scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly and annotation of the *Rhinogobio ventralis*, an endangered endemic fish from the Yangtze River

Yan Zhao[1,3], Xingbing Wu[1,3], Weixin Zheng[2], Xiaoli Li[1], Yongjiu Zhu[1], Tingbing Zhu[1] ✉ & Xuemei Li[1] ✉

***Rhinogobio ventralis*, a critically endangered freshwater fish endemic to the upper Yangtze River, holds significant ecological and economic value. Here, we present a high-quality chromosome-level genome assembly of *R. ventralis*. Using a combination of PacBio HiFi long-read sequencing (>99% single-read accuracy), DNBSEQ-T7 short-read sequencing (95.7% of bases ≥Q30), and Hi-C scaffolding, we assembled a 1,039.34 Mb genome comprising 478 contigs with a contig N50 of 18.1 Mb. Hi-C data anchored 97.9% of the assembled sequence onto 25 chromosomes. The final assembly is highly complete (BUSCO completeness 96.6%) and accurate (QV 48.22). Repetitive elements account for 63.44% of the genome, with DNA transposons alone representing 42.31%. We predicted 23,246 protein-coding genes, of which 98.6% have functional annotations. This high-quality genome assembly provides a valuable resource for future studies on the genetics, adaptive evolution, and conservation of *R. ventralis* and other endangered freshwater fishes.**

Background & Summary

Rhinogobio ventralis is a small Cyprinid (order Cypriniformes) endemic to the upper reaches of the Yangtze River, China[1]. It is a rheophilic freshwater species that spawns drifting eggs and has been appreciated locally as a tasty, nutrient-rich food fish with historical commercial value[2]. Over recent decades, wild populations have fallen markedly as a result of river regulation and channel modification, water pollution and excessive harvesting[3]. Reflecting this decline, the species was assessed as endangered (EN) in 2016[4] Red List of China's Vertebrates and was later listed as a Class II National Key Protected Wild Animal in China in 2021[5], highlighting the urgency of germplasm preservation and targeted conservation measures in the upper Yangtze basin.

Previous research on *R. ventralis* has primarily addressed its biology[6,7], natural populations[8,9], and artificial domestication and breeding[10,11], yet genomic resources for *R. ventralis* are presently in short supply. Reference-grade genomes are indispensable for characterizing population genetic structure, evolutionary history, and adaptive traits, as well as for guiding effective conservation strategies[12,13]. Recent advances in sequencing technologies and genome-scale analyses have enabled the assembly of chromosome-level genomes for multiple fish species, such as the big-head schizothorcin (*Aspiorhynchus laticeps*)[12], striped catfish (*Pangasianodon hypophthalmus*)[14], Schizothoracine fish (*Gymnocypris eckloni*)[15], and largemouth bronze gudgeon (*Coreius guichenoti*)[16], facilitating their management and protection.

In light of the ongoing population decline of *R. ventralis*, generating comprehensive genomic information is increasingly critical. In this study, we assembled a chromosome-level *R. ventralis* genome using PacBio long-read sequencing and Hi-C scaffolding. We further performed genome annotation and analyzed genomic features to

[1]Key Laboratory of Freshwater Biodiversity Conservation, Ministry of Agriculture and Rural Affairs, Yangtze River Fisheries Research Institute, Chinese Academy of Fishery Sciences, Wuhan, 430223, China. [2]Frasergen Bioinformatics Co., Ltd, FRASERGEN, Wuhan, 430070, China. [3]These authors contributed equally: Yan Zhao, Xingbing Wu. ✉e-mail: zhutb@yfi.ac.cn; xmli@yfi.ac.cn

Library	Sequencing platform	Clean data (Gb)	Sequence coverage (×)	GC (%)	Q20 (%)	Q30 (%)
Short reads	BGI T7	128.05	123.20	38.9	98.8	95.7
Long reads	PacBio Sequel II	38.41	36.96	39.0	—	—
Hi-C	BGI T7	93.70	90.15	41.0	98.0	93.3

Table 1. Sequencing data used for the genome assembly of *R. ventralis*.

