



OPEN

DATA DESCRIPTOR

Haplotype resolved chromosome-level genome assembly of the gold barb (*Barbodes semifasciolatus*)

Weitao Chen^{1,2,3,6}, Chao Li^{4,6} , Rong Yang⁴, Yuefei Li^{1,2,3}, Baosheng Wu⁵ & Jie Li^{1,2,3}

The gold barb (*Barbodes semifasciolatus*), a member of the Cyprinidae family, exhibits remarkable adaptability to highly acidic environments, making it an ideal model for studying extreme environmental adaptation. However, its genome has not been previously characterized. To address this, we assembled a high-quality chromosome-scale genome for *B. semifasciolatus* using High-Fidelity (HiFi) sequencing and Hi-C technology. The resulting haplotype-resolved assemblies, spanning 776 Mb and 779 Mb across 25 chromosomes, achieved genome coverages of 99.5% and 99.7%, respectively, and included four gap-free chromosomes. Genome quality assessment using BUSCO indicated a high completeness score of 98.2% for haplotype1 and 98.3% for haplotype2, further validated by strong synteny with the zebrafish (*Danio rerio*), confirming the assembly's integrity and continuity. Through integration of full-length transcriptome data, RNA sequencing, and homology-based annotation, we identified 26,057 protein-coding genes with 2,087 pseudogenes in haplotype 2, and 25,622 protein-coding genes with 2,101 pseudogenes in haplotype 1. This high-resolution genome assembly is a crucial resource for advancing research in the Cyprinidae, particularly for understanding adaptive evolution in extreme environments.

Background & Summary

The Cyprinidae, the largest family of freshwater fishes, has a remarkable global distribution, with particularly high diversity in Asia^{1,2}. This family is characterized by exceptional adaptability, allowing rapid radiation and colonization of diverse habitats, particularly in East Asia^{3–5}. The ecological success of cyprinids in this region can be attributed to their high reproductive capacity, broad ecological tolerance and efficient dispersal mechanisms^{6,7}. These traits have allowed cyprinids to thrive in a wide range of environmental conditions, from mountain streams to lowland rivers, often becoming dominant components of freshwater ecosystems^{8,9}. Understanding the mechanisms that drive their distribution and adaptability not only provides insights into their evolutionary success, but also informs conservation strategies to maintain biodiversity and ecological balance in rapidly changing environments.

Gold barb (*Barbodes semifasciolatus*), a member of the family Cyprinidae, exhibits strong tolerance to extremely acidic environments, making it an excellent model for studying the adaptive mechanisms of cyprinids in response to harsh conditions¹⁰. In this study, we assembled two haplotype resolved chromosome-level genomes (776 and 779 Mb) for the gold barb using PacBio's advanced highly accurate long-read sequencing and chromosome conformation capture technology, achieving a contig N50 of 23.07 Mb and 26.46 Mb. The continuity and accuracy of the genome was confirmed by conserved core gene analysis (BUSCO scores) and synteny assessment, establishing it as one of the highest quality cyprinid genomes assembled to date. Overall, this high quality genome will be a great resource for future research into the adaptive evolution of cyprinids.

¹Pearl River Fisheries Research Institute, Chinese Academy of Fishery Sciences, Guangzhou, 510380, China.

²Key Laboratory of Prevention and Control for Aquatic Invasive Alien Species, Ministry of Agriculture and Rural Affairs, Guangzhou, Guangdong, 510380, China. ³Scientific Observing and Experimental Station of National Fisheries Resources and Environment, Guangzhou, Guangdong, 510380, China. ⁴Guangzhou Key Laboratory of Subtropical Biodiversity and Biomonitoring, Guangdong Provincial Engineering Technology Research Center for Environmentally, Friendly Aquaculture, School of Life Sciences, South China Normal University, Guangzhou, 510631, China. ⁵Guangdong Key Laboratory of Animal Conservation and Resource Utilization, Institute of Zoology, Guangdong Academy of Sciences, Guangzhou, 510260, China. ⁶These authors contributed equally: Weitao Chen, Chao Li. ✉e-mail: wubs@nwpu.edu.cn; lijie1561@163.com

