scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A chromosome-level genome assembly of the panda loach (*Yaoshania pachychilus*)

Feng Lin[1,2,3,6], Liang-Kun Wang[1,6], Zhi-Yong Yuan[1], Jun Shi[2,3], Cai-Xin Liu[4], Dong-Yi Wu[1], Yao-Quan Han[2,3], Yu-Sen Li[2,3], Kun Qin[5], Yu-Yin Huang[2,3], Da-Peng Wang[2,3] ✉ & Feng Shao[1] ✉

Yaoshania pachychilus **(Gastromyzontidae, Cypriniformes) is a benthic species inhabiting torrential mountain streams, where it attaches to rocks using adhesive pelvic and pectoral fins. Its juvenile black-white coloration has made it popular in the ornamental fish trade, yet wild populations face increasing threats. Here, we present the first chromosome-level genome assembly for this species, generated using PacBio HiFi and Hi-C technologies. The final assembly spans 451.24 Mb, with a contig N50 of 9.87 Mb, and 99.2% of the sequences are anchored onto 25 pseudo-chromosomes. We annotated repetitive elements accounting for 27.8% of the genome and predicted 21,816 protein-coding genes. BUSCO completeness exceeded 97% for both the genome assembly and gene annotation. This reference genome provides a valuable resource for investigating torrent adaptation and pigmentation in *Y. pachychilus*, and also supports broader phylogenomic and adaptive evolution studies across Gastromyzontidae.**

Background & Summary

Dayaoshan Mountain, located in the Guangxi Zhuang Autonomous Region of China, rises to nearly 2000 m and harbors a unique subtropical ecosystem with rich biodiversity. Among its endemic fauna, *Yaoshania pachychilus* stands out as a representative species. This loach belongs to the family Gastromyzontidae (Cypriniformes) and is the sole described member of the genus *Yaoshania*[1,2]. Gastromyzontidae exhibit a flattened ventral body with enlarged pectoral and pelvic fins, which together form an integrated sucker for substrate attachment in fast-flowing waters[3]. This striking adaptation reflects their specialization to torrential mountain stream habitats and also offers promising inspiration for biomimetics and underwater adhesion research[4–6].

At the juvenile stage, *Y. pachychilus* displays a distinctive black-white color pattern resembling that of the giant panda (*Ailuropoda melanoleuca*), earning it the nickname "panda loach" (Fig. 1a). During development, this coloration gradually fades, transitioning to a gray-yellow tone in adulthood, reflecting a clear ontogenetic color shift. Its unique juvenile appearance has made *Y. pachychilus* popular in the aquarium trade and widely recognized as ornamental fishes. However, increasing pressures from overfishing and habitat degradation pose significant threats to wild populations, underscoring the need for strengthened conservation efforts. Accordingly, the species is currently classified as Vulnerable (VU) in the China Biodiversity Red List (2020) and is listed as a Class II protected fish in China.

Current research on *Y. pachychilus* has primarily focused on morphology and mitochondrial genomes[1,2], while the genetic basis of its key adaptive traits remains largely unexplored. As a monotypic genus with a distinct phylogenetic position and highly specialized morphology within Gastromyzontidae, *Yaoshania* still lacks nuclear genomic resources, limiting further investigation into torrent adaptation, pigmentation development, and other fundamental biological questions. In this study, we generated the first chromosome-level reference

[1]Key Laboratory of Freshwater Fish Reproduction and Development (Ministry of Education), School of Life Sciences, Southwest University, Chongqing, 400715, China. [2]Guangxi Academy of Fishery Sciences, Nanning, Guangxi, 530021, China. [3]Engineering Research Center of Hongshui River Rare Fish Conservation, Guangxi Zhuang Autonomous Region, Nanning, Guangxi, 530021, China. [4]State Key Laboratory of Genetic Resources and Evolution & Yunnan Key Laboratory of Biodiversity and Ecological Conservation of Gaoligong Mountain, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, 650201, China. [5]Management Center of the Dayao Mountain National Nature Reserve, Laibin, Guangxi, 546100, China. [6]These authors contributed equally: Feng Lin, Liang-Kun Wang. ✉e-mail: oucwdp@163.com; shaofeng@swu.edu.cn

Library type	Raw read counts	Raw bases (bp)	Q30 (%)	Adapter contamination (%)	reads removed by filtering (bp)
Illumina	356,950,904	53,542,635,600	96.55	0.297	163,181,892
Hi-C	417,657,991	141,497,397,300	97.96	0.340	480,939,080
RNA-skin	23,577,744	7,073,323,200	96.00	1.122	106,841,474
RNA-fin	23,191,327	6,957,398,100	96.04	1.096	101,580,670
RNA-gill	22,155,556	6,646,666,800	95.94	1.230	112,797,901
RNA-brain	26,308,831	7,892,649,300	95.11	3.363	292,029,783
RNA-mixed	33,584,074	10,075,222,200	96.08	0.975	132,801,455

