scientific data

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of the *Lepturichthys fimbriata*

Cheng Wang[1,4], Weitao Chen[2,4], Wenjun Chen[1], Zufa Ding[1], Liandong Yang[1]✉ & Ying Wang[3]✉

Lepturichthys fimbriata **is a small benthic fish inhabiting rapid rocky shoals within the Jinsha River system. This species holds considerable significance for studying animal morphology, ecology, and evolution due to its specialized habitat adaptations. We generated the first chromosome-level genome assembly of *L. fimbriata* through integrated HiFi, Illumina, and Hi-C sequencing. The 591 Mb genome features a scaffold N50 of 20.9 Mb, with 94.2% (556.18 Mb) of sequences anchored to 25 pseudo-chromosomes (2n = 50). Genome annotation identified 36.01% (212.72 Mb) repetitive elements and 21,320 protein-coding genes, with assembly completeness evidenced by a 97.3% BUSCO score. This high-quality genome provides a crucial resource for investigating *L. fimbriata* genetic diversity and ecological adaptation mechanisms.**

Background & Summary

Lepturichthys fimbriata belongs to the genus *Lepturichthys* in the family Balitoridae of Cypriniformes. It is a small benthic fish living in rapids and rocky shoals. It is distributed in the Jinsha River water system[1]. It mainly attaches itself to the surface of gravel in rapids with its broad and flattened paired fins and flat and bare chest and abdomen, and scrapes sessile algae and small invertebrates with its sharp keratinized lower jaw. This species spawns in rapids and gorges with extreme water-level fluctuations within the Jinsha River (upper Yangtze above Yibin) and its tributary, the lower Yalong River. Eggs and larvae drift downstream, with larvae dispersing as far as the middle Yangtze River below Yichang[2–4].

While dam construction provides human benefits, it threatens fish survival. Critically, Yangtze River dams lacking fishways (e.g., Three Gorges Dam, Gezhou Dam) severely disrupt *L. fimbriata* habitats and spawning, potentially endangering the species[2,5–7]. Current genetic resources for *L. fimbriata* remain limited to mitochondrial sequences and microsatellite markers[2,8,9], lacking a complete nuclear genome. The advancement of high-throughput sequencing technologies, particularly single-molecule real-time (SMRT) sequencing, has enabled rapid expansion of fish genomic resources in recent years[10,11]. These genomic datasets will provide a core database for research on the adaptive evolution of fish and the innovation of aquatic germplasm.

Here, we present a chromosome-scale *Lepturichthys fimbriata* genome assembly generated through an integrative approach combining Illumina short-read sequencing, PacBio HiFi long-read sequencing, and Hi-C chromatin interaction mapping. This genome will facilitate resource protection, genetic selection, and ecological adaptation studies of *L. fimbriata*.

Methods

Sample collection, preparation and sequencing. An adult fish was obtained in 2024 from Jinsha river, China. The sampled fish in this study was permitted by the Ethics Committee of the Institute of Hydrobiology, Chinese Academy of Science. The tissues including muscle, liver, heart, fin, and eye were collected and preserved in liquid nitrogen until the extraction of DNA and RNA. The muscles tissues were utilized for DNA sequencing to assembled the genome, while other tissues were utilized for RNA sequencing.

[1]State Key Laboratory of Breeding Biotechnology and Sustainable Aquaculture, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, 430072, China. [2]Pearl River Fisheries Research Institute, Chinese Academy of Fishery Sciences, Guangzhou, 510380, China. [3]Hubei Engineering Research Center for Protection and Utilization of Special Biological Resources in the Hanjiang River Basin, College of Life Sciences, Jianghan University, Wuhan, 430056, China. [4]These authors contributed equally: Cheng Wang, Weitao Chen. ✉e-mail: yangliandong1987@163.com; xinyuanwangying@163.com

Library type	Tissue	Raw data (Gb)
WGS Illumina	Muscle	22
PacBio HiFi	Muscle	11
Hi-C	Liver	51
RNA-seq	Fin	4.8
RNA-seq	Heart	5.4
RNA-seq	Eye	4.2

Table 1. Statistics of sequencing data for *L. fimbriata* genome assembly and annotation.