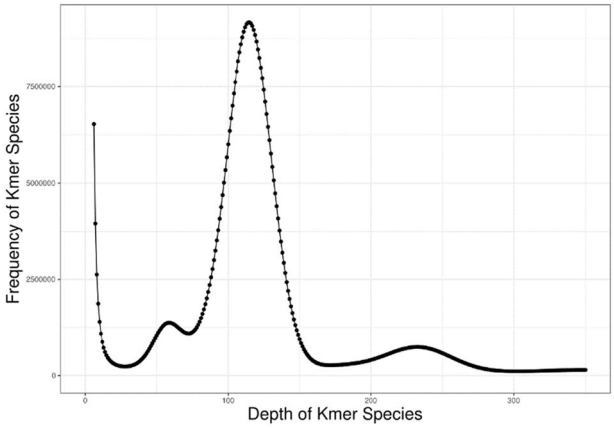

Fig. 1 The 17-mer occurrence distribution of the short reads of the *R. ventralis* genome. The X-axis indicates the sequencing depth of individual Kmer, and the Y-axis shows the number of distinct Kmer observed at each depth.

provide a detailed genetic framework. The resulting high-quality genome assembly not only offers valuable insights into the evolutionary and functional genomics of *R. ventralis*, but also serves as an essential resource to inform future conservation and management strategies for this endangered species.

Methods

Ethics statement. All procedures involving *R. ventralis* were conducted in accordance with the guidelines for the care and use of laboratory animals approved by the Committee on Animal Ethics of the Yangtze River Fisheries Research Institute, Chinese Academy of Fishery Sciences (Approval No. 2022YFI-ZTB-02). These guidelines ensure that all animal care and experimental procedures comply with national and international ethical standards.

Sample collection. A three-year-old female *R. ventralis* (71.7 g, 19.2 cm) was obtained from a domestication and breeding base in Jingzhou, Hubei, China. Following anesthesia with 30 mg/L MS-222, euthanasia was performed by severing the spinal cord immediately after anesthesia to ensure a rapid and humane loss of consciousness. Tissues, including muscle, heart, kidney, gill, liver and spleen, were quickly dissected and snap-frozen in liquid nitrogen for subsequent nucleic-acid extractions. Multiple tissues were sampled to ensure comprehensive transcriptome coverage for gene annotation.

Library preparation and sequencing. *Genomic DNA libraries.* genomic DNA was extracted from muscle tissue. DNA was enzymatically sheared to a target size (300–400 bp), followed by end repair, A-tailing and adapter ligation using the MGIEasy Universal DNA library Prep Set (MGI Tech Co., Ltd., Shenzhen, China). Libraries were PCR-amplified, circularized and subjected to rolling circle amplification to generate DNA nano balls. Libraries were run on the DNBSEQ-T7 (MGI Tech Co., Ltd., Shenzhen, China), producing paired-end short reads with high accuracy and coverage.

PacBio HiFi libraries. High-quality genomic DNA (10 μg) was used to prepare a standard HiFi library for long-read sequencing, in accordance with the SMRTbell Express Template Prep Kit 2.0 protocol (Pacific Biosciences, California, USA). The library was then subjected to sequencing on the PacBio Sequel II platform to generate highly accurate long reads (N50 = 14 kb; >99% accuracy), facilitating assembly of complex and repetitive regions.

Hi-C libraries. To achieve chromosome-level scaffolding, chromatin from multiple tissues was cross-linked with 1% formaldehyde, quenched with glycine, andlysed[16]. Chromatin DNA was digested with MboI, end- labeled with biotin-14-dCTP and ligated with T4 DNA ligase. After reversing cross-links, DNA fragments (300–500 bp) were prepared with end repair, A-tailing, adapter ligation, and biotin-streptavidin enrichment. Libraries were PCR amplified and sequenced on the DNBSEQ-T7 platform to generate high-resolution Hi-C interaction maps.