Table 1. Quality control metrics for all libraries. **Note**: The level of adapter contamination was calculated as the ratio of adapter trimmed bases to the total number of sequenced bases.

genome for *Y. pachychilus* using PacBio sequencing platform and Hi-C techniques. The final assembly spans 451.24 Mb with a contig N50 of 9.87 Mb, anchored onto 25 pseudo-chromosomes. We annotated 21,816 protein-coding genes. BUSCO assessment and whole-genome synteny analysis indicate high completeness and continuity of the assembly. This high-quality genome fills a key genomic gap within Gastromyzontidae, providing an essential foundation for studying adaptive evolution and endangered mechanism in *Y. pachychilus*, while also serving as a valuable genomic resource for phylogenetic studies of Gastromyzontidae and for the genetic conservation of endemic fishes in the karst regions of southwest China.

Methods

Ethics statement. All methods were carried out in accordance with relevant guidelines and regulations. The animal study protocol was approved by the Institutional Animal Care and Use Committee (IACUC) of Southwest University, and the study is reported in accordance with the ARRIVE guidelines.

Sample collection. In February 2025, one adult female *Y. pachychilus* was collected using a hand fishing net from a tributary of the Lipu River in Jinxiu Yao Autonomous County, Laibin City, Guangxi Zhuang Autonomous Region, China (24.1533° N, 110.2778° E). The specimen was assigned the voucher ID Yaoshan_001 and is currently deposited in the Southwest University School of Life Sciences (SWU) in Beibei, Chongqing, China. Sample collection was approved by the Management Center of the Dayao Mountain National Nature Reserve. All procedures adhered to the Implementation Rules of the Fisheries Law of the People's Republic of China and the Laboratory Animal Guidelines for the Ethical Review of Animal Welfare (GB/T 35892–2018). The fish was anesthetized using MS-222 and immediately dissected. High-quality genomic DNA (gDNA) was extracted from muscle tissue using the sodium dodecyl sulfate (SDS) method and subsequently purified with a QIAGEN® Genomic Kit. DNA quality was evaluated by agarose gel electrophoresis and a NanoDrop One spectrophotometer (Thermo Scientific, USA). DNA concentration was then quantified using a Qubit® 3.0 Fluorometer (Thermo Scientific, USA).

While the assembly was generated from a single individual, which is common for reference genome projects, this design does not allow assessment of population-level variation or structural polymorphisms within the species. Future studies incorporating resequencing of multiple individuals and phased assemblies will be valuable for investigating population genomic variation and structural diversity in *Y. pachychilus*.

Genome survey. To estimate the genome size and heterozygosity of *Y. pachychilus*, short-read libraries with a 350 bp insert size were constructed using the Nextera DNA Flex Library Prep Kit (Illumina, CA, USA) and sequenced on the DNBSEQ-T7 platform (MGI, Shenzhen, China) with a paired-end 150-bp (PE-150) strategy. Raw sequencing data were processed using fastp v0.20.0[7]. After filtering, 53.38 Gb of clean reads were retained for genome characterization analysis (Table 1). A 19-mer frequency profile was generated using Jellyfish v2.3.0[8], followed by genome property estimation with GenomeScope v2.0[9]. The results indicated an estimated genome size of approximately 428.99 Mb, a heterozygosity rate of 0.19%, and a repeat content of 20.43% (Fig. 1b).

Genome sequencing. For PacBio HiFi sequencing, genomic DNA was sheared into ~15–20 kb fragments using Megaruptor® 3. SMRTbell libraries were constructed with the SMRTbell® Prep Kit 3.0 (Pacific Biosciences, CA, USA) following the manufacturer's instructions, and sequencing was performed on the PacBio Revio platform (Pacific Biosciences, CA, USA). A total of 38.03 Gb of PacBio CCS reads were generated and used for subsequent genome assembly (Table 2). To anchor scaffolds onto chromosomes, a Hi-C library was constructed following a previously published protocol[10]. Briefly, sample cells were first fixed with formaldehyde to preserve chromatin conformation. The DNA was then digested using the restriction enzyme DpnII to create specific cutting sites, and the resulting sticky ends were labeled with biotinylated nucleotides. The labeled DNA fragments were ligated using T4 DNA ligase and further fragmented to 300–700 bp. Finally, the Hi-C library was sequenced on the DNBSEQ-T7 platform (MGI, Shenzhen, China) with a PE-150 strategy. Hi-C sequencing produced 141.50 Gb of raw data. After adapter and low-quality read filtering using fastp v0.20.0[7], 141.02 Gb of clean Hi-C reads were retained for chromosome scaffolding.

