scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A chromosome-level genome assembly and annotation of the *Pseudorasbora elongata* (Cypriniformes: Cyprinidae)

Pan Wang[1,4], Denghua Yin[1,4], Min Jiang[1,4], Lingli Xie[1], Jie Liu[1], Yukuan Chen[2], Xinyue Wang[3], Yan Shao[1] & Kai Liu[1,2] ✉

Pseudorasbora elongata **is a unique small fish species endemic to China, distinguished by its striking body coloration resembling a "Chinese ink brush." Due to environmental changes and anthropogenic factors, its wild populations have declined, and it has been listed multiple times as an endangered species. However, the absence of a chromosomal-level reference for *P. elongata* has hindered our understanding of its population genetics and conservation biology. To address this gap, we present a chromosome-level genome assembly of *P. elongata*, generated using PacBio HiFi reads, Oxford Nanopore Technologies, and Hi-C data. We get a genome size of 1.4 Gb with a contig N50 of 34.4 Mb and a scafold N50 of 53.7 Mb. Telomeric sequences were identified at the ends of 42 telomeres across 25 chromosomes. Notably, we observed a high degree of collinearity between our assembly and the *Pseudorasbora parva* genome. This study provides valuable insights into the genetics, genomics, and evolutionary history of *P. elongata*, offering a foundation for future research and enabling the development of genetic conservation strategies.**

Background & Summary

Pseudorasbora elongata is a small fish species endemic to China, first identified in the Lijiang River basin of Guangxi Province[1]. The species is characterized by bright light brown body coloration, with a gray hue on the ventral region. A distinctive feature is the presence of broad black stripes that extend from the snout along the lateral line to the caudal fin, leading to its common designation as the "Chinese writing brush fish"[2] (Fig. 1A). The population of *P. elongata* is relatively limited, and it was classified as a vulnerable species in the Red Data Book of Endangered Animals in China: Fish at the beginning of the 20_{th} century[3]. Currently, its native habitat has nearly vanished, and populations in other areas have also significantly declined due to changes in hydrological conditions from water conservancy engineering construction and pollution[4]. In 2023, it was added to the List of Key Protected Wild Animals in Anhui Province by the People's Government of Anhui Province[5]. *P. elongata* has a narrow distribution range and prefers mountain stream environments with slow-flowing water. In our field resource sampling surveys, *P. elongata* was occasionally found in the Qiupu River, Huangpen River, Xin'an River, and Changjiang River watersheds in Anhui Province (Fig. 1B). Therefore, it is imperative to conduct further research on *P. elongata* to address the ongoing decline in its population and to implement effective strategies for the sustainable development and conservation of its resources.

P. elongata is classified within the order Cypriniformes, subfamily Gobioninae, and genus *Pseudorasbora*, yet it exhibits significant differences from other species in the same genus. Based on external morphological characteristics, we observed that this species closely resembles *Pseudopungtungia tenuicorpus* of the genus *Pseudopungtungia*, as both have a cylindrical body shape and broad black longitudinal stripes on the sides. In contrast, other species within the genus possess a laterally compressed body shape and lack these stripes[6].

[1]Key Laboratory of Freshwater Fisheries and Germplasm Resources Utilization, Ministry of Agriculture and Rural Affairs, Freshwater Fisheries Research Center, Chinese Academy of Fishery Sciences, Wuxi, 214081, China. [2]Wuxi Fisheries College, Nanjing Agricultural University, Wuxi, China. [3]School of Ecology and Environment, Anhui Normal University, Wuhu, China. [4]These authors contributed equally: Pan Wang, Denghua Yin, Min Jiang. ✉e-mail: liuk@ffrc.cn

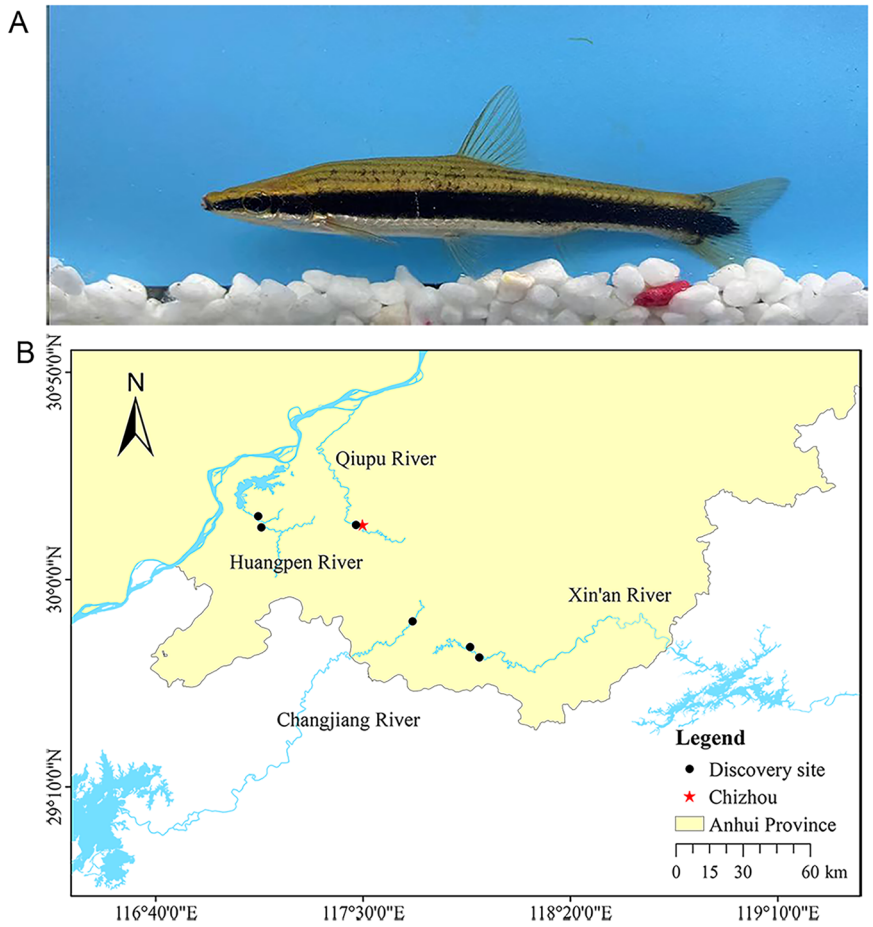

Fig. 1 Morphological characteristics of *P. elongata* (**A**); The round dots are discovery sites and the red star is sample site (**B**).

