scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of *Triplophysa scleroptera*

Xiu Feng[1], Ren Zhu[1], Yintao Jia[1,2], Xiaoyun Sui[1], Xiaoling Wang[1,3], Yu Zhuo[1], Junle Li[1,3] & Yifeng Chen[1,2] ✉

Triplophysa scleroptera **is an endemic fish species in Qinghai Lake and the upper reaches of the Yellow River. However, studies on conservation and evolutionary genetics were seriously impeded by the absence of a reference genome. Here, by using PacBio HiFi sequencing and Hi-C assembly technology, we assembled a chromosome-level genome of *T. scleroptera*, with a total length of 660.22 Mb and 99.82% of the sequence anchored to 25 chromosomes. The contig N50 and scaffold N50 were 9.09 Mb and 24.38 Mb, respectively. The evaluation using BUSCO indicated the genome assembly to be 96.40% complete. About 33.41% of the genome consists of repeat elements. We predicted 26,168 protein-coding genes in the genome, and 99.02% of them were functionally annotated. This high-quality reference genome would serve as a valuable genomic resource for advancing evolutionary conservation genetics studies in this species.**

Background & Summary

The genus *Triplophysa* (Cypriniformes: Nemacheilidae) encompasses more than 160 species, making it the largest group within the family[1]. The *Triplophysa* fishes are among the highest altitude-distributed fish species in the world, and together with the Schizothoracinae fishes, they constitute the main body of the fish fauna of the Qinghai-Tibet Plateau (QTP)[2]. *T. scleroptera* is an endemic fish species in Qinghai Lake and the upper reaches of the Yellow River, and typically inhabits the shallow bays of the lake and slow-flowing sections of the river. However, with the intensification of environmental changes and human activities, populations of fishes (e.g. *T. scleroptera*) in these regions face tremendous challenges[3]. For example, hydropower projects along the upper reaches of the Yellow River have disrupted the river connectivity, resulting in habitat fragmentation and gene flow disruption for fish species[4]. Additionally, the introduction of more than ten non-native fishes, such as the topmouth gudgeon (*Pseudorasbora parva*) and rainbow trout (*Oncorhynchus mykiss*), affects the population dynamics of native fish species through competition and predation[5]. Hence, the conservation of the germplasm resources of *T. scleroptera* is an urgent task. However, its basic genetic information such as genetic diversity and structure remains limited. As one of the few fishes that are simultaneously distributed and adapted to both Qinghai Lake and the upper reaches of the Yellow River, *T. scleroptera* is an ideal species for studying population adaptation and evolution in these regions. However, the scarcity of a high-quality reference genome of *T. scleroptera* has hampered studies on the adaptation and evolutionary history in this species, as well as our understanding of its current population status.

In this study, by using approximate 22.51 Gb of Pacbio HiFi reads and 108.93 Gb of Hi-C reads, a chromosome-level genome of *T. scleroptera* was assembled, with the final assembled genome size of 660.22 Mb and the twenty-five chromosomes contributing 659.01 Mb.

Methods

Sample collection, library construction and sequencing. All experiments were approved by the Animal Care and Use Committee at the Institute of Hydrobiology, Chinese Academy of Sciences. A mature female *T. scleroptera* with a body weight of 22.05 g and a body length of 12.80 cm was collected from the Heima River estuary in the Qinghai Lake basin. Genomic DNA was extracted from fresh muscle tissue using the Qiagen DNeasy Blood and Tissue Kit (Qiagen, USA) and used for whole-genome sequencing, including short-read, long-read and Hi-C sequencing. Total RNA was extracted from nine tissues, including liver, spleen, kidney, heart,

[1]Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, 430072, China. [2]State Key Laboratory of Breeding Biotechnology and Sustainable Aquaculture, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, 430072, China. [3]University of Chinese Academy of Sciences, Beijing, 100049, China. ✉e-mail: chenyf@ihb.ac.cn