RNA extraction and transcriptome sequencing. Total RNA from five different tissues was extracted using TRIzol reagent (Invitrogen, USA). RNA quality and concentration were assessed using a NanoDrop spectrophotometer (Thermo Scientific, USA) and agarose gel electrophoresis. RNA-Seq libraries were prepared using the Illumina TruSeq RNA Library Prep Kit (Illumina, CA, USA) following the manufacturer's instructions, and

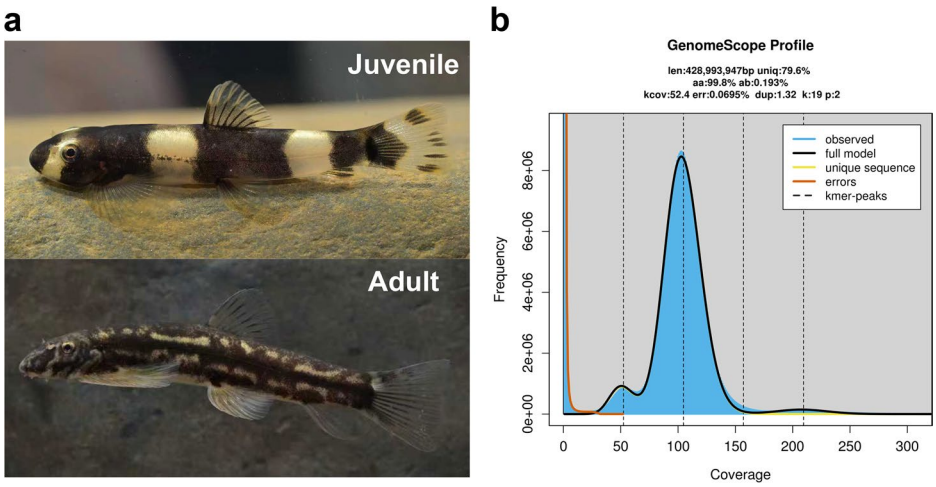

Fig. 1 Genome assembly of *Y. pachychilus*. (**a**) Photographs of juvenile and adult *Y. pachychilus* taken by Liu Caixin; (**b**) The K-mer (K = 19) distribution curve for estimation of the genome size.

Metric	Read counts	Read bases (bp)	Mean read length (bp)	Read N50	HiFi read quality (median)
Value	1,770,568	38,027,259,614	21,477	21,223	32

Table 2. Statistics of PacBio sequencing results.

sequenced on the NovaSeq 6000 platform (Illumina, CA, USA) with a PE-150 strategy. A total of 38.65 Gb of raw transcriptome data was generated and used to support genome annotation.

Genome assembly. The primary assembly of PacBio HiFi reads was generated using HiFiasm v0.19.7[11], producing 313 contigs with a total genome size of 451.24 Mb and a contig N50 of 8.58 Mb. Following the initial assembly, BLASTN v2.15.0 (parameters: -evalue 1e-5 -perc_identity 0.8 -task megablast) searches against the published mitochondrial genome of *Y. pachychilus* were used to identify the mitochondrial sequence[2], and the assembly was screened against the NCBI nt_prok database v20230625 for potential contamination. Approximately 0.2% of the assembled bases were identified as contaminant sequences. To achieve chromosome-level assembly, the HapHiC v1.0.5[12] pipeline was applied. Briefly, Hi-C reads were first aligned to the contig assembly using BWA v0.7.17[13] (parameters: mem -5SP). PCR duplicates were removed using samblaster v0.1.26[14], and secondary, supplementary, and low-quality alignments were filtered using samtools v1.9[15] (parameters: view -F 3340). 99.55% of the Hi-C read pairs were successfully mapped to the genome. The alignments were further filtered with MAPQ $\geq$ 1 (mapping quality $\geq$ 1) and NM $<$ 3 (edit distance $<$ 3). Chromosome scaffolding was then performed using HapHiC (parameters: --quick_view). The order and orientation of contigs were manually reviewed and adjusted using Juicebox v1.11.08[16]. Mis-assembled scaffolds and blocks showing abnormal interaction signals were relocated to appropriate positions based on the Hi-C contact map. As a result, 99.2% of the assembled sequences were anchored onto 25 pseudo-chromosomes, ranging from 12.88 to 25.38 Mb in length (Fig. 2a). However, 77 scaffolds remained unanchored to the chromosomes. The final assembly size remained 451.24 Mb, with a contig N50 of 9.87 Mb and a scaffold N50 of 19.00 Mb (Table 3). Genome completeness was assessed using BUSCO v5.4.7[17] with the actinopterygii_odb10 dataset (version 2024.01.08), showing 97.1% completeness for the final assembly (Table 4).