Several scholars have investigated the phylogenetic relationships within the genus *Pseudorasbora*, revealing that it is not a monophyletic group. *P. elongata* and other species within the genus are distantly related, with *P. elongata* being more closely aligned with the genera *Pungtungia* and *Pseudopungtungia*. This suggests that *P. elongata* may be the first species to diverge from the genus *Pseudorasbora*[7,8]. Despite being classified in the same genus, the morphological differences between *P. elongata* and other species indicate a need for further research to understand the dynamic evolution and biological distinctions within *Pseudorasbora*. Additionally, tracing the evolutionary history of *P. elongata* is essential for a comprehensive understanding of its taxonomy.

Research on *P. elongata* is currently limited, primarily focusing on species differentiation patterns related to specific functional traits, such as life history, and the relationships between functional groups and ecological environments[9–11]. Molecular-level studies on *P. elongata* are scarce, while its close relative, *P. parva*, has garnered significant attention due to its strong invasive capabilities[12–14]. Existing research on the genus *Pseudorasbora* predominantly centers on mitochondrial genomes, with only *P. parva* having its complete genome data published (GenBank assembly accession: ASM2467924)[15]. The genomic data for other species within the genus remain unknown.

In recent years, advances in sequencing technologies have significantly enhanced genetic and genomic research in aquatic species, offering new tools and methods for the study of fish[16,17]. This study aims to leverage genomic sequencing to precisely assemble the genome of *P. elongata*, providing a scientific foundation for the in-depth analysis of its genomic architecture, the spatial distribution of functional genes, the evolutionary trajectories of gene families, and the degree of genetic variability. We present a chromosome-level genome assembly of *P. elongata* using PacBio HiFi reads, Oxford Nanopore Technologies, and Hi-C data. The final assembly spans 1.4 Gb, with the largest contig reaching 66.1 Mb and an N50 value of 34.4 Mb. In summary, this study offers the latest reference genome for *P. elongata*, potentially contributing for future valuable insights into the biological characteristics of the *Pseudorasbora* genus, as well as the genetic diversity and population structure variations across species. It also serves as an essential resource for future conservation research on *P. elongata*.

Methods

Fish material collection and preparation. In July 2023, we collected fish samples from the Chizhou section of the Qiupu River, Anhui Province, China (30.220755°N; 117.498117°E) (Fig. 1B). The experimental fish exhibited a body length of 102.24 mm and a weight of 14.7 g. Multiple tissues (muscle, liver, eye, brain, spleen, gonads, gills, heart and kidney) were harvested from a single fish and flash-frozen in liquid nitrogen for future

Library type	Tissue	Raw data (Gb)	Clean data (Gb)	Average read length (bp)
BGISEQ T7	Muscle	116.66	112.65	150
PacBio HiFi	Muscle	42.68	42.68	15000
Hi-C	Muscle	138.02	136.23	150

Table 1. Statistics of sequencing data for *P. elongata* genome assembly and annotation.

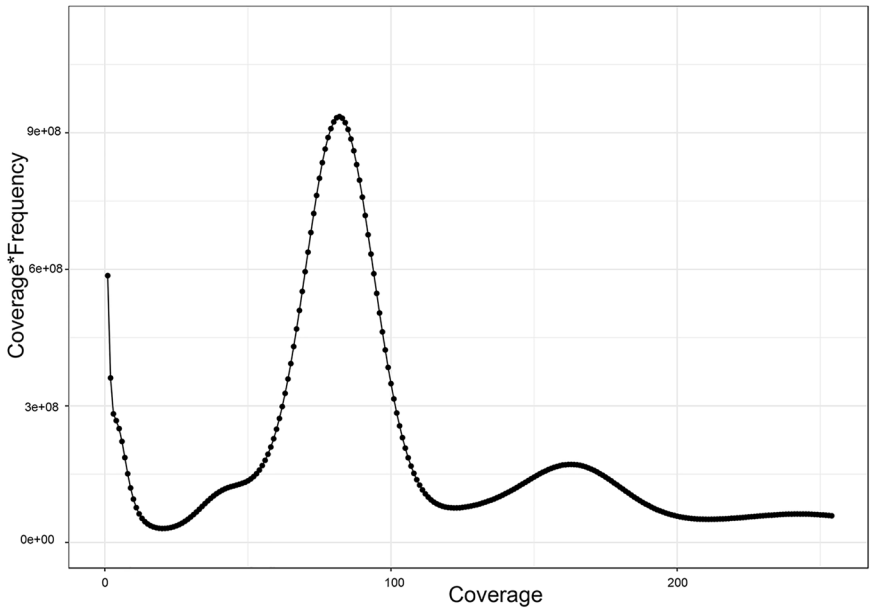

Fig. 2 Distribution of k-mer depth and frequency at 17 k-mer for *P. elongata* genome.

DNA extraction. Genomic DNA (gDNA) for sequencing was obtained from muscle and liver tissues to enable genome assembly.

The experimental samples were obtained during a routine survey under the Monitoring of Aquatic Resources in Key Waters of Anhui Province Project (2023AHNYNC016XQ), with the Special Fishing License ([2023] No. 002) issued by the Anhui Provincial Department of Agriculture and Rural Affairs. Animal welfare and sampling procedures were conducted in compliance with the Guiding Principles on the Humane Treatment of Laboratory Animals (No. 398, 2006) issued by the Chinese Ministry of Science and Technology.

