proportion	0.64	0.68	0.70	0.73	0.63	0.69	0.72	0.74

Table 2. Estimation of genome size using Jellyfish counts.

Sample	M		F	
Type	Contig (kb)	Scaffold (kb)	Contig (kb)	Scaffold (kb)
Number	164	58	197	98
N10	58,325	62,424	58,264	61,254
N50	34,599	42,303	34,440	43,064
N90	13,127	36,061	13,030	34,523
Max length	58,703	83,280	61,071	90,105
Total length	1,097,893	1,100,778	1,092,363	1,092,928

Table 3. Assembly statistics at the contig level and scaffold level for *C.mongolicus*.

for duplicate sequence identification. We merged the resulting libraries to create the final library, which was input into RepeatMasker to identify repeat sequences. Additionally, we applied RepeatMasker with the parameters "-e ncbi." The outputs were combined to obtain the genome after masking the repetitive sequences, along with the repetitive sequence annotation file and the statistics file. The analysis indicated that the male genome of *C. mongolicus* contains 63.27% repetitive sequences, including 36.3% DNA elements, 9.87% long terminal repeats (LTRs), and 3.22% long interspersed nuclear elements (LINEs). Compared to other carps, the genome of *C. mongolicus* has a higher percentage of repeats than *Ctenopharyngodon idellus* (38.06%), and is slightly higher than *Danio rerio* (52.2%) and *Chanodichthys erythropterus* (51.34%)[37].

Gene prediction and annotation. We used three strategies for annotating the genes of the male genome: *ab initio* prediction, homology-based prediction and RNA-Seq prediction. Augustus[38] (v3.2.3) was used for *ab initio* gene prediction. For homology-based prediction, protein sequences of 8 species (*Cyprinus_carpio, Sinocyclocheilus rhinocerous, Sinocyclocheilus grahami, Danio_rerio, Oryzias latipes, Gadus morhua, Oreochromis niloticus*) were compared using GeMoMa (v1.9)[39]. Exonerate[40] (v2.2.0) was then employed to select 10 kb of sequences surrounding the best comparison gene for exon region prediction. For RNA-Seq prediction, Trinity[41] (v2.4.0) assembled the transcriptome, and PASA[42] (v2.1.0) aligned RNA-seq data to the genome, producing redundant transcript sequences, transcript-genome alignment results, and variable splicing information, as well as predict open reading frame (ORF) regions. We employed Tophat[43] to align transcriptome sequencing data and subsequently used Cufflinks[44] (v2.1.1) for transcriptome prediction. Finally, Evidence Modeler (EVM)[45] (v1.1.1) combined the gene prediction results, with PASA updating EVM predictions to include UTR and splicing annotations. A total of 33,581 putative genes were predicted in the *Chanodichthys mongolicus* genome. Functional annotation of these genes was performed by aligning them with the SwissProt[46], NCBI nonredundant protein (Nr), KEGG[47], and GO[48] databases using BLAST + (v2.11.0) software (e-value $\leq 1e^{-5}$)[49]. As a result, a total of 29,602 genes (88.15%) were successfully annotated (Table 5).

Data Records

The female genomic Illumina sequencing data were deposited in the SRA at NCBI SRX24981704[50].
The male genomic Illumina sequencing data were deposited in the SRA at NCBI SRX24981703[51].
The female genomic PacBio sequencing data were deposited in the SRA at NCBI SRX24981712[52].
The male genomic PacBio sequencing data were deposited in the SRA at NCBI SRX24981711[53].
The transcriptomic sequencing data were deposited in the SRA at NCBI SRX24981715[54].
The female Hi-C sequencing data were deposited in the SRA at NCBI SRX24981714[55].
The male Hi-C sequencing data were deposited in the SRA at NCBI SRX24981713[56].
The male chromosome assembly was deposited in GenBank at NCBI JBEUTB000000000[57].
The female chromosome assembly was deposited in GenBank at NCBI JBEUTA000000000[58].
The genome annotation file is available in figshare[59].