Repeat annotation. For repeat annotation, we applied a combined strategy of *de novo* prediction and homology-based annotation. For *de novo* identification, tandem repeats were first detected using Tandem Repeats Finder (TRF) v4.09[18] (parameters: 2 7 7 80 10 50 500 -d -h -ngs). A *de novo* transposable element (TE) library was then constructed using RepeatModeler v2.0.4[19] (parameters: -LTRStruct). For homology-based annotation, the RepBase v20181026[20] and Dfam v3.8[21] databases were integrated with the *de novo* TE library (parameters: -e rmblast -gff -s -a -species Cypriniformes), and genome-wide repeat annotation was performed using RepeatMasker v4.1.5[22] along with its built-in RepeatProteinMask module (parameters: -e rmblast -gff -s -a -lib panda-families.fa). The annotated genome was then soft-masked using BEDTools v2.30.0[23] (parameters: maskFastaFromBed -soft). In total, approximately 125.51 Mb of repetitive sequences were identified, accounting for 27.8% of the genome (Table 5). These included 36.37 Mb (8.06%) of DNA transposons, 16.31 Mb (3.61%) of long interspersed nuclear elements (LINEs), 8.58 Mb (1.90%) of short interspersed nuclear elements (SINEs), 20.92 Mb (4.64%) of long terminal repeat elements (LTRs), 27.54 Mb (6.10%) of tandem repeats (TRs), and 21.73 Mb (4.82%) of unclassified repetitive sequences (Fig. 2b).

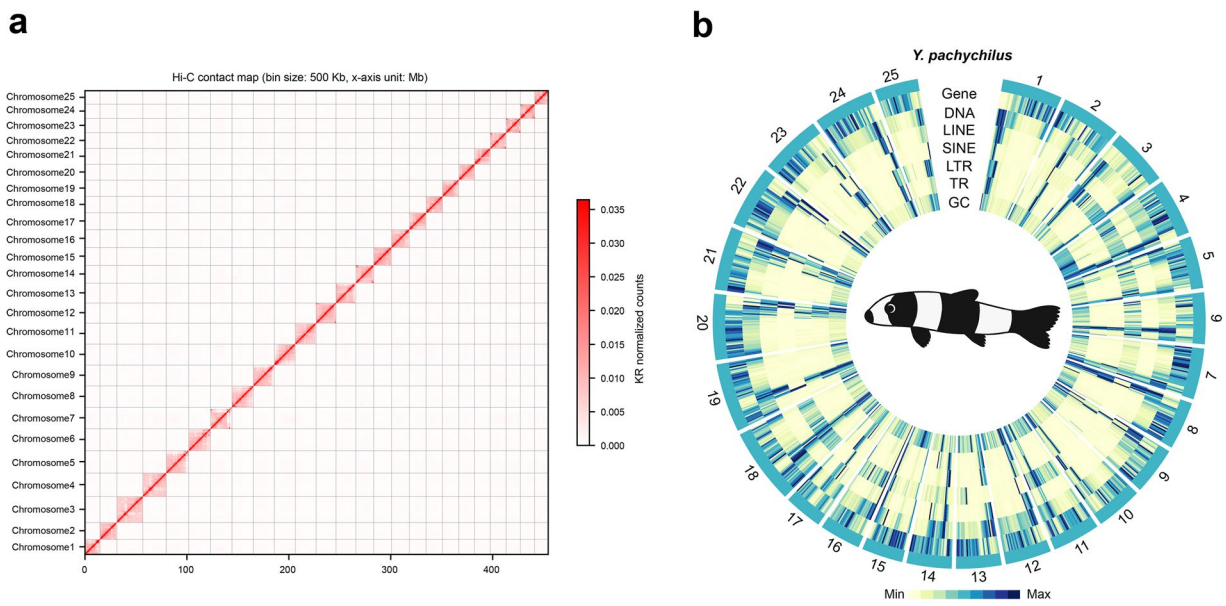

Fig. 2 Landscape of the *Y. pachychilus* genome. (**a**) Genome-wide chromatin interactions at a 500-kb resolution. The color bar shows the intensity of interaction from low (white) to high (red); (**b**) A Circos plot of the main genomic features. From outside to inside, the tracks show the 25 pseudo-chromosomes, gene density, densities of repetitive elements (including DNA transposons, LINEs, SINEs, LTRs, and tandem repeats), and GC content, all calculated in 500-kb windows.