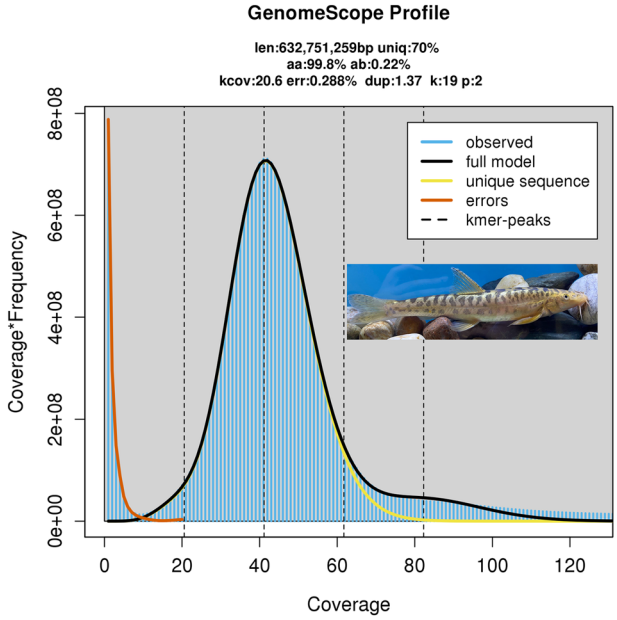

Fig. 1 Genome size estimated by GenomeScope using 19-mers.

brain, muscle, skin, ovary and gill, and used for transcriptome sequencing. The purity and concentration of the DNA and RNA were evaluated with a Qubit 3.0 Fluorometer (Thermo Fisher Scientific, USA).

Short-read libraries with 350 bp insert size were constructed using the Nextera DNA Flex Library Prep Kit and sequenced on the Illumina Novaseq. 6000 platform (Illumina, USA) with a paired-end 150-bp (PE-150) strategy. For PacBio HiFi sequencing, long-read libraries were prepared using the SMRTbell Express Template Prep Kit 2.0 following the manufacturer's instructions and then sequenced using the PacBio Sequel II platform (Pacific Biosciences, USA). The Hi-C library was constructed using the GrandOmics Hi-C kit with the *DpnII* restriction endonuclease (GrandOmics, China) following the manufacturer's protocol and sequenced with a PE-150 strategy on the Illumina Novaseq 6000 platform. The Illumina and PacBio sequencing data were filtered using fastp (v0.20.0)[6] and the CCS algorithm (v3.4.1), respectively, to obtain the clean data for analysis. In total, 22.51 Gb of clean PacBio HiFi data, 108.93 Gb of clean Hi-C data were obtained for the genome assembly. In addition, 31.96 Gb of clean short-read sequencing data were obtained for the genome survey.

To facilitate genome annotation, RNA-seq libraries were constructed using a mixed RNA sample with the TruSeq RNA Sample Prep Kit (Illumina, USA) following the manufacturer's instructions. The libraries were sequenced with a PE-150 strategy on the Illumina Novaseq. 6000 platform and yielded 12.65 Gb of clean sequencing data.

Genome survey and assembly. The genome size, heterozygosity and duplication ratio were estimated based on the clean short-read sequencing data using Jellyfish (v2.3.0)[7] and GenomeScope (v2.0)[8]. As determined from 19-mers histogram, the estimated genome size of *T. scleroptera* was 632.75 Mb, with a heterozygous ratio of 0.22% (Fig. 1). The clean HiFi data (35 × sequencing depth) contained 1.5 million reads with an average read length of 14.9 Kb, the maximum read length of 52.3 Kb and a read N50 of 15.1 Kb. The initial contig assembly based on the clean HiFi data by Hifiasm (v0.16.1)[9] yielded 547 contigs with a total length of 660.20 Mb and an N50 of 9.09 Mb (Table 1). The contigs was then scaffolded with Hi-C reads using Juicer (v1.6)[10] and 3D-DNA pipeline[11], to achieve a near-chromosome level. Subsequently, manual curation and correction were performed based on chromosome interaction strengths using Juicebox software (v1.11.08)[12], to ultimately obtain a chromosome-level genome. The final assembled genome size was 660.22 Mb with a scaffold N50 of 24.38 Mb (Table 1), and 659.01 Mb of the sequence (99.82%) was anchored to 25 chromosomes (Table 2). The chromosome numbers were assigned based on the chromosomal collinearity with *T. dalaica* (GCF_015846415.1)[13], to facilitate further genomic studies among *Triplophysa* species. The Hi-C heatmap clearly showed interactions between the 25 chromosomes, with strong signals near the diagonal, indicating the high-quality of the pseudochromosomes assembly (Fig. 2). The statistics of the assembled genome were similar to those of published *Triplophysa* genomes[13–15] which were also assembled based on HiFi and Hi-C data (Table 1). The mitochondrial genome, which was assembled using our previous described procedures[3,16] based on short reads, was 16,572 bp in size, with 13 protein-coding genes, 22 tRNAs and 2 rRNAs (Fig. S1).