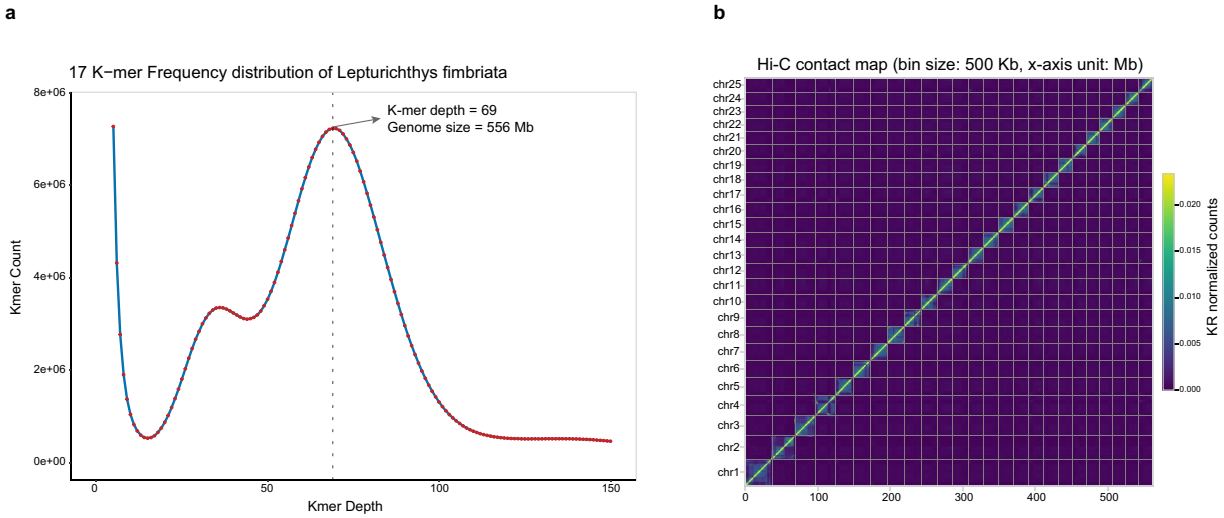

Fig. 1 Genome assembly of the *L. fimbriata*. (**a**) A 17-mer distribution of the Illumina short reads for genome size estimation. The x-axis represents the sequencing depth of each unique 17-mer, and the y-axis represents the percentage of unique 17-mers. The peak depth was confirmed at 69. (**b**) The Hi-C heatmap illustrates the interaction frequencies among various chromosomes in *L. fimbriata* (2n = 50).

We constructed and sequenced a SMRT cell sequencing library containing fragments approximately 15–20 kb in size on a PacBio Sequel II platform. We prepared second-generation sequencing libraries using Illumina technology (Illumina, San Diego, USA) with 350 bp inserts. We generated RNA sequencing libraries with an NEBNext® Ultra™ RNA Library Prep Kit (Illumina®) following the manufacturer's recommendations and sequenced RNA on an Illumina platform (Wuhan Benagen Technology Co., Ltd.), yielding 150 bp paired-end reads. We prepared Hi-C libraries from liver tissue and conducted sequencing on an Illumina NovaSeq 6000 platform. As a result, 22 Gb of Illumina read data, 11 Gb of PacBio HiFi read data, 14.4 Gb of RNA read data, and 51 Gb of Hi-C read data were obtained (Table 1).

Genome size estimation and *de novo* assembly. Prior to *de novo* genome assembly, we conducted a genome survey to estimate genome size using Illumina short-read data analyzed through jellyfish (v2.3.0)[12] and GCE (v1.0.2)[13] software with a k-mer size of 17. This k-mer analysis estimated the *L. fimbriata* genome size at 556 Mb (Fig. 1a).

The *de novo* assembly pipeline initiated with HiFi long-read processing in hifiasm (v0.18.5)[14] under default parameters, followed by systematic error correction using SMRT Link's gcpp (v11.0) and three rounds of Racon (v1.5.0)[15] polishing. Sequence redundancy was minimized through Purge_haplotigs (v1.1.2)[16] optimization, with residual gaps resolved using TGS-GapCloser[17] (v1.2.1). Chromosomal scaffolding was achieved through Hi-C data integration, where clean paired-end reads were aligned with BWA-MEM[18] (v0.7.17), processed through the Juicer[19] (v1.6)-3D-DNA[20] (v180922) pipeline, and manually curated using Juicebox Assembly Tools[21] (v2.20.00).

The *L. fimbriata* genome assembly spans 591 Mb with a scaffold N50 of 20.9 Mb, of which 94.2% was successfully anchored into 25 chromosomes (2n = 50) (Fig. 1b). The assembly exhibits a GC content of 39.5% and demonstrates exceptional completeness through BUSCO[22] assessment (actinopterygii_odb10, n = 3,640), with 97.3% of conserved genes identified as complete (94.7% single-copy, 2.6% duplicated) (Fig. 2). These metrics collectively validate the chromosomal integrity and high contiguity of this chromosome-scale genome resource.