BGISEQ library and PacBio library construction, sequencing and contig-level assembly. DNA was extracted from *P. elongata* muscle samples using the conventional phenol/chloroform method for the whole-genome sequencing (WGS) libraries. The quality of the extracted DNA was assessed using 1% agarose gel electrophoresis and quantified with a Pultton DNA/Protein Analyzer spectrophotometer (Plextech). BGISEQ short reads were generated by constructing a paired-end library with 300–350 bp inserts, adhering to the BGISEQ standard protocol. Then we sequenced DNA from both ends using the BGISEQ-T7 sequencing platform, resulting in a total of 116.66 Gb of raw reads (Table 1). After filtering low-quality reads, short reads, adapters, and redundant sequences using fastp v0.23.2[18] with default parameters, a total of 112.65 Gb of clean reads were obtained (Table 1). Genome size and heterozygosity estimates for *P. elongata* were determined using K-mer analysis with GCE v1.0.0 software[19], resulting in an estimated genome size of 1,204 Mb and a heterozygosity rate of 0.48% (Fig. 2). For PacBio long reads, a 20 kb fragment size library was constructed using the SMRTBell template preparation kit 1.0, following the manufacturer's instructions. Sequencing was performed on the PacBio Revio platform (https://www.pacb.com/) in Circular Consensus Sequence (CCS) mode. After filtering low-quality sequences using the CCS v6.0.0 algorithm with default parameters, we obtained 42.68 Gb of high-precision reads with an N50 value of 15.47 kb (Table S1, Fig. S1).

High-quality genomic DNA was isolated from the muscle tissue using the SDS method. The DNA quality and concentration were tested by 0.75% agarose gel electrophoresis, NanoDrop One spectrophotometer (Thermo Fisher Scientific) and Qubit 3.0 Fluorometer (Life Technologies, Carlsbad, CA, USA). Qualified products were prepared for the next step of library construction. Using the SQK LSK110 ligation kit, libraries were prepared for Oxford Nanopore sequencing according to the standard protocol. We transferred the purified library to primed R9.4 Spot-On Flow Cells and performed sequencing on a PromethION device (Oxford Nanopore Technologies, Oxford, UK). Then, base calling analysis of raw data was performed using the Oxford Nanopore GUPPY software (v0.3.0)[20]. Finally, the total sequencing volume for Oxford Nanopore Technology (ONT) was 34.33 Gb, with 20.8 Gb of ultra-long reads (N50 $\geq$ 100 kb) (Table S2, Fig. S2). Using these HiFi reads, the initial genome

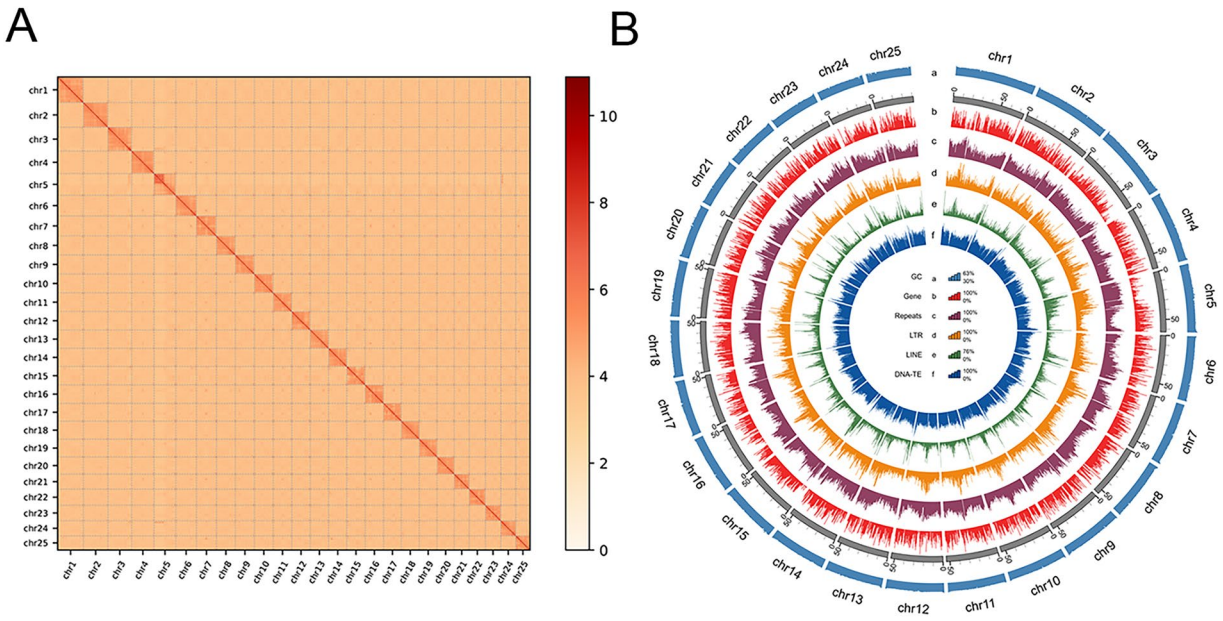

Fig. 3 Heatmap of interactions within and among chromosomes based on Hi-C data (**A**); Features the *P. elongata* genome arranged from the outermost (a) to innermost (f) rings: (a) GC content, (b) gene density, (c) total repeat sequence density, (d) LTR density, (e) LINE density, (f) DNA-TE density. The window size is 500 kb (**B**).

contigs were assembled with the Hifiasm v0.19.6 (https://github.com/chhylp123/hifiasm) and purge_haplotigs algorithms (default settings)[21]. The preliminary assembly produced a 1,419.31 Mb genome, featuring a largest contig size of 52.26 Mb and an N50 length of 27.95 Mb (Table S3).

Hi-C library preparation, sequencing and chromosomal-level assembly. Hi-C data were used to anchor the contigs produced in the earlier step to chromosomes. To facilitate chromatin interactions, 1 g of *P. elongata* liver tissue was cross-linked with 1% formaldehyde for 20 minutes at 20–25 °C, promoting protein coagulation. DNA was then digested with the 4-cutter restriction enzyme (400 units of MboI), and biotinylated nucleotides were used to label the overhangs of the restriction fragments, which were subsequently ligated in a confined space. After reversing the cross-linking, purified ligated DNA was sheared into fragments of 300–500 bp in size. Streptavidin beads were employed to capture ligation junctions, followed by paired-end sequencing on the BGISEQ-T7 platform. Following quality control with fastp v0.23.2[18], low-quality reads and adapters were filtered out, keeping only paired-end reads with lengths exceeding 50 bp. The HiCUP pipeline[22] was used to generate a non-redundant contig interaction matrix, and the contigs were anchored to chromosomes using the 3D-DNA pipeline[23]. Manual error correction was performed using Juicebox Assembly Tools to address any chromosome inversions or translocations[24]. Using the Hi-C data (Table S4), the preliminary assembly was refined and anchored to 25 chromosomes, achieving an anchoring rate of 99.13% (Fig. 3A, Table S5). Using TGS-GapCloser v1.2.0[25] to fill gaps between contigs by leveraging the coverage relationship between ONT ultra-long reads and the already assembled contigs, thereby extending the contigs. This resulted in the first chromosome-level genome assembly of *P. elongata*, with a total genome size of 1.4 Gb (Fig. 3B, Table 2).