Term	Hap1		Hap2	
	Contig assembly	Hi-C assembly	Contig assembly	Hi-C assembly
Genome size (bp)	776,018,489	776,025,089	779,091,036	779,095,936
N50	23,066,965	30,078,511	26,458,349	30,708,761
N90	7,328,842	30,708,761	8,088,531	26,458,349
Chromosomes	—	25	—	25
Contigs	143	52	100	26
Max length (bp)	41,374,509	47,254,445	32,294,908	46,208,965
Min length (bp)	19,111	19,111	18,950	18,950

Table 1. Statistics of the genome assembly.

Methods

Sample collection. We obtained a gold barb (*Barbodes semifasciolatus*) from the highly acidic waters of the Guangzhou Conghua Nature Reserve for *Tanichthys albonubes* in China. The fish were anaesthetized with MS-222 as in our previous study¹¹, and samples of muscle, liver, blood, gills, brain and kidney tissues were rapidly collected. These tissues were immediately frozen in liquid nitrogen and stored at −80 °C. Muscle tissue was used for PacBio’s advanced highly accurate long-read sequencing (HiFi), liver tissue for chromosome conformation capture (Hi-C) technology and a pooled transcriptome of all collected tissues was prepared for RNA-seq sequencing.

Genomic long-read sequencing (HiFi). We followed PacBio’s standard protocol (Pacific Biosciences, California, USA) to generate genomic data using the PacBio Sequel II platform. The sequencing yielded 1,895,794 clean reads with a total of 34.76 Gb of genomic data, achieving an average read length of 18.34 kb. To ensure high-quality data, reads were filtered for adapter sequences and low-quality bases. The filtering criteria were set to remove reads with a length shorter than 0.5 kb and those with a quality score lower than 0.80.

Iso-Seq Library Construction and Sequencing. Iso-Seq libraries were prepared using the PacBio SMRTbell prep kit 3.0, following the manufacturer’s guidelines. This process resulted in the generation of 49,885,551 long reads totaling 117.61 Gb, with an average read length of 2.35 kb. The libraries were loaded onto the PacBio Sequel II platform, and sequencing was carried out using the SMRT cells. Post-sequencing, long reads were filtered using the PacBio Circular Consensus Sequence (CCS) algorithm to ensure only high-quality consensus sequences were retained.

RNA Extraction and cDNA Library Construction. Total RNA was extracted from pooled tissue samples using the TRIzol reagent (TIANGEN, Cat # DP424, China) following the manufacturer’s protocol. The extracted RNA was quantified and assessed for integrity using the Agilent 2100 Bioanalyzer. Only high-quality RNA samples (RIN > 7.0) were used for further processing. Reverse transcription was performed to synthesize cDNA from the RNA samples, following standard protocols with random hexamers and SuperScript III Reverse Transcriptase (Thermo Fisher Scientific, USA). The cDNA libraries were then constructed using the Illumina TruSeq Stranded mRNA Library Prep Kit (Illumina, USA). The constructed cDNA libraries were sequenced using the Illumina NovaSeq 6000 platform (Illumina, USA). Paired-end sequencing was performed with a read length of 150 bp. During sequencing, the Illumina software automatically filtered out reads of low quality, ensuring that only high-quality reads were retained for downstream analysis. After sequencing, raw reads were processed using the fastp¹² tool (v0.20.0) to remove adapter sequences, low-quality bases (with Phred score < 20), and short reads. The resulting dataset consisted of 58,426,944 high-quality short reads, corresponding to 8.7 Gb of sequencing data.