Repetitive element and protein-coding genes annotation in the *L. fimbriata* genome. Repetitive element annotation was conducted through an integrated computational pipeline combining *de novo* repeat library construction with curated database searches. A custom repeat library was first generated using

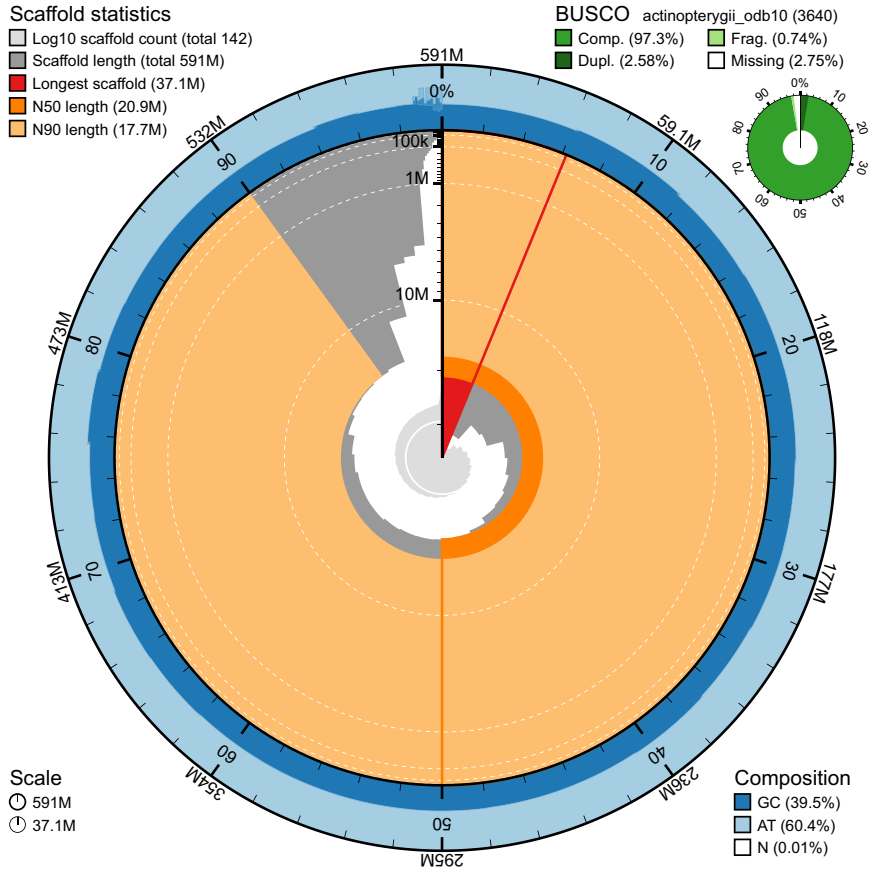

Fig. 2 Snail plot summary of the assembly statistics and Busco scores for the *L. fimbriata* genome. Dark grey arcs display record lengths in descending size order, with plot radius scaled to the longest assembly record (red). N50 and N90 record lengths are represented by dark and pale orange arcs, respectively. A pale grey spiral indicates cumulative record counts on a logarithmic scale, with white scale lines marking orders of magnitude. The outer ring (blue/pale-blue) shows GC, AT, and N content distributions binned identically to the inner plot. BUSCO scores (top right) were assessed against the actinopterygii core gene set (odb10; n = 3,640).

Type	Length (bp)	Percentage (%)
SINE	2,691,554	0.46
LINE	41,003,810	6.94
DNA	56,850,564	9.62
LTR	37,427,498	6.34
Unclassified	40,247,782	6.81
Others	25,909,044	5.84
Total	212,716,407	36.01

Table 2. Statistics on transposable elements in *L. fimbriata* genome.

RepeatModeler[23] (v2.0.4) under default parameters. Subsequent integration of this species-specific library with the Dfam database enabled comprehensive transposable element (TE) detection via RepeatMasker[24] (v4.1.5). The analysis revealed 212.72 Mb of repetitive sequences (36.01% genome coverage), with TE composition dominated by DNA transposons (9.62%), followed by LINE retrotransposons (6.94%), LTRs (6.34%), and SINEs (0.46%) (Table 2).