Repetitive sequence annotation. Repetitive sequences were annotated by using both *de novo* prediction and homology-based methods[17,18]. RepeatModeler (v2.0.1)[19] was firstly employed to detect repetitive elements. LTR sequences were also predicted using LTR_retriever (v2.9.0)[20]. These *de novo* predicted sequences were then merged with the RepBase library and deduplicated to create a repeat library. RepeatMasker (v4.1.2)[21] was utilized to identify final repetitive elements. In addition, tandem repeats were identified using MISA (v2.1)[22] and TRF (v4.09)[23]. In total, 220.59 Mb (33.41% of the genome) of repetitive sequences were identified, including LTR

Feature	*Triplophysa scleroptera*	*Triplophysa dalaica*[13]	*Triplophysa rosa*[14]	*Triplophysa tibetana*[15]
Assembly genome length (Mb)	660.22	607.91	684.8	652.9
GC (%)	38.87	39	39.5	39
Number of contigs	547	166	6,299	1,325
Contig N50 (Mb)	9.09	9.27	0.38	3.05
Contig N90 (Mb)	0.75	—	—	0.22
Number of scaffolds	357	141	1,038	293
Scaffold N50 (Mb)	24.38	23.6	24.84	24.89
Scaffold N90 (Mb)	19.68	—	—	21.05
Number of chromosomes	25	25	25	25
Max chromosome length (Mb)	46.28	36.52	40.13	39.52
Min chromosome length (Mb)	18.20	16.5	16	—
Percentage of nucleotides assigned to chromosomes	99.82%	96.30%	94.52%	98.70%

Table 1. Summary statistics of the assembled genome of *Triplophysa scleroptera* and other published *Triplophysa* genomes.

Chromosome	Length (bp)	Contig number (bp)	Gene number
Chr01	37,178,414	18	1219
Chr02	30,591,935	14	1178
Chr03	31,721,367	15	1414
Chr04	32,640,930	29	1300
Chr05	46,279,922	29	1493
Chr06	26,972,843	27	974
Chr07	26,445,896	18	1196
Chr08	25,493,853	20	843
Chr09	25,600,424	24	975
Chr10	33,851,409	47	1026
Chr11	23,998,962	14	877
Chr12	25,679,262	9	1121
Chr13	24,730,667	39	935
Chr14	25,304,596	14	871
Chr15	23,370,065	10	950
Chr16	23,586,636	32	798
Chr17	23,923,161	23	873
Chr18	23,386,455	20	1063
Chr19	22,015,722	16	950
Chr20	21,328,367	18	915
Chr21	22,323,180	15	847
Chr22	22,098,943	29	895
Chr23	20,771,475	14	719
Chr24	18,201,912	22	798
Chr25	21,513,396	15	924

Table 2. Chromosome lengths and features.

transposons (6.30%), LINE transposons (3.99%), DIRS transposons (1.37%), SINE transposons (0.30%), DNA transposons (14.93%), Tandem Repeats (6.47%) and other types of elements (Table 3 and Fig. 3).

Gene prediction and function assignment. Gene prediction of *T. scleroptera* was performed through three methodologies, including homology-based prediction, transcriptome-based prediction, and *ab initio* prediction[24,25]. GeMoMa (v1.7)[26] was employed for gene prediction based on protein homologous, utilizing protein sequences from a model species and three species of the same genus: *Danio rerio* (GCF_000002035.6)[27], *T. dalaica* (GCF_015846415.1)[13], *T. rosa* (GCF_024868665.1)[14] and *T. tibetana* (GCA_008369825.1)[15]. Two approaches were adopted for transcriptome-based prediction: 1) the first involved assembling transcripts using Hisat (v2.1.0)[28] and StringTie (v2.1.4)[29], followed by gene prediction with GeneMarkS-T (v5.1)[30]; 2) the second utilized Trinity (v2.11)[31] for transcript assembly, and PASA (v2.4.1)[32] for gene prediction. The *ab initio* prediction was performed using Augustus (v3.1.0)[33] and SNAP (v2013.11.2927)[34]. The prediction results from the three methodologies were integrated and refined using EVM (v1.1.1)[32] and PASA (v2.4.1). A total of 26,168 genes were predicted in the