Term	Contig assembly		Hi-C assembly	
	Size (bp)	Number	Size (bp)	Number
N90	1,271,240	64	1,312,469	22
N80	2,996,580	43	3,512,362	19
N70	5,067,209	31	5,171,720	17
N60	6,233,582	23	7,881,937	14
N50	8,584,184	17	9,874,448	11
Max length (bp)	18,578,622	—	18,578,622	—
Total size (bp)	451,236,549	—	451,237,849	—
Total number	—	313	—	300

Table 3. Summary statistics of *Y. pachychilus* genome assembly.

actinopterygii_odb10 (version 2024.01.08)	Genome	Protein
Complete BUSCOs (C)	3,532 (97.1%)	3,555 (97.7%)
Complete and single-copy BUSCOs (S)	3,468 (95.3%)	3,484 (95.7%)
Complete and duplicated BUSCOs (D)	64 (1.8%)	71 (2.0%)
Fragmented BUSCOs (F)	24 (0.7%)	8 (0.2%)
Missing BUSCOs (M)	84 (2.2%)	77 (2.1%)
Total BUSCO groups searched	3,640	3,640

Table 4. Statistics of the genome assembly and annotated proteins for the presence of conserved BUSCO orthologs.

Gene prediction and functional annotation. The gene structure annotation was performed using BRAKER3 v3.0.3[24] (parameters: --gff3), integrating *de novo* prediction, transcriptomic evidence, and homologous protein evidence. Specifically, RNA-seq reads were aligned to the reference genome using HISAT2 v2.1.0[25], and the resulting BAM files were used as transcript evidence. For homology-based prediction, protein sequences from multiple species were collected, including *Danio rerio* (GRCz11, Ensembl), *Oryzias latipes* (ASM223467v1, NCBI), *Astyanax mexicanus* (AstMex3_surface, NCBI), *Triplophysa rosa* (Trosa_1v2, NCBI), *Beaufortia kweichowensis* (ASM1915518v1, NCBI), and *Beaufortia pingi* (ASM4404849v1, NCBI). DIAMOND v2.1.8.162[26] was used for homology-based protein alignment to support gene structure prediction. In the BRAKER3 workflow, transcripts were first assembled from RNA-seq alignments using StringTie v2.2.1[27], and protein-coding

Class		Count	Size (bp)	% in genome
DNA	hAT	69,048	9,458,575	2.1
	TcMar	19,995	3,943,270	0.87
	Others	198,054	22,015,772	4.88
LINE	L1	5,127	3,802,625	0.84
	L2	28,966	7,537,747	1.67
	Others	10,533	4,610,217	1.02
LTR	Gypsy	15,035	5,633,132	1.25
	Copia	642	190,148	0.04
	Others	33,492	14,498,944	3.2
SINE	5S	771	3,781,862	0.84
	tRNA	14,819	2,778,840	0.62
	Others	6,451	1,829,561	0.4
Tandem repeats (TR)		93,049	27,543,210	6.1
Unknown		109,679	21,728,688	4.82
Total		736,576	125,514,390	27.81

Table 5. Statistics of genome repeat sequence of *Y. pachychilus*.

Type	Number
Protein-coding genes	21,816
Exons	221,234
Single exon genes	1,519
Mean gene length (bp)	8,858
Mean exons per mRNA	10.1
Total gene length (bp)	193,801,851
Percentage of genome covered by genes (%)	42.9

Table 6. Statistics of annotated protein-coding genes of *Y. pachychilus*.

RNA Type	Subtype	Copy number	Total length (bp)	Mean length (bp)
rRNA	5S rRNA	3,283	378,968	115.43
	5.8S rRNA	12	1,848	154
	18S rRNA	9	12,159	1,351
	28S rRNA	10	18,707	1,870.7
tRNA	—	6,438	494,680	76.84
miRNA	—	1,215	86,853	71.48
snRNA	splicing	1,582	242,197	153.1
	CD-box	424	75,824	178.83
	HACA-box	67	10,138	151.31
	scaRNA	11	2,199	199.91
	Others	4	273	68.25

Table 7. Statistical results of non-coding RNA annotation.

genes were predicted from these transcripts using GeneMarkS-T[28]. BAMTOOLS v2.5.1[29] was used to generate intron hints via bam2hints. Structural evidence from homologous proteins was generated using ProtHint. GeneMark-ETP[30] integrated transcript, intron hints, and protein homology evidence, and was further combined with *de novo* predictions from AUGUSTUS v3.5.0[31,32]. All evidence was merged and filtered using TSEBRA[33] to produce a high-confidence, non-redundant gene set. In total, BRAKER3 predicted 21,816 protein-coding genes, with 97.7% BUSCO completeness based on the actinopterygii_odb10 dataset (version 2024.01.08) (Table 6).