RNA-Seq libraries. Total RNA from multiple tissues was extracted using Trizol together and assessed for purity and integrity with an Agilent Fragment Analyzer (Agilent Technologies, CA, USA). Following enrichment with

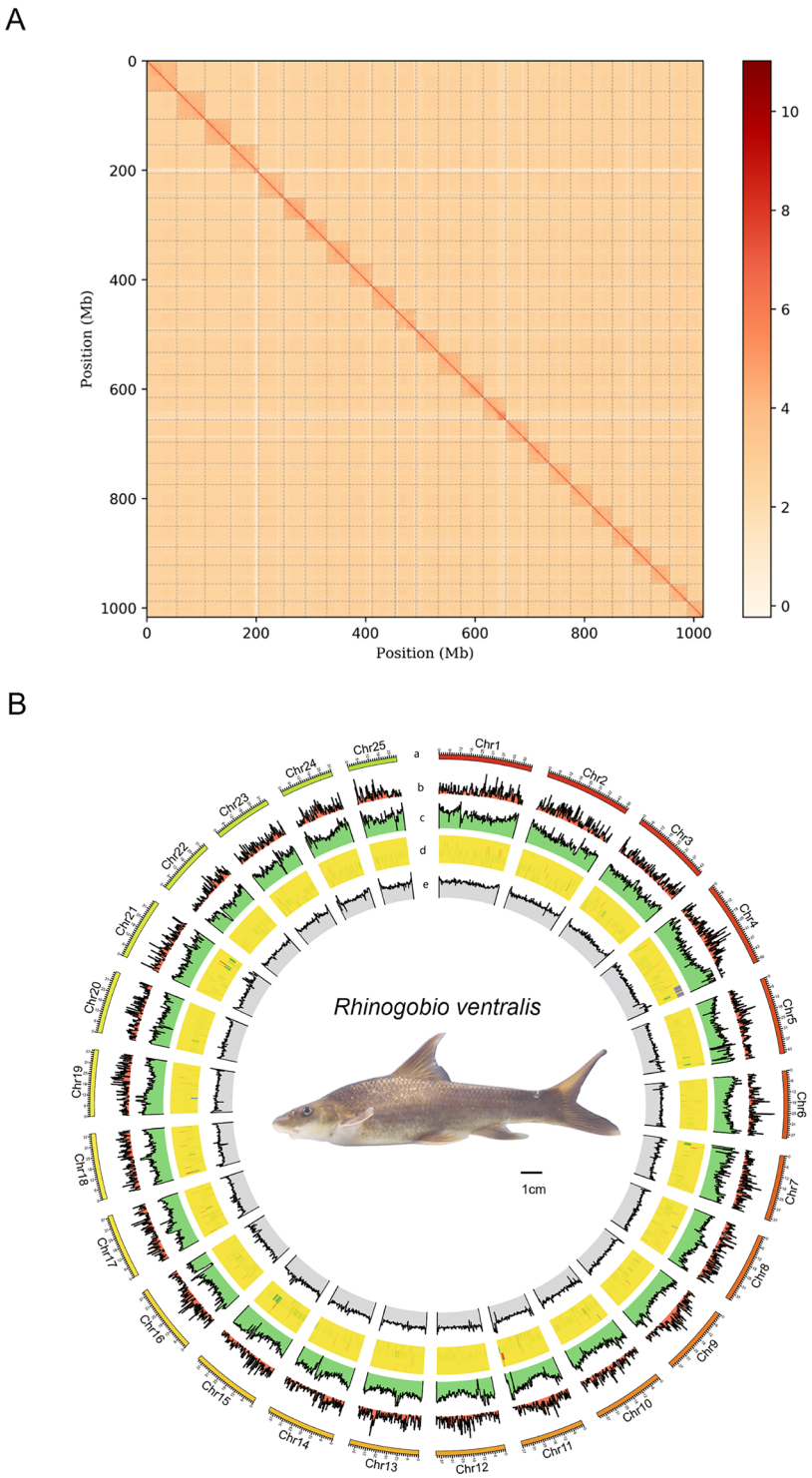

Fig. 2 Hi-C intra-chromosomal contact map (**A**) and circus plot (**B**) of the *R. ventralis* genome assembly. a. chromosome; b. gene density; c. repeat coverage; d. rRNA/SnRNA /tRNA/miRNA; e. GC ratio.

oligo (dT) beads and fragmentation, the Poly(A) + mRNA was reverse-transcribed into cDNA. The resulting cDNA was PCR-amplified and purified to construct transcriptome libraries for sequencing[17].