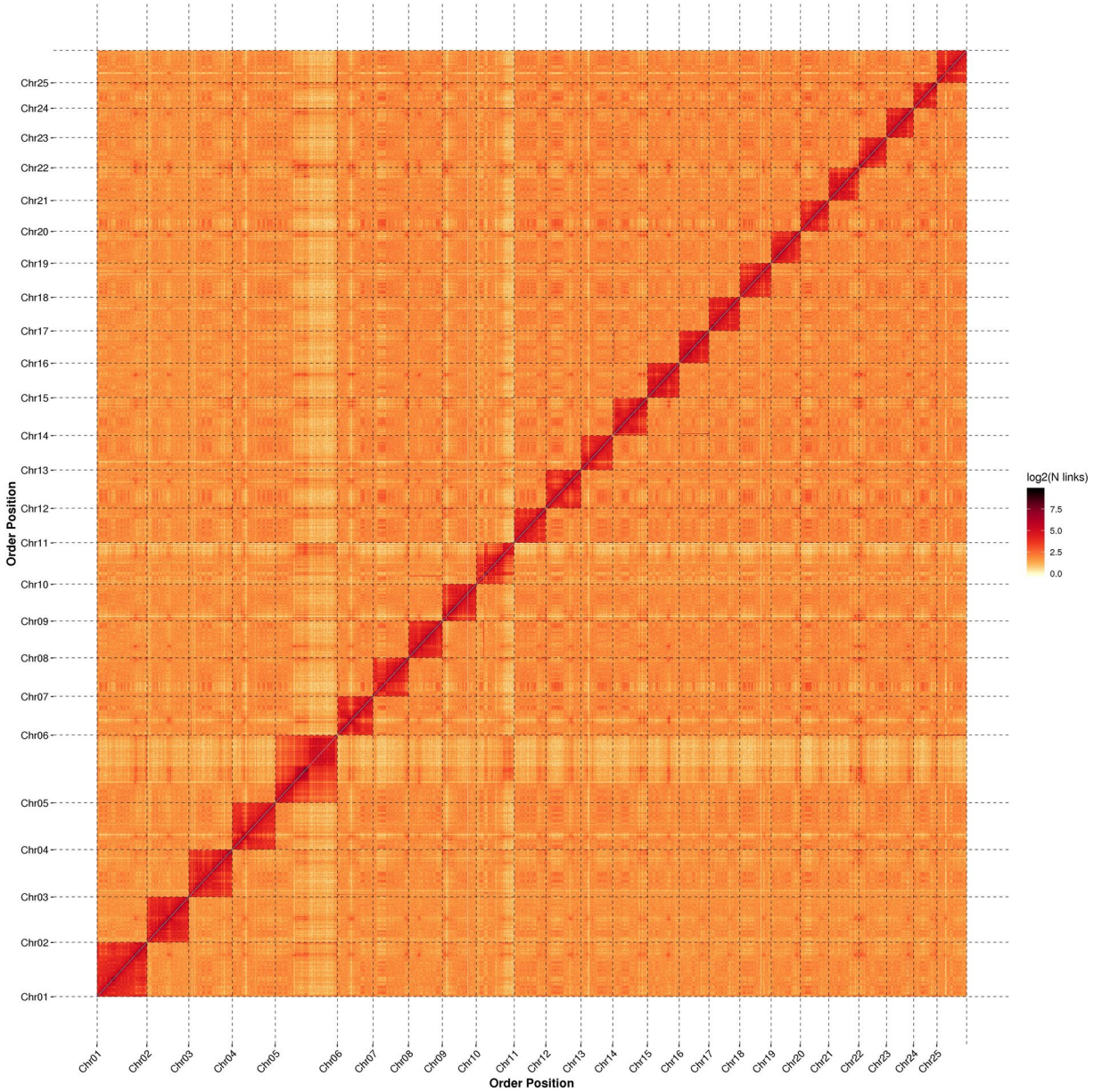

Fig. 2 Genome-wide Hi-C heatmap of *Triplophysa scleroptera*.