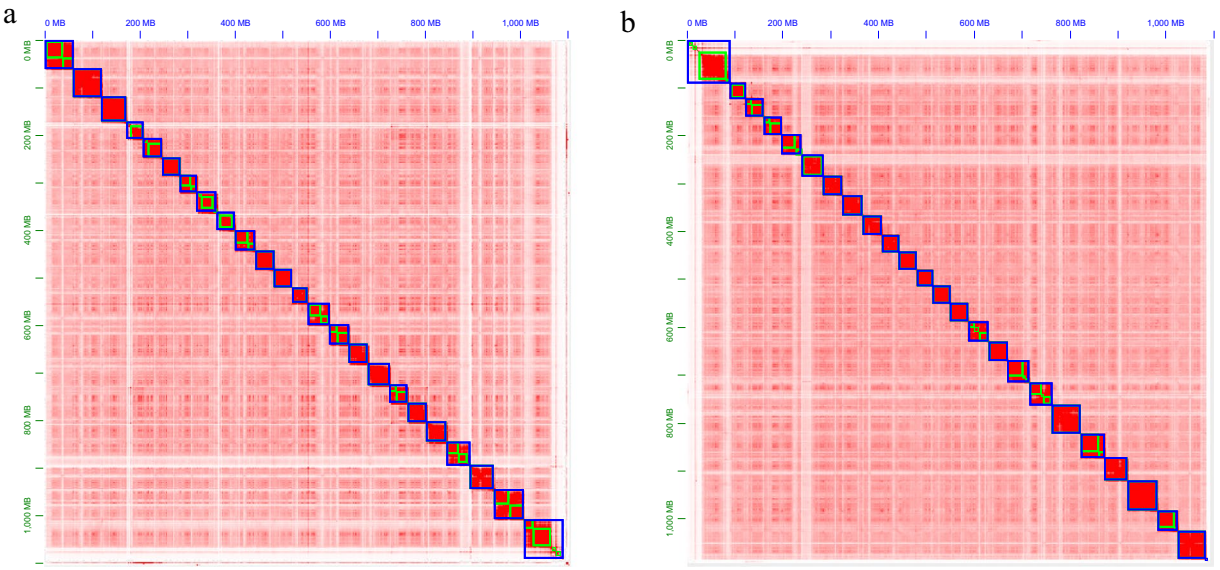

Fig. 2 Overview of the genome assemblies for *Chanodichthys mongolicus*. (**a**) Heatmap of chromosomal genome assembly for the male. (**b**) Heatmap of chromosomal genome assembly for the female. In both heatmaps, the blue boxes represent chromosomes, where the green boxes indicate scaffolds. Both the male and female genomes have 24 chromosomes.

Sample	M		F	
Type	**Number**	**Percentage (%)**	**Number**	**Percentage (%)**
Complete BUSCOs (C)	3,585	98.50%	3,586	98.50%
Complete and single-copy BUSCOs (S)	3,517	96.60%	3,537	97.20%
Complete and duplicated BUSCOs (D)	68	1.90%	49	1.30%
Fragmented BUSCOs (F)	19	0.50%	21	0.60%
Missing BUSCOs (M)	36	1.00%	33	0.90%
Total BUSCO groups searched	3,640	\	3,640	\

Table 4. Results of BUSCO assessment.

Database	Gene number	Percentage (%)
KEGG	25,278	75.27
GO	19,228	57.26
NR	29,383	87.50
Swissport	19,589	58.33
All Annotated	29,602	88.15
Predicted Genes	33,581	

Table 5. Statistics of functional annotation.

Technical Validation

The DNA samples prepared for PacBio sequencing were validated by agarose gel electrophoresis, exhibiting a prominent band exceeding 20 kb. DNA concentration was quantified with a Qubit Fluorometer (Thermo Fisher Scientific, USA), and the 260/280 nm absorbance ratio was measured at 1.804 with a NanoDrop ND-1000 spectrophotometer (LabTech, USA).

For RNA sequencing, total RNA was isolated with TRIzol reagent (Invitrogen, MA, USA), adhering to the supplier's instructions. The integrity of the RNA was evaluated on an Agilent 2100 Bioanalyzer (Agilent Technologies, CA, USA). The RNA sample selected for this investigation presented an RNA Integrity Number (RIN) exceeding 9.0.

Code availability

No specific code was used in this study. Data analyses were performed using standard bioinformatics tools as described in the methods.