Telomere identification. Telomere sequences were utilized the Telomere Identification quarTeT to identify from the genome assembly based on the characteristic motif (CCCTAA/TTAGGG), with a minimum repetition of four times. A total of 42 telomeres were annotated across the 25 chromosomes, and telomeres were detected on both ends of 1 chromosome (Fig. 4A, Table S6).

Repeat sequence annotation. A comprehensive identification of repetitive elements within the *P. elongata* genome was conducted using a dual approach combining homology searches with known repetitive element databases and *de novo* predictions. Repetitive elements identified through ab initio prediction were detected using Tandem Repeat Finder v4.0.9[26] and LTR_FINDER_parallel v 1.0.7[27]. We constructed a *de novo* repeat library using LTR_FINDER_parallel v1.0.7 and RepeatModeler v1.0.11[28], and further predicted novel repeats with RepeatMasker v4.0.9[29]. To identify known repetitive elements, we employed RepeatMasker v4.0.9 and RepeatProteinMask v 4.1.0 (http://www.repeatmasker.org), querying the Repbase database with the genome sequence[30]. The integration of Repbase with our *de novo* transposable element (TE) library revealed that 60.84% of the assembled *P. elongata* genome was annotated as repetitive elements (Table S7, Fig. 4B). Of these, DNA transposons, LINEs, SINEs, and LTRs accounted for 35%, 4.59%, 0.62%, and 15.31% of the total genome, respectively (Table 3).

Gene prediction and annotation. Protein-coding genes within the *P. elongata* genome were predicted using a combination of two distinct approaches: ab initio and homologous methods. Ab initio gene prediction was performed using Augustus v 3.3.2[31] and Genscan[32] with closed species models based on

	Contig Length (bp)	Contig Number	Scaffold Length (bp)	Scaffold Number
N90	12,967,241	40	43,670,000	23
N80	19,199,792	31	49,614,007	20
N70	27,354,052	25	52,278,822	17
N60	30,997,610	20	53,562,927	14
N50	34,397,543	16	53,708,891	12
Total length	1,384,761,903	—	1,384,769,703	—
Number(>=100 bp)	—	244	—	166
Number(>=2 kb)	—	244	—	166
Max length	66,101,788	—	74,637,848	—

Table 2. Genomic features of *P. elongata* assembly.

the organism's taxonomic classification. For homologous prediction, protein sequences from various species, including *Pseudorasbora parva* (GenBank assembly accession: GCF_024679245.1_ASM2467924v1)[15], *Danio rerio* (GenBank assembly accession: GCF_000002035.6_GRCz11)[33], *Megalobrama amblycephala* (GenBank assembly accession: GCF_018812025.1_ASM1881202v1)[34], *Puntigrus tetrazona* (GenBank assembly accession: GCF_018831695.1_ASM1883169v1)[35], *Rhinichthys klamathensis* (GenBank assembly accession: GCF_029890125.1_OSU_Roscu_1.1)[36], were retrieved from the NCBI database. These sequences were aligned to the *P. elongata* genome using tblastn (E-value ≤ 1e-5). The homologous genome sequences were then mapped to their corresponding proteins using miniprot v 0.11-r23[37] and liftoff v 1.6.3 software[38] for accurate splicing. All gene models were merged to eliminate redundancy using MAKER2 v2.31.10[39] and HiFAP (Wuhan OneMore Tech Co., Ltd., https://www.onemore-tech.com/) with default settings. This process identified 25,180 and 24,730 genes, respectively (Table S8).

To annotate the functional roles of these protein sequences, we employed several databases: NCBI nonredundant protein (NR), INTERPRO, Swiss-Prot[40] (http://www.gpmaw.com/html/swiss-prot.html), TrEMBL (http://www.uniprot.org), eukaryotic orthologous groups of proteins (KOG) (https://ftp.ncbi.nih.gov/pub/COG/KOG/), Kyoto Encyclopedia of Genes and Genomes (KEGG) (http://www.genome.jp/kegg/), Pfam and TF databases. Protein domains were analyzed using InterProScan v5.61-93.0[41]. Gene sequences were compared with the KEGG, Swiss-Prot, TrEMBL, KOG, and NR databases using DIAMOND v2.0.14[42] with an E-value threshold of 1e-5. Functional annotation of the final gene sets was successfully achieved for 99.64% (24,642 genes) of the predicted genes (Fig. S3 and Table 4).

Non-coding RNA prediction and annotation. We used INFERNAL based on the rfam v 14.8[43] and miRBase[44] to predict ribosomal RNA (rRNA), microRNA (miRNA) and small nuclear RNA (snRNA) genes. Additionally, tRNAscan-SE v1.3.1 software[45] was applied to identify Transfer RNA (tRNA) genes. In total, we identified four types of noncoding RNAs in the *P. elongata* genome: 1,227 miRNAs, 22,040 tRNAs, 14,628 rRNAs, and 1,368 snRNAs (Table 5).

Genomic collinearity. Genome synteny analysis of *P. elongata* and *P. parva* was performed using WGDI v0.5.6[46] to assess the accuracy of the genome assembly. Comparative syntenic maps were visualized using the jcvi.graphics.karyotype module in JCVI v1.4.1445[47].

Data Records

The sequencing dataset and genome assembly of *P. elongata* have been deposited in the NCBI SRA database under project number PRJNA1173790[48]. The data are as follows: Hi-C data (SRX26407526)[49]; DNBSEQ-T7 genome sequencing data (SRX26407527)[50]; PacBio Revio genome sequencing data (SRX26407528)[51]; Sequel IIe genome sequencing data (SRX26407529)[52]; OXFORD_NANOPORE genome sequencing data (SRX26407530)[53]. The assembled genome was deposited in the NCBI Genome with the accession number GCA_046055825[54]. Genome annotations, along with predicted coding sequences and protein sequences, can be accessed through the Figshare (https://doi.org/10.6084/m9.figshare.27418155)[55].