Type		Number	Length (bp)	% of the genome
Transposable elements	LTR	140,892	41,610,846	6.30
	LINE	77,292	26,360,518	3.99
	DIRS	13,873	9,048,992	1.37
	SINE	16,435	1,984,512	0.30
	DNA transposon	518,912	98,596,998	14.93
	Unknown	1,281	290,987	0.04
	Total	768,685	177,892,853	26.94
Tandem Repeats		207,148	42,692,684	6.47
Total Repeats		975,833	220,585,537	33.41

Table 3. Summary of repetitive sequences in the genome of *Triplophysa scleroptera*.

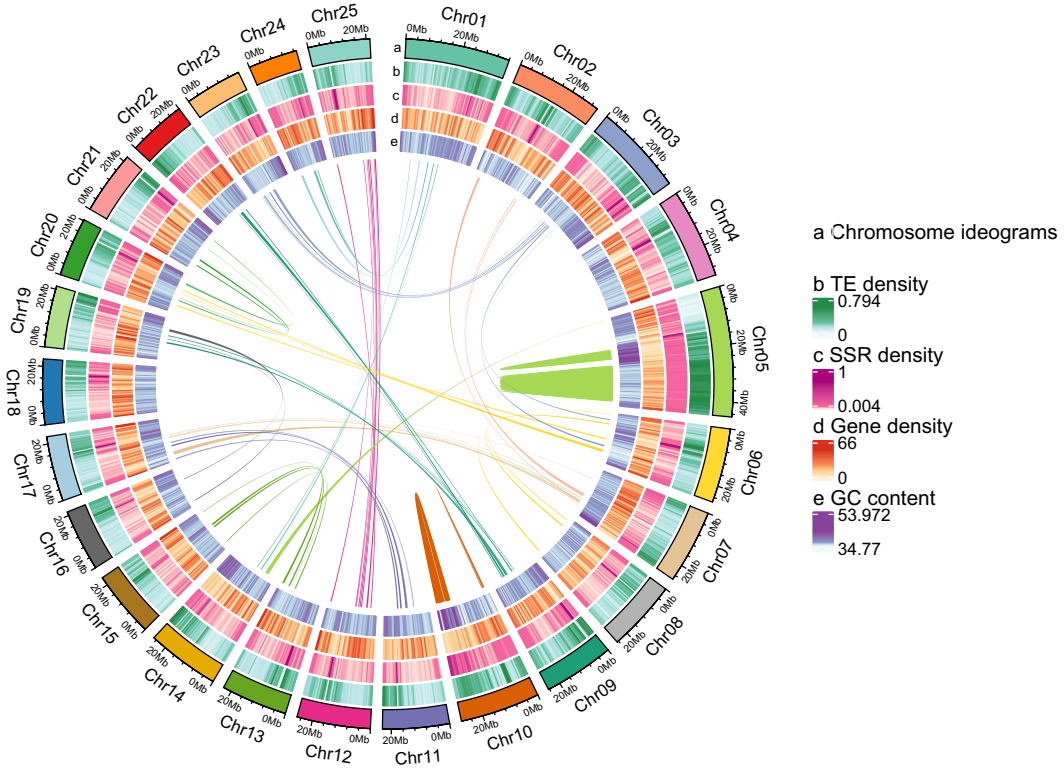

Fig. 3 Circos plot showing characteristics of the *Triplophysa scleroptera* genome. Tracks from the outer to the inner layers represent the 25 chromosomes (**a**), transposable element (TE) density (**b**), tandem repeat (SSR) density (**c**), gene density (**d**) and GC density (**e**), and each line in the inner circle connects paralogous genes on different chromosomes.

Item	Value	Database	Number of annotated genes
Number of genes	26,168	Nr	25,079
Total length of genes (bp)	359,564,642	Swissprot	16,931
Average length per gene (bp)	13,741	TrEMBL	25,896
Number of exons	256,613	COG/KOG	16,387
Average exon number per gene	9.81	KEGG	22,550
Average length per exon (bp)	2385	Pfam	23,276
Number of introns	230,445	eggNOG	22,027
Average intron number per gene	8.81	GO	22,466
Average length per intron (bp)	11,355	All	25,911

Table 4. Summary of predicted genes and function annotation based on different databases.

genome, with an average length of 13,740.62 bp, an average coding sequence length of 1,728.74 bp, and an average exon number of 9.81 (Table 4). In addition, 23,232 tRNAs, 16,118 rRNAs and 198 miRNAs were identified by using tRNAscan-SE (v1.3.1)[35], barrnap (v0.9)[36] and Infenal (v1.1)[37], respectively.