For non-coding RNA annotation, barrnap v0.9[34] (parameters: --kingdom euk) and tRNAscan-SE v2.0.12[35] (parameters: -E) were used to identify rRNA and tRNA genes, respectively. Other ncRNA types were predicted using Infernal v1.1.4[36] based on the Rfam[37] database. Partial and low-confidence predictions were removed. rRNA genes were further filtered by BLASTN alignment against an rRNA reference database derived from Rfam[37] (parameters: -evalue 1e-5 -task megablast). In total, 13,055 non-coding RNA (ncRNA) genes were annotated, including 3,314 rRNAs, 6,438 tRNAs, 1,215 miRNAs, and 2,088 snRNAs (Table 7).

Functional annotation of protein-coding genes was performed using InterPro v5.63-95.0[38], eggNOG-mapper v2.1.12[39,40], and several public databases, including the Kyoto Encyclopedia of Genes and

Term	Number	Percentage (%)
InterPro	21,247	97.4
eggNOG	21,010	96.3
KEGG	16,227	74.4
Swissprot	19,493	89.4
TrEMBL	21,317	97.7
NR	21,334	97.8
Annotated	21,476	98.4
Total	21,816	—

Table 8. Statistics of the functional annotation of protein-coding genes.

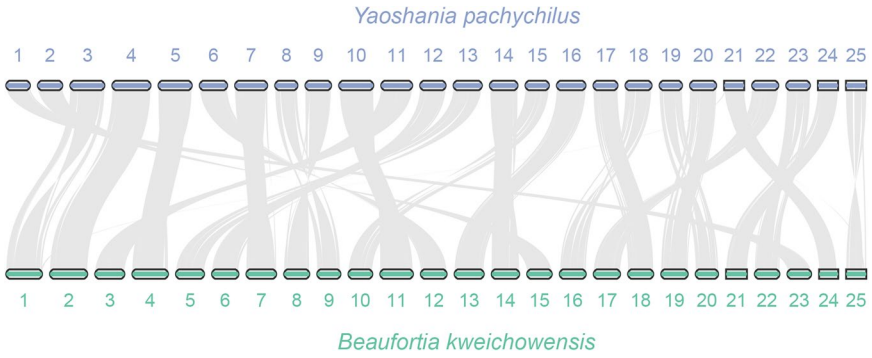

Fig. 3 Genome synteny between *Y. pachychilus* and *B. kweichowensis*. Colored bars represent chromosomes, and grey links denote syntenic regions between the two genomes.

Genomes (KEGG: https://www.kegg.jp), Swiss-Prot (https://www.uniprot.org), TrEMBL (https://www.uniprot.org), and the NCBI non-redundant protein database (NR: https://ftp.ncbi.nlm.nih.gov/blast/db). In total, 21,476 protein-coding genes (98.4% of all predicted genes) were successfully annotated with at least one functional assignment (Table 8).

Data Records

All sequencing data generated in this study have been deposited in the NCBI SRA database under the BioProject accession number PRJNA1347039, which includes the following datasets: SRR35824702–09[41]. The genome assembly is deposited into the GenBank with accession number JBTNTC000000000.1[42]. Furthermore, the genome assembly, annotation files, predicted coding sequences and protein sequences are available at Figshare database (https://doi.org/10.6084/m9.figshare.30403003.v4)[43].

Technical Validation

The final genome assembly of *Y. pachychilus* has a total size of 451.24 Mb with a contig N50 of 9.87 Mb and was successfully anchored into 25 pseudo-chromosomes. Both the genome size and chromosome number are consistent with those reported for two species of *Beaufortia* within Gastromyzontidae[44,45]. The assembled genome size is slightly larger than the 428.99 Mb estimated by k-mer analysis, likely due to the underestimation of repetitive sequence content in *Y. pachychilus* using short-read based genome profiling[46].