Gene annotation was performed through an integrative approach combining *ab initio* prediction, homology-based inference, and transcriptomic evidence. *Ab initio* gene models were generated using Augustus[25] (v3.4.0) with parameters trained through PASA[26] (v2.5.2)-processed RNA sequencing data. For homology-based annotation, protein sequences from six reference species (*Danio rerio*, *Oryzias latipes*, *Triplophysa dalaica*, and *Misgurnus anguillicaudatus, Takifugu flavidus, Chanos chanos*) were aligned to the genome using Exonerate (v2.4.0) and GenomeThreader[27] (v1.7.3) with default splice-site parameters. Transcript-based evidence was derived from StringTie[28] (v2.2.1)-assembled isoforms, which were structurally refined through iterative PASA analysis. EvidenceModeler[29] (v1.1.1) integrated these multi-source predictions using weighted evidence support

		Number	Percent (%)
Total		21,320	
Annotated	EggNog	18,154	85.15
	Swissprot	16,709	78.37
	TrEMBL	19,672	92.27
	NR	19,688	92.35
	InterPro	19,827	93
Unannotated		928	4.35

Table 3. Summary of functional annotations for predicted genes.

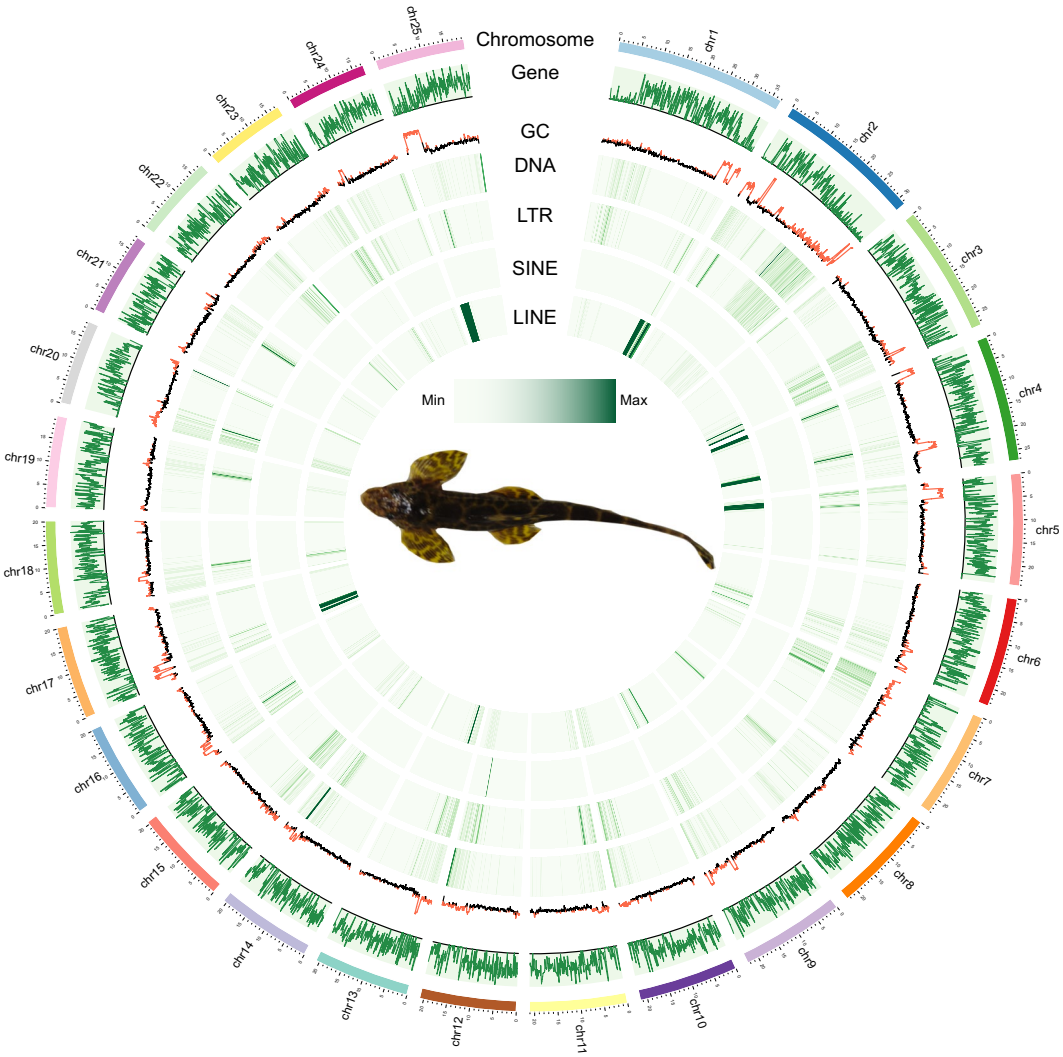

Fig. 3 Genomic landscape of the *L. fimbriata*. The rings, from the outermost to the innermost layer, represent the chromosome of the *L. fimbriata* genome, gene density, GC content, DNA transposons, LTRs, SINEs, and LINEs. The analysis was conducted using 100-kb genomic sliding windows.