Gene functional annotations were carried out by aligning the predicted sequences against entries in databases of Nr, Swissprot, TrEMBL, COG/KOG and KEGG using Diamond (v 2.1.1)[38], Pfam using Hmmscan[39], and eggNOG using eggNOG-mapper (v2)[40], The GO terms for genes were obtained from the corresponding Swissprot entry. As a result, 25,911 genes (99.02%) were functionally annotated in at least one database, and 10,910 genes 41.69%) were annotated in all databases (Table 4).

Data Records

The sequencing data of *T. scleroptera*, including the PacBio long-read data, and Illumina Hi-C data and RNA-seq data have been deposited in the NCBI Sequence Read Archive (SRA) database under the accessions SRR32801341[41], SRR32801340[42] and SRR32234133[43], respectively, with the associated BioProject accession of PRJNA1219334[44]. The assembled genome sequences have been deposited in the NCBI GenBank with the accession number GCA_049739055.1[45]. Moreover, data of the genome annotations, including predicted coding sequences and protein sequences have been deposited at Figshare[46].

Type	Assembled genome		Predicted genes	
	Number	Percent (%)	Number	Percent (%)
Complete BUSCOs (C)	3,489	95.85	3,494	95.99
Complete and single-copy BUSCOs (S)	3,420	93.96	3,428	94.18
Complete and duplicated BUSCOs (D)	69	1.90	66	1.81
Fragmented BUSCOs (F)	20	0.55	23	0.63
Missing BUSCOs (M)	131	3.60	123	3.38
Total BUSCO groups searched	3,640	100.00	3,640	100.00

Table 5. BUSCO analysis of the assembled genome with HiFi data and predicted genes.

Technical Validation

Evaluation the genome assembly and gene annotation. The quality of the genome assembly and the integrity of the gene annotation of *T. scleroptera* was evaluated by Benchmarking Universal Single-Copy Orthologs (BUSCO, v5.4.3)[47] based on the conserved Actinopterygii gene set (actinopterygii_odb10). The analysis of assembled genome based on HiFi data showed that 96.40% of all BUSCOs were assembled, including 95.85% and 0.55% of BUSCOs which were completely and partially assembled, respectively (Table 5). In addition, the results of analysis of predicted genes showed that 96.62% of all BUSCOs were predicted, including 95.99% and 0.63% of BUSCOs which were completely and partially predicted, respectively. These results indicated a high level of completeness for the genome assembly the gene annotation.

Code availability

No specific code or script was used in this work, and all commands and pipelines used in data processing were performed based on the manuals and protocols of the applied bioinformatic software. The versions and parameters of all used software programs were described in the Methods section.