Hi-C Library Preparation and Sequencing. Genomic DNA was extracted from muscle and cross-linked with biotinylated nucleotides using formaldehyde. The DNA was then digested with the restriction enzyme DpnII, followed by ligation of the DNA ends to form chimeric fragments. The biotinylated ligation products were captured on streptavidin beads and the libraries were amplified by PCR. Then Hi-C libraries were sequenced on the Illumina NovaSeq 6000 platform using paired-end sequencing (150 bp). Library concentration and fragment size were verified prior to sequencing. Raw sequencing data were processed to remove low quality reads, adapter sequences using fastp¹² tool (v0.20.0), followed by alignment to the reference genome using the BWA-MEM v0.7.12¹³.

Chromosome level genome assembly of gold barb. By integrating HiFi reads and Hi-C reads, we assembled the gold barb genome using Hifiasm v0.16.1¹⁴ with default parameters. This process produced two haplotype-resolved assemblies (hap1 and hap2) (Table 1). The resulting haplotype-resolved assemblies were 776 Mb and 779 Mb in size, containing 143 and 100 contigs respectively, with N50 of 23.07 Mb and 26.46 Mb (Table 1). The genome size is very comparable to that of its closely related cyprinid species (*Puntigrus tetrazona*). We then aligned the Hi-C reads to the two haplotype-resolved assemblies using BWA v0.7.12¹³. Chromosome-level assemblies were then constructed for both haplotypes using Haphic¹⁵, resulting in 25 high quality chromosomes for each haplotype. Next, manual curation of potential assembly errors was performed using JuiceBox v2.20.00¹⁶, with contigs lacking obvious interaction relationships treated as single scaffolds. The haplotype-resolved assemblies had anchoring rates of 99.6% and 99.7%, respectively. Each chromosome consisted of between 1 and 6 contigs, with four chromosomes being gap-free (Fig. 1A,B). We further employed