Received: 14 November 2024; Accepted: 1 May 2025;
Published online: 06 June 2025

References

1. Chen, Y. Y. *Fauna sinica, Osteichthyes. Cypriniformes* **3**, 40–49 (1998).
2. Xia, C. & Jiang, Z. The population characteristics of *Chanodichthys mongolicus* in Jingbo Lake and its impact on free range fish species. *Freshwater fisheries* **23**, 13–16 (1993).
3. Xie, J., Yan, Y., Yang, Y. & Lin, S. Analysis on genetic structure of *Chanodichthys mongolicus* populations by mitochondrial COI gene sequences. *Freshwater Fisheries* **49**, 3–7 (2019).
4. Yan, B., Xiong, C., Jin, F. & Du, G. Artificial breeding technology of Mongolian culter. *Aquaculture* **36**, 37–38 (2015).
5. Xu, W. *et al*. Preliminary study on artificial breeding technology of Mongolian culter (*Chanodichthys mongolicus*) in Jingpo Lake. *Freshwater Fisheries* **39**, 63–66 (2009).
6. Huai, X. J. *et al*. Parent fish cultivation and artificial breeding technology of Mongolian culter (*Chanodichthys mongolicus*). *Shanghai Agricultural Science and Technology* **6**, 3 (2014).
7. Jiang, H. F. *et al*. Artificial propagation and observation of embryonic and postembryonic development in pond-farmed Mongolian culter (*Chanodichthys mongolicus*) from Jingpo Lake. *Fisheries Science* **35**, 130–135 (2016).
8. Shi, Q., *et al*. *Economically Important Fishes In China*. *Huazhong University of Science & Technology Press* **1**, 29-30 (2015).
9. Yu, H. *et al*. Meat quality analysis of three culter species in Liangzi Lake. *Acta Hydrobiologica Sinica* **29**, 502–506 (2005).
10. Chen, Q. H. Seasonal variation of digestive enzyme activity and muscle nutritional composition of four culter species in Xingkai Lake. *Shanghai Ocean University* (2011).
11. Lin, M. *et al*. Effects of Two Anesthetics on Survival of Juvenile *Culter mongolicus* during a Simulated Transport Experiment. *North American Journal of Aquaculture* **74**, 137–146 (2012).
12. Zhang, X. G., Ruan, Z. J. & Xiong, B. X. Age and growth characteristics of *Chanodichthys mongolicus* in Poyang Lake. *Transactions of Oceanology and Limnology* **3**, 137–143 (2008).
13. Tong, G., Kuang, Y., Geng, L., Xu, W. & Yin, J. Mitochondrial DNA sequence of Mongolian redfin (*Chanodichthys mongolicus*). *Mitochondrial DNA* **25**, 407–409 (2014).
14. Saitoh, K. *et al*. Mitogenomic Evolution and Interrelationships of the Cypriniformes (Actinopterygii: Ostariophysi): The First Evidence Toward Resolution of Higher-Level Relationships of the World's Largest Freshwater Fish Clade Based on 59 Whole Mitogenome Sequences. *Journal of Molecular Evolution* **63**, 826–841 (2006).
15. Liu, K., Feng, X. Y., Ma, H. J. & Xie, N. Comparative mitochondrial genome analysis of the Mongolian redfin, *Chanodichthys mongolicus* (Xenocyprididae) from China reveals heteroplasmy. *Mitochondrial DNA Part B* **6**, 2601–2604 (2021).
16. Liu, K., Feng, X., Ma, H. & Xie, N. Development and characterization of 13 microsatellite markers of *Chanodichthys mongolicus* (Cypriniformes: Cyprinidae) by RAD-seq. *Journal of Applied Ichthyology* **37**, 975–979 (2021).
17. Liu, K. *et al*. Genetic structure analysis of *Megalobrama terminalis*, *Culter alburnus*, *Chanodichthys mongolicus* and their hybrids based on genotyping by sequencing. *Journal of Fisheries of China* **45**, 1307–1316 (2021).
18. Miao, C. Q. & Han, Y. Genetic diversity analysis of *Chanodichthys mongolicus* populations in three regions based on mtDNA d-Loop gene sequences. *Heilongjiang Animal Science and Veterinary Medicine* **11**, 17–22 (2016).
19. Xiong, Y. *et al*. Genetic structure and demographic histories of two sympatric Culter species in eastern China. *Journal of Oceanology and Limnology* **38**, 106–147 (2020).
20. Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. C. & Hunkapiller, M. W. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature Biotechnology* **37**, 1155–1162 (2019).
21. Zheng, J. *et al*. Chromosome-level genome assembly of *Acrossocheilus fasciatus* using PacBio sequencing and Hi-C technology. *Scientific Data* **11**, 166 (2024).
22. Wu, L. *et al*. Chromosome-level genome assembly and annotation of the *Spinibarbus caldwelli*. *Scientific Data* **11**, 933 (2024).
23. Li, S. *et al*. Chromosome-level genome assembly of the yellow-cheek carp *Elopichthys bambusa*. *Scientific Data* **11**, 426 (2024).
24. Wu, Y., Sha, H., Luo, X., Zou, G. & Liang, H. Chromosome-level genome assembly of *Plagiognathops microlepis* based on PacBio HiFi and Hi-C sequencing. *Scientific Data* **11**, 802 (2024).
25. Zhang, W. *et al*. Chromosome-level genome assembly and annotation of the yellow grouper, *Epinephelus awoara*. *Scientific Data* **11**, 151 (2024).
26. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinforma. Oxf. Engl.* **27**, 764–770 (2011).
27. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 1–6 (2021).
28. Steven, W. *et al*. HiCUP: pipeline for mapping and processing Hi-C data. *F1000res* **4**, 1310 (2015).
29. Dudchenko, O. *et al*. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92 (2017).
30. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
31. Manni, M., Berkeley, M. R., Mathieu, S., Simo, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral Genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
32. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
33. Zhao, X. & Hao, W. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* **35**, W265–268 (2007).
34. Ou, S. & Jiang, N. LTR_retriever: A Highly Accurate And Sensitive Program For Identification of Long Terminal Repeat Retrotransposons. *Plant Physiology* **176**, 2 (2017).
35. Flynn, J. M., Hubley, R., Goubert, C., Rosen, J. & Smit, A. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA*. **117**, 9451–9457 (2020).
36. Ou, S. *et al*. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **20** (2019).
37. Zhao, S. *et al*. A chromosome-level genome assembly of the redfin culter (*Chanodichthys erythropterus*). *Scientific data* **9**, 535, https://doi.org/10.1038/s41597-022-01648-0 (2022).
38. Lange *et al*. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**, 767–769 (2016).
39. Jens, K. *et al*. Using intron position conservation for homology-based gene prediction. *Nucleic Acids Research* **9**, 89 (2016).
40. Slater, G. S. C., Birney, E., Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005). *BMC Bioinformatics* **6**, 31.
41. Grabherr, M. G., Haas, B. J., Yassour, M. & Levin, J. Z. & others. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnology* **29**, 644 (2013).