Technical Validation

Genome assembly and assessment. The quality of the *P. elongata* genome assembly was assessed in terms of sequence correctness, consistency, and integrity. To confirm that the assembly corresponds to the target species, the genome was segmented into 50,000 kb intervals, and each segment was aligned to the NCBI nucleotide database (NT library) using BLAST. The results showed that the median identity with sequences from the same genus reached 91.74%, indicating a low error rate in the assembly (Table S9). Next, Illumina short and long reads were mapped to the assembled genome, and mapping rates and coverage were calculated to evaluate the assembly's accuracy. The short reads achieved a mapping rate of 98.96%, and the long reads provided 99.47% genome coverage (Table S10). Additionally, we assessed k-mer-based quality estimates using Merqury v1.353[56], which incorporated both short and long reads, yielding a high QV value (Table S11, Fig. S4). The completeness of the genome assembly was evaluated using BUSCO v5.7.154[57] and Compleasm v0.2.6[58]. Approximately 99.15% of the core conserved genes were found to be complete (Table 6). The assembly showed a low homozygosity rate (Table S12) and a strong interaction signal around the diagonal of the Hi-C heatmap (Fig. 3A). These results collectively demonstrate the high quality, reliability, and accuracy of the *P. elongata* genome assembly. A high

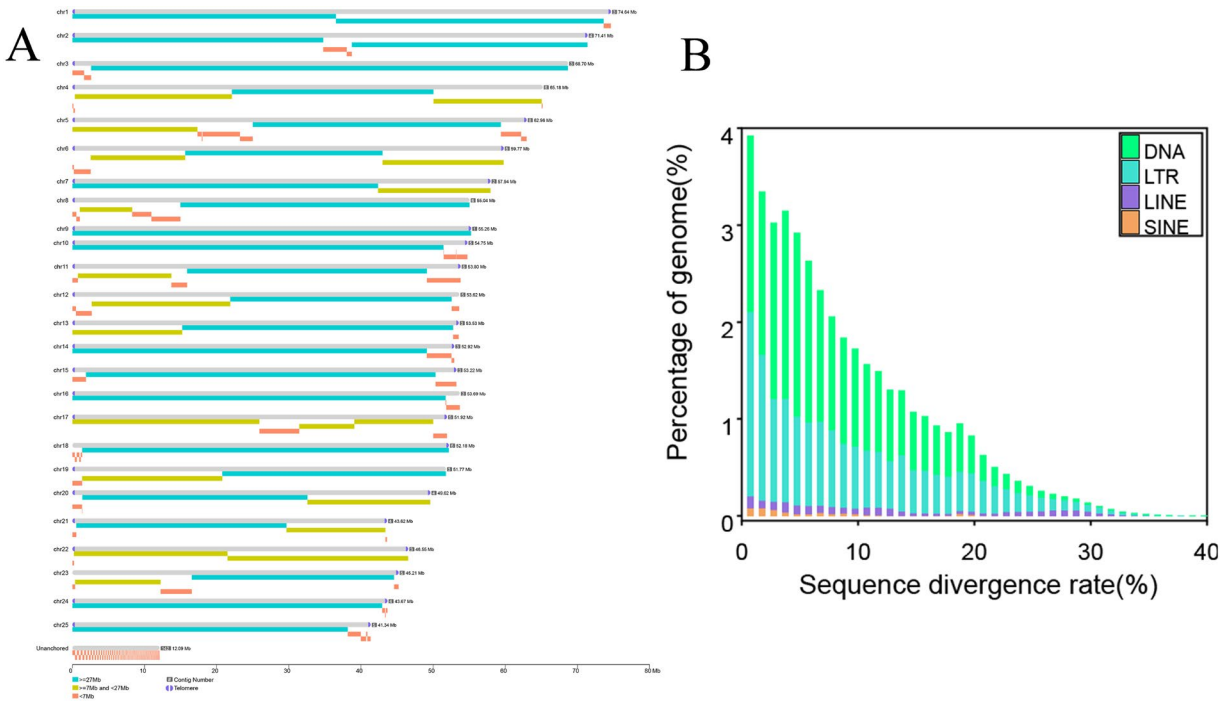

Fig. 4 Distribution of divergence of four TE sequences predicted by *De novo* method (**A**); Chromosome-level genome display for *P. elongata*. (**B**).

	RepBase TEs		TE Proteins		*De novo*		Combined TEs	
	Length (bp)	% in Genome	Length (bp)	% in Genome	Length (bp)	% in Genome	Length (bp)	% in Genome
DNA	357,994,606	25.85	65,294,978	4.72	283,233,427	20.45	484,676,027	35
LINE	55,082,892	3.98	32,924,396	2.38	20,965,626	1.51	63,606,553	4.59
SINE	6,017,310	0.43	0	0	5,033,159	0.36	8,636,987	0.62
LTR	80,895,431	5.84	48,321,364	3.49	161,143,406	11.64	211,966,089	15.31
Satellite	18,496,291	1.34	0	0	4,122,656	0.3	22,217,139	1.6
Simple_repeat	0	0	0	0	4,739	0	4,739	0
Other	12,685	0	0	0	0	0	12,685	0
Unknown	9,947,344	0.72	513	0	102,312,582	7.39	110,307,223	7.97
Total	502,533,863	36.29	146,501,790	10.58	566,978,635	40.94	842,446,516	60.84

Table 3. Statistics of repeated sequence classification for *P. elongata* genome.

		Gene		mRNA	
		Number	Percent (%)	Number	Percent (%)
	Total	24,730	100	24,730	100
Annotated		24,642	99.64	24,642	99.64
	NR	24,598	99.47	24,598	99.47
	SwissProt	21,214	85.78	21,214	85.78
	TrEMBL	24,582	99.4	24,582	99.4
	KOG	19,641	79.42	19,641	79.42
	TF	6,183	25	6,183	25
	InterPro	23,404	94.64	23,404	94.64
	GO	17,696	71.56	17,696	71.56
	KEGG_ALL	24,549	99.27	24,549	99.27
	KEGG_KO	16,540	66.88	16,540	66.88
	Pfam	22,038	89.11	22,038	89.11
Unannotated		88	0.36	88	0.36

Table 4. The Functional annotation of *P. elongata* genome.