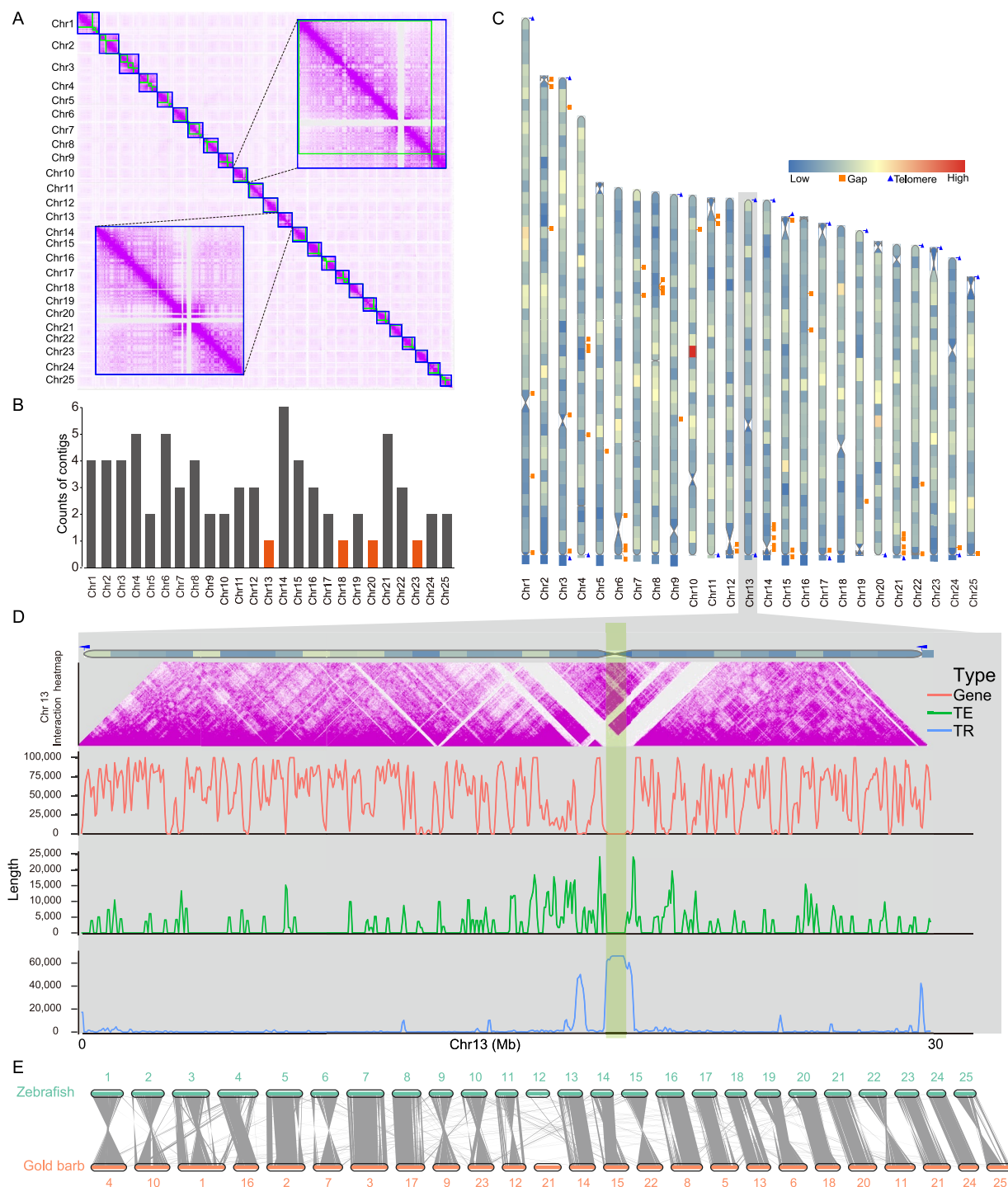


Fig. 1 Genome assembly of haplotype 2 for the gold barb (*B. semifasciatus*). **(A)** Hi-C linkage density heat map, where the x and y axes represent genomic positions. Red dots indicate regions of high linkage density, suggesting that these regions are more likely to belong to the same chromosome. Blue boxes indicate potential chromosomes, while green boxes indicate contigs. **(B)** Distribution of contigs among the 25 chromosomes. Chromosomes with free gaps are shown in orange. **(C)** Assembly of the 25 chromosomes showing the distribution of gaps, gene density and putative telomeric and centromeric regions. Blue triangular regions indicate putative telomeric regions, orange boxes represent the locations of assembly gaps, and concave regions mark potential centromeric regions. **(D)** Detailed view of chromosome 13 showing the distribution of genes, repetitive sequences, transposable elements and the Hi-C linkage density heat map. Abbreviations: TE, transposable elements; TR, tandem repeats. **(E)** Synteny between the chromosomes of the gold barb and the 25 chromosomes of zebrafish. Each line represents a syntenic block, highlighting homologous regions between the two species. The 25 zebrafish chromosomes are shown on one axis, while the corresponding gold barb chromosomes are displayed on the other axis.

Class	Count (hap1)	Masked (bp) (hap1)	Percentages (hap1)	Count (hap2)	Masked (bp) (hap2)	Percentages (hap2)
DNA	3,742	1,130,000	0.15%	11,248	1,831,127	0.24%
LINE	66,701	30,137,446	3.89%	68,682	25,292,327	3.24%
LTR	17,7455	99,229,249	12.78%	190,152	99,991,077	12.85%
SINE	27,044	5,207,716	0.68%	21,085	4,380,438	0.56%
TIR	305,331	76,894,289	9.90%	281,108	70,467,345	9.04%
Other	158,918	35,774,276	4.61%	176,109	41,211,754	5.29%
Total	739,191	248,372,976	32.01%	748,384	243,174,068	31.22%

Table 2. Statistics of repeat elements genome assembly.