Multiple approaches were applied to comprehensively evaluate the quality of the assembled genome. First, reads generated from different sequencing platforms were mapped back to the assembly. HiFi reads were aligned to the genome using Minimap2 v2.28[47], and Illumina DNA reads were aligned using BWA v0.7.17[13]. The mapping rates for HiFi and Illumina DNA reads were 99.85% and 99.53%, respectively. RNA-seq reads were mapped to the genome using HISAT2 v2.1.0[25], with alignment rates ranging from 95.31% to 98.22% across tissues. Secondly, we assessed the completeness of both the genome assembly and annotation using BUSCO v5.4.7, comparing against the Actinopterygii conserved gene set (actinopterygii_odb10, version 2024.01.08). The BUSCO completeness scores were 97.1% for the genome assembly and 97.7% for the annotated protein-coding genes, indicating high completeness for both assembly and annotation. Finally, using annotated gene structure information, chromosome-level synteny analysis between *Y. pachychilus* and its close relative *B. kweichowensis* was performed using JCVI v1.1.125[48] to identify homologous regions. We identified a total of 19,075 single-copy orthologous gene pairs between *Y. pachychilus* and *B. kweichowensis*. The results revealed strong syntenic relationships between the two genomes, further supporting the high quality of the genome assembly and annotation (Fig. 3).

Data availability

All raw sequencing reads from all libraries have been deposited in the NCBI Sequence Read Archive under accession number SRP636306. The genome assembly has been deposited in NCBI GenBank under accession number JBTNTC000000000.1 (https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_054948905.1). Annotation files are available via Figshare (https://doi.org/10.6084/m9.figshare.30403003.v4).

Code availability

All analyses were conducted with publicly available bioinformatics tools according to official guidelines, using default parameters unless noted otherwise. The core code is also available at https://github.com/wangliangkun501/Yaoshania_pachychilus/.

Received: 13 November 2025; Accepted: 26 March 2026;
Published online: 10 April 2026

References

1. Yang, J., Kottelat, M., Yang, J.-X. & Chen, X.-Y. *Yaoshania* and *Erromyzon kalotaenia*, a new genus and a new species of balitorid loaches from Guangxi, China (Teleostei: Cypriniformes). *Zootaxa* **3586**, 173–186, https://doi.org/10.11646/zootaxa.3586.1.16 (2012).
2. Xu, K. & Hu, F. The complete mitochondrial genome of *Yaoshania pachychilus* (CHEN, 1980) (Cypriniformes, Balitoridae). *Mitochondrial DNA B Resour* **1**, 207–209, https://doi.org/10.1080/23802359.2016.1155088 (2016).
3. Wang, J. *et al*. An adhesive locomotion model for the rock-climbing fish, *Beaufortia kweichowensis*. *Sci Rep* **9**, 16571, https://doi.org/10.1038/s41598-019-53027-2 (2019).
4. Wang, J., Xi, Y., Ji, C. & Zou, J. A biomimetic robot crawling bidirectionally with load inspired by rock-climbing fish. *J. Zhejiang Univ. Sci. A* **23**, 14–26 (2022).
5. Wu, J. *et al*. Light-driven soft climbing robot based on negative pressure adsorption. *Chem. Eng. J.* **466**, 143131, https://doi.org/10.1016/j.cej.2023.143131 (2023).
6. Zou, J., Wang, J. & Ji, C. The Adhesive System and Anisotropic Shear Force of Guizhou Gastromyzontidae. *Sci Rep* **6**, 37221, https://doi.org/10.1038/srep37221 (2016).
7. 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).
8. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770, https://doi.org/10.1093/bioinformatics/btr011 (2011).
9. Vurture, G. W. *et al*. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204, https://doi.org/10.1093/bioinformatics/btx153 (2017).
10. Belton, J. M. *et al*. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276, https://doi.org/10.1016/j.ymeth.2012.05.001 (2012).
11. 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).
12. Zeng, X. *et al*. Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. *Nat Plants* **10**, 1184–1200, https://doi.org/10.1038/s41477-024-01755-3 (2024).
13. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* 1303.3997. https://doi.org/10.48550/arXiv.1303.3997 (2013).
14. Faust, G. G. & Hall, I. M. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics* **30**, 2503–2505, https://doi.org/10.1093/bioinformatics/btu314 (2014).
15. Danecek, P. *et al*. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008, https://doi.org/10.1093/gigascience/giab008 (2021).
16. Robinson, J. T. *et al*. Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. *Cell Syst* **6**, 256–258.e251, https://doi.org/10.1016/j.cels.2018.01.001 (2018).
17. Manni, M., Berkeley, M. R., Seppey, M. & Zdobnov, E. M. BUSCO: Assessing Genomic Data Quality and Beyond. *Curr Protoc* **1**, e323, https://doi.org/10.1002/cpz1.323 (2021).
18. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580, https://doi.org/10.1093/nar/27.2.573 (1999).
19. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
20. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**, 11, https://doi.org/10.1186/s13100-015-0041-9 (2015).
21. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA* **12**, 2, https://doi.org/10.1186/s13100-020-00230-y (2021).
22. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* Chapter 4, 4.10.11–4.10.14, https://doi.org/10.1002/0471250953.bi0410s25 (2009).
23. 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).
24. Gabriel, L. *et al*. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Res.* **34**, 769–777, https://doi.org/10.1101/gr.278090.123 (2024).
25. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915, https://doi.org/10.1038/s41587-019-0201-4 (2019).
26. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368, https://doi.org/10.1038/s41592-021-01101-x (2021).
27. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol* **20**, 278, https://doi.org/10.1186/s13059-019-1910-1 (2019).
28. Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res.* **43**, e78, https://doi.org/10.1093/nar/gkv227 (2015).
29. Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P. & Marth, G. T. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**, 1691–1692, https://doi.org/10.1093/bioinformatics/btr174 (2011).
30. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-ETP significantly improves the accuracy of automatic annotation of large eukaryotic genomes. *Genome Res.* **34**, 757–768, https://doi.org/10.1101/gr.278373.123 (2024).
31. Stanke, M., Schöffmann, O., Morgenstern, B. & Waack, S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics* **7**, 62, https://doi.org/10.1186/1471-2105-7-62 (2006).

32. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644, https://doi.org/10.1093/bioinformatics/btn013 (2008).
33. Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector for BRAKER. *BMC Bioinformatics* **22**, 566, https://doi.org/10.1186/s12859-021-04482-0 (2021).
34. Loman, T. *Lund University*. *A novel method for predicting ribosomal RNA genes in prokaryotic genomes*. https://lup.lub.lu.se/student-papers/search/publication/8914064 (2017).
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, https://doi.org/10.1093/nar/25.5.955 (1997).
36. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935, https://doi.org/10.1093/bioinformatics/btt509 (2013).
37. Kalvari, I. *et al*. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.* **49**, D192–d200, https://doi.org/10.1093/nar/gkaa1047 (2021).
38. Paysan-Lafosse, T. *et al*. InterPro in 2022. *Nucleic Acids Res.* **51**, D418–d427, https://doi.org/10.1093/nar/gkac993 (2023).
39. 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, https://doi.org/10.1093/molbev/msab293 (2021).
40. Huerta-Cepas, J. *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–d314, https://doi.org/10.1093/nar/gky1085 (2019).
41. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP636306 (2025).
42. Wang, L. *GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_054948905.1 (2025).
43. Wang, L. Genome annotation, predicted CDS, and protein sequences of *Yaoshania pachychilus*. *Figshare* https://doi.org/10.6084/m9.figshare.30403003.v4 (2025).
44. Deng, Y. *et al*. Genome of the butterfly hillstream loach provides insights into adaptations to torrential mountain stream life. *Mol Ecol Resour* **21**, 1922–1935, https://doi.org/10.1111/1755-0998.13400 (2021).
45. Shen, Q. *et al*. Chromosome-level genome assembly of the butterfly hillstream loach *Beaufortia pingi*. *Sci. Data* **11**, 1260, https://doi.org/10.1038/s41597-024-04144-9 (2024).
46. 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**, 1432, https://doi.org/10.1038/s41467-020-14998-3 (2020).
47. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100, https://doi.org/10.1093/bioinformatics/bty191 (2018).
48. Tang, H. *et al*. JCVI: A versatile toolkit for comparative genomics analysis. *Imeta* **3**, e211, https://doi.org/10.1002/imt2.211 (2024).

Acknowledgements

We extend our sincere gratitude to the staff of the Management Center of the Dayao Mountain National Nature Reserve for their assistance. This study was funded by the Guangxi Natural Science Foundation under Grant No. 2026GXNSFBA00640092, the Project of Financial Funds of the Ministry of Agriculture and Rural Affairs: Investigation of Fishery Resources and Habitat in the Pearl River Basin (ZJZX-04), and Special survey on comprehensive scientific investigation of Guangxi Dayao Mountain National Nature Reserve (LKWT-2025-057).

Author contributions

F.L., F.S. and D.P.W. conceived this project; F.L., J.S., C.X.L. and K.Q. collected samples; D.Y.W., Y.Q.H., and Y,S.L. prepared the sequencing samples; L.K.W. and F.S. analyzed the data; L.K.W. and F.L. wrote the manuscript; F.S., Z.Y.Y. and F.L. revised the manuscript. All authors have read and approved the final manuscript for publication.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.-P.W. or F.S.