Type		Copy	Average length (bp)	Total length (bp)	Percentage of genome (%)
miRNA		1,227	89	108,753	0.007854
tRNA		22,040	76	1,670,471	0.120632
rRNA	rRNA	14,628	118	1,731,055	0.125007
	18S	0	0	0	0
	28S	0	0	0	0
	5.8S	35	154	5,398	0.00039
	5S	14,593	118	1,725,657	0.124617
snRNA	snRNA	1,368	149	204,227	0.014748
	CD-box	247	144	35,482	0.002562
	HACA-box	77	149	11,491	0.00083
	splicing	1,032	150	154,682	0.01117
	scaRNA	12	214	2,572	0.000186

Table 5. Statistics of non-coding RNA annotation for *P. elongata* genome.

Software	Busco		Compleasm	
Type	Proteins	Percentage (%)	Proteins	Percentage (%)
Complete BUSCOs (C)	3,609	99.15	3,633	99.81
Complete and single-copy BUSCOs (S)	3,573	98.16	3,606	99.07
Complete and duplicated BUSCOs (D)	36	0.99	27	0.74
Fragmented BUSCOs (F)	22	0.6	3	0.08
Missing BUSCOs (M)	9	0.25	4	0.11
Total BUSCO groups searched	3,640	100	3,640	100
Database		actinopterygii_odb10		actinopterygii_odb10

Table 6. The evaluation of genome assembly in *P. elongata*.

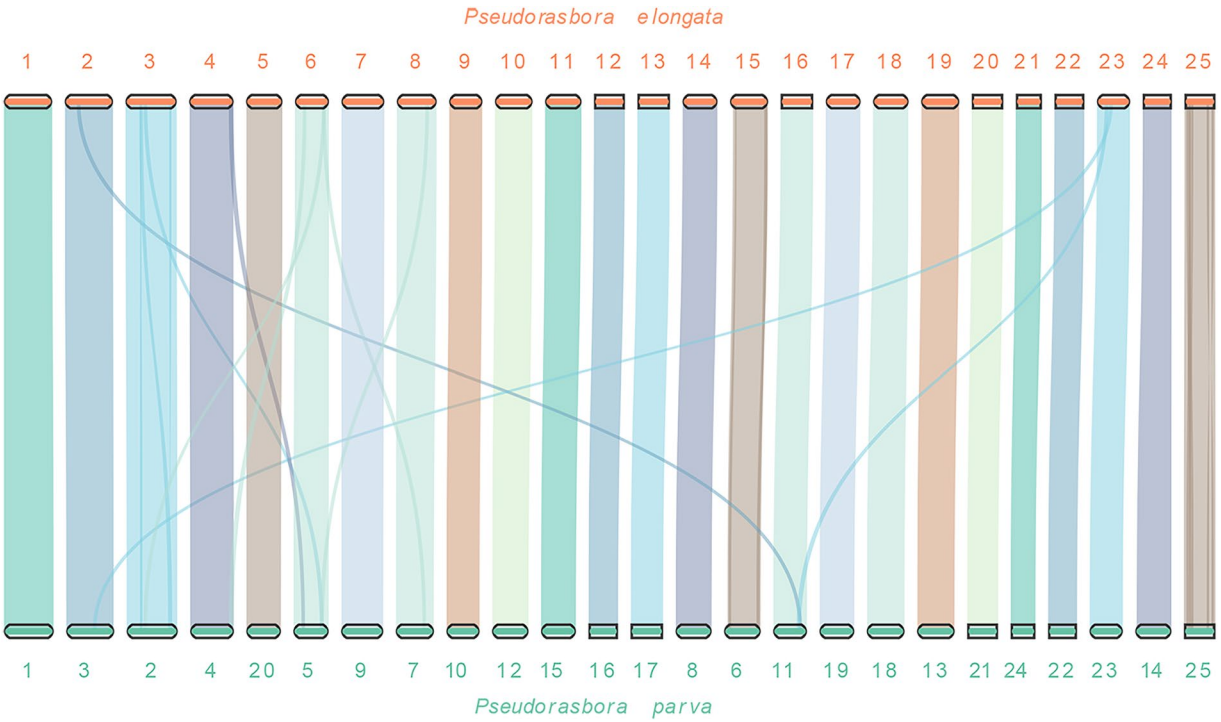

Fig. 5 Chromosome-level genomic synteny between *P. elongata* and *P. parva*.

degree of collinearity was observed between our assembly and the *P. parva* genome (Fig. 5), further supporting the accuracy of the assembly.

Software	BUSCO		Compleasm	
Type	Proteins	Percentage (%)	Proteins	Percentage (%)
Complete BUSCOs	3,548	97.47	3,549	97.5
Complete Single-Copy BUSCOs	3,493	95.96	3,009	82.66
Complete Duplicated BUSCOs	55	1.51	540	14.84
Fragmented BUSCOs	26	0.71	26	0.71
Missing BUSCOs	66	1.81	65	1.79
Total BUSCO groups searched	3,640	100	3,640	100
Database		actinopterygii_odb10		actinopterygii_odb10

Table 7. The evaluation of genome annotation in *P. elongata*.