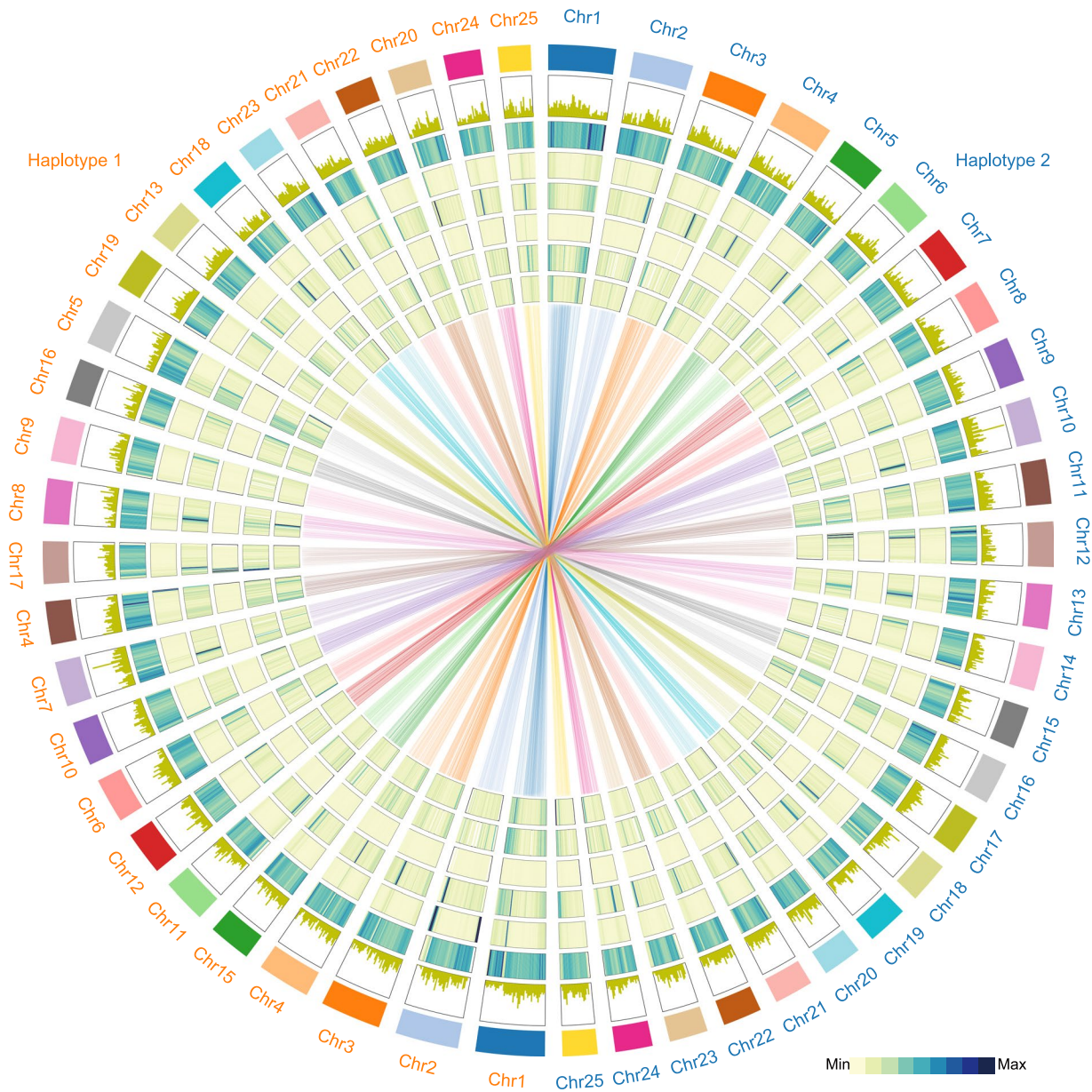


Fig. 2 Circos plots show the distribution of genomic components in the gold barb. Genomic features are shown in 1,000,000 bp windows. From the outermost circle to the innermost circle, the following features are shown: gold barb chromosomes, gene frequency across the genome, density of DNA transposable elements, density of long interspersed nuclear elements (LINEs), density of long terminal repeats (LTRs), density of short interspersed nuclear elements (SINEs), density of other transposable elements and density of unknown genomic elements. The innermost links represent the collinearity relationships between different haplotypes.