(weights: transcript = 8, homology = 6, *ab initio* = 1), followed by PASA-mediated incorporation of alternative splicing variants and UTR annotations to produce the final gene set.

For functional annotation of predicted genes, we aligned the predicted genes to the NR (non-redundant protein database), eggNOG, Swiss-Prot, TrEMBL, and InterPro databases using BLASTP[30]. In total, we successfully predicted 21,320 protein-coding genes within the genome. Further, 20,392 genes, which accounts for 95.65% of the total number of predicted genes, were successfully assigned with at least one functional annotation (Table 3). Finally, chromosomal distributions of gene density, GC content, and repeat sequence density were plotted across all 25 *L. fimbriata* chromosomes. (Fig. 3).

Data Records

All the raw sequencing data have been deposited in the NCBI database with accession number SRP590038[31] under BioProject accession number PRJNA1272245. The whole genome shotgun project has been deposited at GenBank under the accession JBPGZU000000000[32]. Furthermore, files of the assembled genome, gene annotation were deposited in Figshare database[33].

Technical Validation

The BUSCO was used to evaluate the quality of the genome assembly. We assessed assembly completeness using BUSCO[22] (v4.1.2) with the reference actinopterygii gene set (n = 3,640). The final genome assembly showed a BUSCO completeness of 97.3%, consisting of 3,446 (94.7%) single-copy BUSCOs, 94 (2.6%) duplicated BUSCOs, 27 (0.7%) fragmented BUSCOs, and 73 (2%) missing BUSCOs. Comparison of BUSCO results with *Gymnocypris eckloni*[34] (87.5%) and *Beaufortia pingi*[35] (97.0%) revealed the high genome assembly quality of *L. fimbriata*. Furthermore, we remapped Illumina paired-end short reads and HiFi long reads to the final assembled genome using BWA[18] (v0.7.17) and minimap2[36] (v2.28), resulting in mapping rates of 99.48% and 99.92%, respectively.

Data availability

The raw sequencing datasets has been deposited to NCBI (PRJNA1272245), the assembled genome of *L. fimbriata* was deposited at GenBank (JBPGZU000000000), and the assembled genome and gene annotation has been deposited to figshare database (https://doi.org/10.6084/m9.figshare.29243024).

Code availability

No specific code was used in this study. The data analyses adhered to the manuals and protocols offered by the creators of the corresponding bioinformatics tools, the parameter settings of which were outlined in the methods section.

Received: 30 June 2025; Accepted: 22 October 2025;
Published online: 28 November 2025