42. Haas, B. J. *et al*. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
43. Trapnell, C., Goff, R. A., Pertea, L., Kim, G. & Kelley, D. DR. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7**, 562–578 (2012).
44. Ghosh, S. & Chan, C.-K. K. Analysis of RNA-Seq Data Using TopHat and Cufflinks. *Methods in molecular biology.* **1374**, 339–361 (2016).
45. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
46. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic acids research* **28**, 45–48 (2000).
47. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research* **28**, 27–30 (2000).
48. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. *Nature genetics* **25**, 25–29 (2000).
49. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology* **215**, 403–410 (1990).
50. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRX24981704 (2024).
51. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRX24981703 (2024).
52. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRX24981712 (2024).
53. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRX24981711 (2024).
54. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRX24981715 (2024).
55. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRX24981714 (2024).
56. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRX24981713 (2024).
57. *NCBI Genbank* https://identifiers.org/ncbi/insdc.gca:GCA_040802225.1 (2024).
58. *NCBI Genbank* https://identifiers.org/ncbi/insdc.gca:GCA_040802255.1 (2024).
59. Liu, Q. & Li, W. MG1.finally.gff3. *figshare* https://doi.org/10.6084/m9.figshare.27601491.v1 (2024).

Acknowledgements

This work was financially supported by the earmarked fund for Jiangxi Agriculture Research System (No. JXARS-03), the indigenous fish germplasm resources excavating project of Jiangxi Province (No. 2022-03) and Jimei University (grant number ZQ2022022).

Author contributions

X.X. and W.L. conceived of the project. Q.L., D.Y., J.W., K.Z., Y.Z., H.F., Z.Z., J.F. collected the samples and extracted the genomic DNA and RNA. Q.L., X.W., W.L. and X.X. performed the data analysis and wrote the manuscript. D.Y. contributed to the data analyses. X.W. and X.X. revised the manuscript. All authors read and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to X.W., W.L. or X.X.