quarTeT¹⁷ to predict the telomeric and centromeric regions of these 25 chromosomes. The centromeric regions were subsequently validated by their intra-chromosomal interaction patterns. For example, in the 18,754,673–19,992,527 bp region on chromosome 13, we found that this region is largely devoid of coding genes, enriched with tandem repeat sequences (TR), and flanked by transposable elements (TE). Moreover, Hi-C data revealed minimal chromatin interactions in this region, suggesting its potential as a centromeric region (Fig. 1C,D). The BUSCO (v5.5.0) score based on actinopterygii_odb10 is 98.1%, comprising 96.2% Complete and single-copy and 1.9% Complete and duplicated. Finally, we performed a synteny analysis of the genomes of zebrafish and gold barb using JCVI v0.9.13¹⁸, which revealed a strong syntenic relationship between the two species (Fig. 1E).

Annotation of repetitive sequences. We used both homologous and *de novo*-based approaches to annotate the repetitive sequences in the gold barb genome. For homologous methods, we used RepeatProteinMask¹⁹ and RepeatMasker²⁰ to align transposable elements (TEs) at the protein and DNA levels, respectively. Tandem repeats were then annotated using TRF²¹ with the following parameters: trf 2 7 7 80 10 50 2000 -d -h. For the species-specific repeat library, we used RepeatModeler (version 1.73) to identify consistent classification sequences, which were then used to run RepeatMasker. Finally, we merged all annotated repeat sequences using bedtools²² (v2.25.0). In total, repeat elements account for 31.22% of the genome (Table 2 and Fig. 2).

Annotation of protein-coding genes. For protein-coding gene prediction, we used a combination of three approaches as in previous studies^{23–25}: *ab initio* prediction, homology-based prediction and transcriptome-based prediction. Transcriptome data were generated using two sequencing methods: Paired-end RNA-seq and full-length transcriptome sequencing. TransDecoder²⁶ v5.5.0 was used to predict proteins from paired-end RNA-seq transcripts, while IsoQuant²⁷ v3.6.2 was used to predict proteins from full-length transcripts. For homology-based prediction, genomic data from related species were downloaded from NCBI, including *Carassius gibelio* (GCF_023724105.1), *Cyprinus carpio* (GCF_018340385.1), *Carassius auratus* (GCF_003368295.1), *Labeo rohita* (GCF_022985175.1), *Carassius carassius* (GCF_963082965.1), *Onychostoma macrolepis* (GCF_012432095.1), *Puntigrus tetrazona* (GCF_018831695.1), *Sinocyclocheilus graham* (GCF_001515645.1), *Sinocyclocheilus rhinoceros* (GCF_001515625.1) and *Sinocyclocheilus anshuiensis* (GCF_001515605.1). These genomes were used for comparative gene structure alignment. Finally, EVidenceModeler²⁸ v1.1.1 was used to integrate the results of the three prediction approaches into the final gene set, resulting in 26,057 protein-coding genes and 2,087 pseudogenes.

Data Records

Hi-C data, full length Full-length transcriptome data and transcriptome data were deposited in the National Center for Biotechnology Information (NCBI) SRA database²⁹ under the accession numbers SRR31825773, SRR31825774 and SRR31825775, respectively. HiFi data were deposited in the NCBI SRA database²⁹ under the accession numbers SRR31825776 and SRR31825777. The assembly genome data of the two haplotypes were deposited at GenBank under accession JBMAGD000000000³⁰ and JBMAGE000000000³¹. Genome annotations were deposited in the Figshare database³².