References

1. Ding, R. The fishes of Sichuan, China. *Chengdu: Sichuan Publishing House of Science and Technology* **397**, 406 (1994).
2. Zhang, F., Duan, Y., Cao, S., Wang, J. & Tan, D. High genetic diversity in population of Lepturichthys fimbriata from the Yangtze River revealed by microsatellite DNA analysis. *Chinese Science Bulletin* **57**, 487–491, https://doi.org/10.1007/s11434-011-4704-8 (2012).
3. Tang, Q.-Y., Yu, D., Yang, J.-Q. & Liu, H.-Z. Geographical speciation process of the genus Lepturichthys (Cypriniformes: Balitoridae). (2010).
4. Jiang, W., Liu, H.-Z., Duan, Z.-H. & Cao, W.-X. Seasonal variation in drifting eggs and larvae in the Upper Yangtze, China. *Zoological Science* **27**, 402–409 (2010).
5. He, T., Deng, Y., Tuo, Y., Yang, Y. & Liang, N. Impact of the dam construction on the downstream thermal conditions of the Yangtze River. *International journal of environmental research and public health* **17**, 2973 (2020).
6. Yi, Y., Yang, Z. & Zhang, S. Ecological influence of dam construction and river-lake connectivity on migration fish habitat in the Yangtze River basin, China. *Procedia Environmental Sciences* **2**, 1942–1954 (2010).
7. Gao, X., Fujiwara, M., Zhang, W., Lin, P. & Liu, H. The impact of dams on the population viability of a migratory fish in the Yangtze River, China. *Aquatic Conservation: Marine and Freshwater Ecosystems* **32**, 1509–1519 (2022).
8. Wang, Y. *et al*. Complete mitochondrial genome of Lepturichthys fimbriata and phylogenetic analysis of family Balitoridae. *Mitochondrial DNA Part B* **5**, 1315–1316 (2020).
9. Xu, D.-M. *et al*. The complete mitochondrial genome of Lepturichthys fimbriata (Teleostei, Balitoridae, Balitorinae). *Mitochondrial DNA Part A* **27**, 1064–1065 (2016).
10. Nakano, K. *et al*. Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. *Human cell* **30**, 149–161 (2017).
11. Yue, G. & Wang, L. Current status of genome sequencing and its applications in aquaculture. *Aquaculture* **468**, 337–347 (2017).
12. Marcais, G. & Kingsford, C. Jellyfish: A fast k-mer counter. *Tutorialis e Manuais* **1**, 1038 (2012).
13. Liu, B. *et al*. Estimation of genomic characteristics by analyzing kmer frequency in de novo genome projects. *Quantitative Biology* (2013).
14. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175, https://doi.org/10.1038/s41592-020-01056-5 (2021).
15. Vaser, R., Sovic, I., Nagarajan, N. & Sikic, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* **27**, 737–746, https://doi.org/10.1101/gr.214270.116 (2017).
16. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460, https://doi.org/10.1186/s12859-018-2485-7 (2018).
17. Xu, M. *et al*. TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *GigaScience* **9**. https://doi.org/10.1093/gigascience/giaa094 (2020).
18. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760, https://doi.org/10.1093/bioinformatics/btp324 (2009).
19. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95–98, https://doi.org/10.1016/j.cels.2016.07.002 (2016).
20. Dudchenko, O. *et al*. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *science* (2017).
21. Durand, N. C. *et al*. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst* **3**, 99–101, https://doi.org/10.1016/j.cels.2015.07.012 (2016).
22. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212, https://doi.org/10.1093/bioinformatics/btv351 (2015).
23. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA* **117**, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
24. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* Chapter 4, Unit 4 10, https://doi.org/10.1002/0471250953.bi0410s25 (2009).

25. Hoff, K. J. & Stanke, M. Predicting Genes in Single Genomes with AUGUSTUS. *Curr Protoc Bioinformatics* **65**, e57, https://doi.org/10.1002/cpbi.57 (2019).
26. Haas, B. J. *et al*. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* **31**, 5654–5666, https://doi.org/10.1093/nar/gkg770 (2003).
27. Gremme, G., Brendel, V., Sparks, M. E. & Kurtz, S. Engineering a software tool for gene structure prediction in higher organisms. *Information and Software Technology* **47**, 965–978, https://doi.org/10.1016/j.infsof.2005.09.005 (2005).
28. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295, https://doi.org/10.1038/nbt.3122 (2015).
29. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* **9**, R7, https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
30. Camacho, C. *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421, https://doi.org/10.1186/1471-2105-10-421 (2009).
31. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP590038 (2025).
32. Cheng, W. *et al*. NCBI GenBank. *The genome assembly of Lepturichthys fimbriata genome* https://identifiers.org/ncbi/insdc:JBPGZU000000000 (2025).
33. *Lepturichthys fimbriata* genome and annotation files. *Figshare* https://doi.org/10.6084/m9.figshare.29243024 (2025).
34. Wang, F. *et al*. Chromosome-level assembly of Gymnocypris eckloni genome. *Sci Data* **9**, 464, https://doi.org/10.1038/s41597-022-01595-w (2022).
35. Shen, Q. *et al*. Chromosome-level genome assembly of the butterfly hillstream loach Beaufortia pingi. *Scientific Data* **11**, 1260, https://doi.org/10.1038/s41597-024-04144-9 (2024).
36. Li, H. & Birol, I. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100, https://doi.org/10.1093/bioinformatics/bty191 (2018).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant Nos. 32422010, 32470457).

Author contributions

Y.W. and L.Y. conceived this study. C.W., W.C., W.C. and Z.D. collected the samples and performed the experiments; C.W. and W.C. performed the research and analyzed the data. C.W. drafted the manuscript. All authors have read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

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

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.