Received: 13 February 2025; Accepted: 1 May 2025;
Published online: 01 July 2025

References

1. Zhang, C. Y., Luo, P., Huang, F. & Zhang, E. Revision of the loach genus *Claea* Kottelat, 2010 (Teleostei: Nemacheilidae) in China, with a description of a new species from the Chang-Jiang basin. *Zootaxa* **5543**(3), 404–422 (2024).
2. Wu, Y. & Tan, Q. Characteristics of the fish-fauna of the characteristics of Qinghai-Xlzang Plateau and its geological distribution and formation. *Acta Zoologica Sinica* **37**(2), 135–152 (1991).
3. Feng, X. *et al*. Genetic responses in sexual diploid and unisexual triploid goldfish (*Carassius auratus*) introduced into a high-altitude environment. *Mol. Ecol.* **32**(8), 1955–1971 (2023).
4. Jia, Y. *et al*. Human disturbance and long-term changes in fish taxonomic, functional and phylogenetic diversity in the Yellow River, China. *Hydrobiologia* **847**(18), 3711–3725 (2020).
5. Jia, Y. *et al*. Understanding invasion success of *Pseudorasbora parva* in the Qinghai-Tibetan Plateau: Insights from life-history and environmental filters. *Sci. Total Environ.* **694**, 133739 (2019).
6. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**(17), 884–890 (2018).
7. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**(6), 764–770 (2011).
8. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**(1), 1432 (2020).
9. 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 (2021).
10. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
11. Dudchenko, O. *et al*. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
12. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
13. Zhou, C. *et al*. The chromosome-level genome of *Triplophysa dalaica* (Cypriniformes: Cobitidae) provides insights into its survival in extremely alkaline environment. *Genome Biol. Evol.* **13**, evab153 (2021).
14. Zhao, Q., Shao, F., Li, Y., Yi, S. V. & Peng, Z. Novel genome sequence of Chinese cavefish (*Triplophysa rosa*) reveals pervasive relaxation of natural selection in cavefish genomes. *Mol. Ecol.* **31**, 5831–5845 (2022).
15. Yang, X. *et al*. Chromosome-level genome assembly of *Triplophysa tibetana*, a fish adapted to the harsh high-altitude environment of the Tibetan Plateau. *Mol. Ecol. Resour.* **19**, 1027–1036 (2019).
16. Liu, S. L., Hansen, M. M. & Jacobsen, M. W. Region-wide and ecotype-specific differences in demographic histories of threespine stickleback populations, estimated from whole genome sequences. *Mol. Ecol.* **25**(20), 5187–5202 (2016).
17. Sun, B. *et al*. Chromosome-scale and haplotype-resolved genome assembly of the autotetraploid *Misgurnus anguillicaudatus*. *Sci. Data* **11**, 1059 (2024).
18. Dong, X. *et al*. A chromosome-level genome assembly of Cape hare (*Lepus capensis*). *Sci. Data* **11**, 1081 (2024).
19. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2020).
20. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).
21. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* **25**, 4–10 (2009).
22. Beier, S., Thiel, T., Münch, T., Scholz, U. & Mascher, M. MISA-web: a web server for microsatellite prediction. *Bioinformatics* **33**, 2583–2585 (2017).
23. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).

24. Liu, H. *et al*. Chromosome-level genome assembly of the mud carp (*Cirrhinus molitorella*) using PacBio HiFi and Hi-C sequencing. *Sci. Data* **11**, 1249 (2024).
25. Liang, Y. *et al*. Chromosome-level genome assembly of the smallscale yellowfin (*Plagiognathops microlepis*). *Sci. Data* **11**, 1234 (2024).
26. Keilwagen, J., Hartung, F. & Grau, J. GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Gene prediction: Methods and protocols* **1962**, 161–177 (2019).
27. Howe, K. *et al*. The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**, 498–503 (2013).
28. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
29. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnolo.* **33**, 290–295 (2015).
30. Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res.* **43**, e78–e78 (2015).
31. Grabherr, M. G. *et al*. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat. Biotechnolo.* **29**, 644 (2011).
32. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, 1–22 (2008).
33. Stanke, M. *et al*. AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
34. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
35. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
36. Loman, T. *Lund University*. *A novel method for predicting ribosomal RNA genes in prokaryotic genomes*, (2017).
37. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
38. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2015).
39. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).
40. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol.* **38**, 5825–5829 (2021).
41. *NCBI Sequence Read Archive*. https://identifiers.org/ncbi/insdc.sra:SRR32801341 (2025).
42. *NCBI Sequence Read Archive*. https://identifiers.org/ncbi/insdc.sra:SRR32801340 (2025).
43. *NCBI Sequence Read Archive*. https://identifiers.org/ncbi/insdc.sra:SRR32234133 (2025).
44. *NCBI BioProject* https://identifiers.org/ncbi/bioproject:PRJNA1219334 (2025).
45. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_049739055.1 (2025).
46. Feng, X. Chromosome-level genome assembly of *Triplophysa scleroptera*. *Figshare* https://doi.org/10.6084/m9.figshare.28350143 (2025).
47. Simão, 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 (2015).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (U22A20454 and 32370559), the Second Tibetan Plateau Scientific Expedition and Research (STEP) program (2019QZKK0304), the Finance Special Fund of Chinese Ministry of Agriculture and Rural Affairs of the People's Republic of China (Fisheries resources and environment survey in the key water areas of Southwest China).

Author contributions

Y.C. conceived and supervised the study; X.F., R.Z., Y.J., X.S., X.W., Y.Z. and J.L. collected samples; X.F. and X.W. performed the experiments; X.F. conducted the genome assembly and bioinformatics analysis; X.F. drafted the manuscript, Y.C. revised the manuscript, and all authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-05122-5.

Correspondence and requests for materials should be addressed to Y.C.