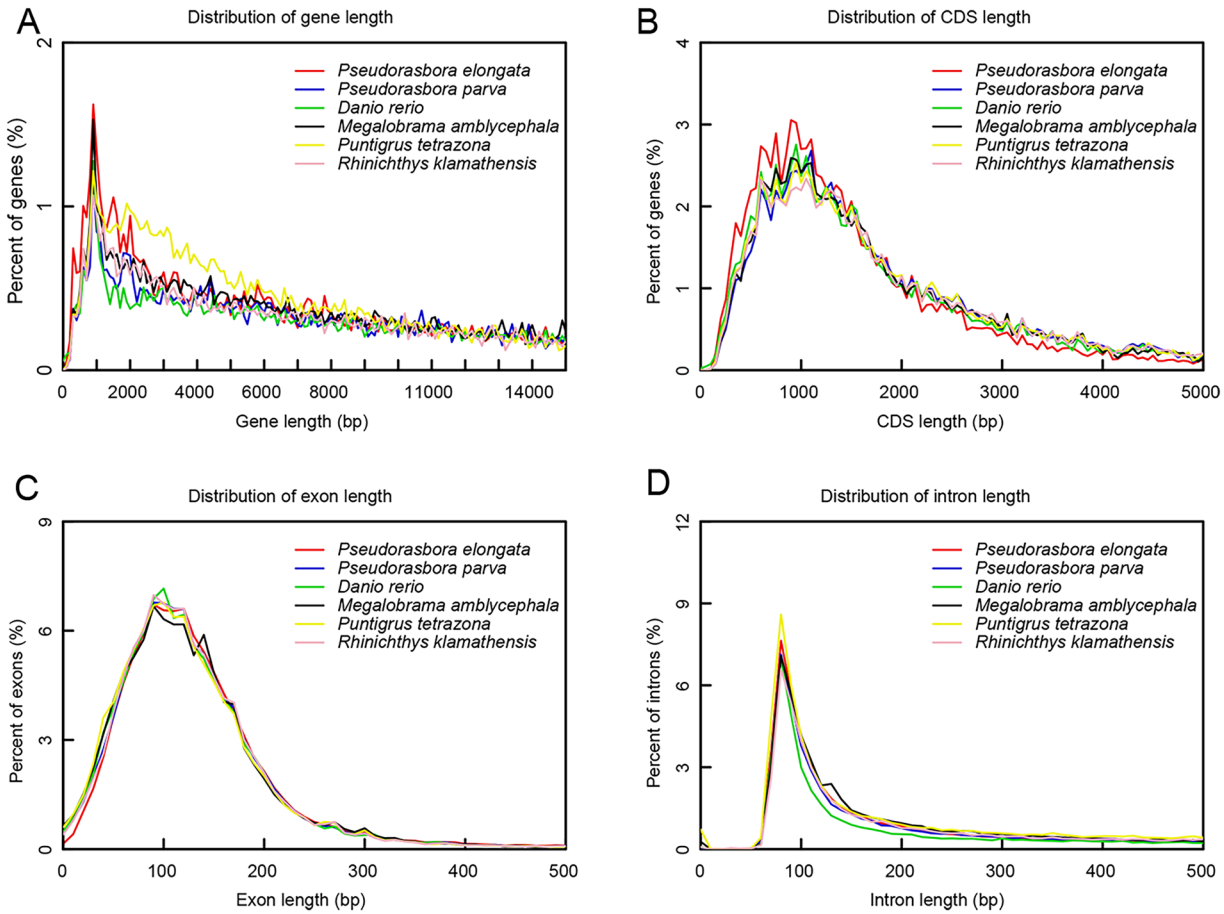

Fig. 6 Comparison of gene length (**A**), CDS length (**B**), exon length (**C**) and intron length (**D**) among five related gene sets between *P. elongata* and *P. parva*.

Structural and functional annotation. We evaluated the quality of the gene annotation using BUSCO v5.7.1 and Compleasm v0.2.6. A total of 3,548 (97.5%) highly conserved core proteins from Actinopterygii were identified in the gene annotation, of which 95.96% were single-copy genes and 1.51% were duplicated. For the remaining conserved genes, 26 (0.71%) were fragmented and 66 (1.81%) were missing (Table 7). Additionally, structural annotations showed a high degree of similarity to those of closely related species (Fig. 6). In summary, the gene annotation demonstrates high accuracy and completeness.

Code availability

No specific code was used in this study. All commands were executed according to the manuals and protocols of the respective bioinformatics software.

Received: 25 November 2024; Accepted: 24 March 2025;
Published online: 01 April 2025