Technical Validation

The completeness of the genome assemblies was assessed using BUSCO v5.5.0³³. The Hap2 results showed a BUSCO completeness of 98.1%, including 96.2% single-copy genes, 1.9% duplicated genes, 0.7% fragmented genes, and 1.2% missing genes. The analysis involved mapping the PacBio and Illumina sequencing reads to Hap2 using Minimap2 v2.28³⁴, BWA v0.7.12¹³ and SAMtools v1.2.0³⁵. The mapping rates for HiFi, Hi-C, and RNA-Seq data were 100%, 99.91%, and 99.84%, respectively.

Code availability

No specific script was used in this work. The corresponding bioinformatics software and the specific versions of software have been described in Methods.

Received: 30 December 2024; Accepted: 9 May 2025;

Published online: 29 May 2025

References

- He, S. *et al.* Molecular phylogenetics of the family Cyprinidae (Actinopterygii: Cypriniformes) as evidenced by sequence variation in the first intron of S7 ribosomal protein-coding gene: further evidence from a nuclear gene of the systematic chaos in the family. *Mol Phylogenet Evol* **46**, 818–829, <https://doi.org/10.1016/j.ympev.2007.06.001> (2008).
- Feng, C. *et al.* Monsoon boosted radiation of the endemic East Asian carps. *Sci China Life Sci* **66**, 563–578, <https://doi.org/10.1007/s11427-022-2141-1> (2023).
- Levin, B. A., Simonov, E., Dgebuadze, Y. Y., Levina, M. & Golubtsov, A. S. In the rivers: Multiple adaptive radiations of cyprinid fishes (*Labeobarbus*) in Ethiopian Highlands. *Scientific Reports* **10**, 7192, <https://doi.org/10.1038/s41598-020-64350-4> (2020).
- Wang, Y. *et al.* Genomic insights into the seawater adaptation in Cyprinidae. *BMC Biol* **22**, 87, <https://doi.org/10.1186/s12915-024-01885-2> (2024).
- Wang, C. *et al.* Genomic features for adaptation and evolutionary dynamics of four major Asian domestic carps. *Science China Life Sciences* **67**, 1308–1310, <https://doi.org/10.1007/s11427-023-2479-2> (2024).
- Nelson, J. A. *et al.* Thermal tolerance of cyprinids along an urban-rural gradient: Plasticity, repeatability and effects of swimming and temperature shock. *J Therm Biol* **100**, 103047, <https://doi.org/10.1016/j.jtherbio.2021.103047> (2021).
- Adeoba, M. I. & Yessoufou, K. Analysis of temporal diversification of African Cyprinidae (Teleostei, Cypriniformes). *ZooKeys* **13**, 141–161, <https://doi.org/10.3897/zookeys.806.25844> (2018).
- Kenthao, A. & Jearanaiprepame, P. Ecomorphological diversification of some barbs and carps (Cyprininae, Cyprinidae) in the Lower Mekong Basin of Thailand. *Zoology* **143**, 125830, <https://doi.org/10.1016/j.zool.2020.125830> (2020).
- Langerhans, R. B., Chapman, L. J. & Dewitt, T. J. Complex phenotype–environment associations revealed in an East African cyprinid. *J Evol Biol* **20**, 1171–1181, <https://doi.org/10.1111/j.1420-9101.2007.01282.x> (2007).