References

1. Cai, D. *et al*. Investigation on fish resources and analysis of species diversity in Lijiang river. *Journal of Guangxi Normal University* **27**, 130–136, https://doi.org/10.3969/j.issn.1001-6600.2009.02.029 (2009).
2. Chen, Y. *Cypriniformes Osteichthys of China (Middle volume)* Vol. 265 (Beijing: Science Press, 1998).
3. Yue, P. & Chen, Y. *China Red Data Book of Endangered Animals: Pisces*. Vol. 247 (Beijing: Science Press, 1998).
4. Kong, D., Cui, G. & Yang, J. Threatened Fishes of the World: *Pseudorasbora elongata* Wu, 1939 (Cyprinidae). *Environ Biol Fish* **76**, 69–70, https://doi.org/10.1007/s10641-006-9025-4 (2006).
5. *Notice of the Anhui Provincial Government on Publishing the List of Key Protected Wild Animals in Anhui Province*. Report No. 00298627-2/202301-00039 (The People's Government of Anhui Province, 2023).
6. Gan, X., Lan, J., Wu, T. & Yang, J. *Primary color map of freshwater fishes in Southern China*. Vol. 110 (Henan Science and Technology Press, 2017).
7. Yang, J. Molecular phylogeny, evolutionary processand biogeography of the gobioninae (pisces:cyprinidae). *Graduate School of Chinese Academy of Sciences (Institute of Hydrobiology)* **01**, 162 http://ir.ihb.ac.cn/handle/342005/15871 (2005).
8. Ma, J. Phylogenetic relationship and genetic diversity of Ustilaginae population in Gobionae. *South China Normal University* (2014).
9. Yang, X. & Lian, Y. Conservation biology teaching combined with the protection of endangered species discussion. *Journal of anhui agricultural science bulletin* **25**, 127–128, https://doi.org/10.16377/j.cnki.issn1007-7731.2019.17.051 (2019).
10. Yang, X., Lian, Y., An, W., Zheng, A. & Yu, P. Morphological and histological observation of digestive system in slender morphological and histological top-mouth gudgeon *Pseudorasbora elongate*. *Chinese Journal of Fisheries* **34**, 45–50, https://doi.org/10.3969/j.issn.1005-3832.2021.02.009 (2021).
11. Yang, X. *et al*. Genetic diversity of *Pseudorasbora elongata* based on mitochondrial *Cyt b* gene and D-loop region sequences. *Journal of Fishery Sciences of China* **30**, 1031–1041, https://doi.org/10.12264/JFSC2023-0153 (2023).
12. Küçük, F. *et al*. Spatio-temporal variation in reproductive characteristics of invasive fish *Pseudorasbora parva* (Temmick & Schlegel, 1846) in the lakes region of Türkiye. *Journal of Fish Biology* **106**, 305–314, https://doi.org/10.1111/jfb.15932 (2025).
13. Adámek, Z., Všetičková, L., Mikl, L. & Šlapanský, L. Habitat preferences and limnological impact of topmouth gudgeon (*Pseudorasbora parva*) population in a small pond. *Biologia* **79**, 3107–3117, https://doi.org/10.1007/s11756-024-01756-9 (2024).
14. Rakauskas, V., Virbickas, T. & Steponėnas, A. Several decades of two invasive fish species (*Perccottus glenii, Pseudorasbora parva*) of European concern in Lithuanian inland waters; from first appearance to current state. *Journal of Vertebrate Biology* **70**, 21048.1–14, https://doi.org/10.25225/jvb.21048 (2021).
15. *NCBI GenBank. Whole Genome Assembly* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_024679245.1/ (2022).
16. Xia, H. *et al*. MultiPrime: A reliable and efficient tool for targeted next-generation sequencing. *Imeta* **2**, e143, https://doi.org/10.1002/imt2.143 (2023).
17. Nguinkal, J. A., Zoclanclounon, Y. A. B., Brunner, R. M., Chen, Y. & Goldammer, T. Haplotype-resolved and near-T2T genome assembly of the African catfish (*Clarias gariepinus*). *Sci Data* **11**, 1095, https://doi.org/10.1038/s41597-024-03906-9 (2024).
18. 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).
19. Liu, B. *et al*. Estimation of genomic characteristics by analyzing k-mer frequency in *de novo* genome projects. *arXiv: Genomics* https://doi.org/10.48550/arXiv.1308.2012 (2013).
20. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biol* **20**, 129, https://doi.org/10.1186/s13059-019-1727-y (2019).
21. 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).
22. Wingett, S. *et al*. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Res* **4**, 1310, https://doi.org/10.12688/f1000research.7334.1 (2015).
23. 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).
24. Dudchenko, O. *et al*. *De novo* assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95, https://doi.org/10.1126/science.aal3327 (2017).
25. 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).
26. 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).
27. Ou, S. & Jiang, N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mob DNA* **10**, 48, https://doi.org/10.1186/s13100-019-0193-0 (2019).
28. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America* **117**, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
29. Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics* **Chapter 4**, Unit 4 10, https://doi.org/10.1002/0471250953.bi0410s05 (2004).
30. Jurka, J. *et al*. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**, 462–467, https://doi.org/10.1159/000084979 (2005).
31. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34**, W435–439, https://doi.org/10.1093/nar/gkl200 (2006).
32. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *Journal of molecular biology* **268**, 78–94, https://doi.org/10.1006/jmbi.1997.0951 (1997).
33. *NCBI GenBank. Whole Genome Assembly* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000002035.6/ (2017).
34. *NCBI GenBank. Whole Genome Assembly* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_018812025.1/ (2021).
35. *NCBI GenBank. Whole Genome Assembly* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_018831695.1/ (2021).
36. *NCBI GenBank. Whole Genome Assembly* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_029890125.1/ (2023).
37. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res* **14**, 988–995, https://doi.org/10.1101/gr.1865504 (2004).
38. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics (Oxford, England)* **37**, 1639–1643, https://doi.org/10.1093/bioinformatics/btaa1016 (2021).
39. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491, https://doi.org/10.1186/1471-2105-12-491 (2011).
40. Boeckmann, B. *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* **31**, 365–370, https://doi.org/10.1093/nar/gkg095 (2003).
41. Zdobnov, E. M. & Apweiler, R. InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics (Oxford, England)* **17**, 847–848, https://doi.org/10.1093/bioinformatics/17.9.847 (2001).
42. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods* **18**, 366–368, https://doi.org/10.1038/s41592-021-01101-x (2021).
43. Griffiths-Jones, S. *et al*. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* **33**, D121–124, https://doi.org/10.1093/nar/gki081 (2005).

44. Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A. & Enright, A. J. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* **34**, D140–144, https://doi.org/10.1093/nar/gkj112 (2006).
45. 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).
46. Sun, P. *et al*. WGDI: A user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *Molecular Plant* **15**, 1841–1851, https://doi.org/10.1016/j.molp.2022.10.018 (2022).
47. Tang, H. *et al*. Synteny and Collinearity in Plant Genomes. *Science* **320**, 486–488, https://doi.org/10.1126/science.1153917 (2008).
48. *NCBI GenBank* https://identifiers.org/ncbi/bioproject:PRJNA1173790 (2024).
49. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRX26407526 (2024).
50. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRX26407527 (2024).
51. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRX26407528 (2024).
52. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRX26407529 (2024).
53. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRX26407530 (2024).
54. *NCBI Assembly* https://identifiers.org/ncbi/insdc.gca:GCA_046055825.1 (2024).
55. Yin, D. H. Chromosome-level genome assembly and annotation of live sharksucker, Pseudorasbora elongata. *Figshare* https://doi.org/10.6084/m9.figshare.27418155.v2 (2024).
56. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21**, 245, https://doi.org/10.1186/s13059-020-02134-9 (2020).
57. Kollmar, M. *Gene Prediction: Methods and Protocols* (Humana Press, 2019).
58. Huang, N. & Li, H. compleasm: a faster and more accurate reimplementation of BUSCO. *Bioinformatics (Oxford, England)* **39**, btad595, https://doi.org/10.1093/bioinformatics/btad595 (2023).

Acknowledgements

This project was funded by the Key Waters Aquatic Biological Monitoring Project in Anhui Province (2023AHNYNC016XQ) and the Central Public-interest Scientific Institution Basal Research Fund, CAFS (NO.2023TD11). We thank Wuhan Onemore-tech Co., Ltd. for their assistance with genome sequencing and analysis.

Author contributions

K.L., D.H.Y. and P.W. conceptualized and developed experimental plan. M.J. and X.Y.W. conducted animal experiments and prepared biological samples. P.W., L.L.X. and M.J. completed the data acquisition, processing, validation. P.W. and L.L.X. wrote the paper. K.L., D.H.Y. and L.J. revised the paper. Y.S. helped process the image. All authors reviewed and endorsed 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/s41597-025-04890-4.

Correspondence and requests for materials should be addressed to K.L.