10. Lin, X., Chen, G. & Liu, H. Biology and protection of endangered fish *Tanichthys albonubes*. (2018).
11. Wu, B. *et al.* Single-cell analysis of the amphioxus hepatic caecum and vertebrate liver reveals genetic mechanisms of vertebrate liver evolution. *Nat Ecol Evol* **8**, 1972–1990, <https://doi.org/10.1038/s41559-024-02510-9> (2024).
12. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890, <https://doi.org/10.1093/bioinformatics/bty560> (2018).
13. 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).
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. Zeng, X. *et al.* Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. *Nature plants* **10**, 1184–1200, <https://doi.org/10.1038/s41477-024-01755-3> (2024).
16. 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).
17. Lin, Y. *et al.* quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Horticulture Research* **10**, uhad127, <https://doi.org/10.1093/hr/uhad127> (2023).
18. Tang, H. *et al.* Synteny and collinearity in plant genomes. *Science* **320**, 486–488, <https://doi.org/10.1126/science.1153917> (2008).
19. Saha, S., Bridges, S., Magbanua, Z. V. & Peterson, D. G. Empirical comparison of ab initio repeat finding programs. *Nucleic Acids Res* **36**, 2284–2294, <https://doi.org/10.1093/nar/gkn064> (2008).
20. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **25**, 4–10, <https://doi.org/10.1002/0471250953.bi0410s25> (2009).
21. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573–580, <https://doi.org/10.1093/nar/27.2.573> (1999).
22. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842, <https://doi.org/10.1093/bioinformatics/btq033> (2010).
23. Wu, B. *et al.* The genomes of two billfishes provide insights into the evolution of endothermy in teleosts. *Mol Biol Evol* **38**, 2413–2427, <https://doi.org/10.1093/molbev/msab035> (2021).
24. Wu, B. *et al.* Distinct and shared endothermic strategies in the heat producing tissues of tuna and other teleosts. *Sci China Life Sci* **66**, 2629–2645, <https://doi.org/10.1007/s11427-022-2312-1> (2023).
25. Wu, B. *et al.* Resequencing of reindeer genomes provides clues to their docile habits. *Evolution Letters*, <https://doi.org/10.1093/evlett/qrae006> (2024).
26. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**, 1494–1512, <https://doi.org/10.1038/nprot.2013.084> (2013).
27. Pribelski, A. D. *et al.* Accurate isoform discovery with IsoQuant using long reads. *Nat. Biotechnol.* **41**, 915–918, <https://doi.org/10.1038/s41587-022-01565-y> (2023).
28. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using evidencemodeler and the program to assemble spliced alignments. *Genome Biol* **9**, R7, <https://doi.org/10.1186/gb-2008-9-1-r7> (2008).
29. NCB Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRP553764> (2024).
30. Chen, W. *Barbodes semifasciolatus* isolate GOLD_1, whole genome shotgun sequencing project. Haplotype 1. *Genbank* <https://identifiers.org/ncbi/insdc:JBMAJD000000000> (2025).
31. Chen, W. *Barbodes semifasciolatus* isolate GOLD_1, whole genome shotgun sequencing project. Haplotype 2. *Genbank* <https://identifiers.org/ncbi/insdc:JBMAJE000000000> (2025).
32. Chen, W. *et al.* Assembly and annotation files of (*Barbodes semifasciolatus*). *figshare. Dataset* <https://doi.org/10.6084/m9.figshare.28090190.v3> (2024).
33. Waterhouse, R. M. *et al.* BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mo Biol Evol* **35**, 543–548, <https://doi.org/10.1093/molbev/msx319> (2017).
34. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100, <https://doi.org/10.1093/bioinformatics/bty191> (2018).
35. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079, <https://doi.org/10.1093/bioinformatics/btp352> (2009).

Acknowledgements

This work was supported by the Project of Innovation Team of Survey and Assessment of the Pearl River Fishery Resources (2023TD-10), The Ministry of Agriculture and Rural Affairs Financial Special Project-Fishery Resources and Habitat Survey in the Pearl River Basin (2024–2026), Guangdong Academy of Sciences Program (No. 2024GDASZH-2024010101), Science and Technology Program of Guangzhou (2025A04J3507) and the Natural Science Foundation of China (32300366).

Author contributions

W.C., C.L., B.W. and J.L. conceived and designed the project. W.C., Y.L. and R.Y. collected the samples. C.L. and R.Y. performed the DNA and RNA extraction, library preparation, and genome sequencing. C.L. and B.W. performed the bioinformatics analysis and visualized the results. W.C. and B.W. wrote the manuscript. W.C., C.L., B.W. and J.L. revised and edited the manuscript. All authors have read and approved the final version of manuscript.

Competing interests

The authors declare no competing interests.

Additional information

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

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.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025