Genome survey and assembly. The raw DNBSEQ-T7 short-reads were quality-filtered using SOAPnuke[18], removing adaptors, low-quality bases ($Q < 20$), and short reads (<75 bp), resulting in 128.05 Gb of high-quality data with 123.2 coverage, Q20 of 98.8%, Q30 of 95.7%, and 38.9% GC content (Table 1). The genome size was estimated to be ~958.3 Mb with 55.54% repetitive content and 0.16% heterozygosity using a 17-merdistribution (GCE v1.0.2)[19] (Fig. 1).

Name	Length (bp)		Number	
	Scaffold	Contig	Scaffold	Contig
Max	55,202,275	55,202,275	—	—
Total	1,039,347,816	1,039,222,316	227	478
N10	51,581.740	34,103,384	2	3
N20	45,681,185	29.632.619	5	7
N30	41,715.417	28,295,410	7	10
N40	41,275,764	19,128,862	9	15
N50	40,824,786	18,099,218	12	20
N60	39,284,368	13,702,650	15	27
N70	38,888.844	9,880,376	17	35
N80	37,108,327	4,840,823	20	50
N90	33,758,425	1.597.982	23	90

Table 2. The statistics of length and number for the de novo assembled *R. ventralis* genome.

	Assembly No.	Assembly ratio (%)	Annotation No.	Annotation ratio (%)
Complete BUSCOs	3,223	96.1	3,207	95.6
Complete and single-copy BUSCOs	3,188	95.1	3,156	94.1
Complete and duplicated BUSCOs	35	1.0	51	1.5
Fragmented BUSCOs	71	2.1	75	2.2
Missing BUSCOs	60	1.8	72	2.1
Total BUSCO groups searched	3,354	100	3,354	100

Table 3. Assessment of the gene coverage rate in *R. ventralis* genome using BUSCO.

Assembly	Consensus quality value	Error rate
Superscaffold1	49.8814	0.00001
Superscaffold2	49.1989	0.00001
Superscaffold3	48.2190	0.00002
Superscaffold4	53.0761	0.00000
Superscaffold5	48.9585	0.00001
Superscaffold6	48.4775	0.00001
Superscaffold7	48.8273	0.00001
Superscaffold8	49.2782	0.00001
Superscaffold9	48.8969	0.00001
Superscafold10	48.0024	0.00002
Superscafold11	49.5974	0.00001
Superscafold12	49.634	0.00001
Superscafold13	47.8199	0.00002
Superscafold14	48.2723	0.00001
Superscafold15	48.2865	0.00001
Superscafold16	47.9085	0.00002
Superscafold17	47.3686	0.00001
Superscafold18	48.5915	0.00001
Superscafold19	47.7876	0.00002
Superscafold20	46.4342	0.00002
Superscafold21	49.1495	0.00001
Superscafold22	48.7268	0.00001
Superscafold23	49.7171	0.00001
Superscafold24	46.8400	0.00002
Superscafold25	47.7062	0.00002
Assembly final	48.2185	0.00001

Table 4. Evaluation of consensus quality of the *R. ventralis* genome.

PacBio HiFi reads (37.4 Gb; N50 = 14 kb; >99% accuracy) were assembled with hifiasm (v.0.16)[20] to produce long, accurate contigs. And GFA-format sequence graphs were converted to FASTA using gfatools. Hi-C data

Type	Repeat Size	% of genome
Tandem Repeat Finder	116,641,712	11.22
Repeatmasker	341,950,904	32.90
Proteinmask	104,102,948	10.02
De nove	502,820,035	48.38
Total	659,287,935	63.43

Table 5. Summary of repetitive sequences identified in the *R. ventralis* genome.

Type		DNA	LINE	SINE	LTR	Other	Unknown	Total TE
Length (bp)	RepeatMasker TEs	242,416,399	34,673,352	3,951,843	56,802,706	5,267	4,330,687	326,296,532
	RepeatProteinMask TEs	53,816,852	19,217,639	0	31,105,647	0	0	104,102,948
	De novo	328,696,948	55,406,812	2,083,364	107,306,648	0	32,048,316	489,000,840
	Combined TEs	439,705,013	72,701,645	5,316,358	130,444,065	5,267	36,191,614	610,024,674
% in genome	RepeatMasker TEs	23.32	3.34	0.38	5.47	0	0.42	31.39
	RepeatProteinMask TEs	5.18	1.85	0	2.99	0	0	10.02
	De novo	31.63	5.33	0.20	10.32	0	3.08	47.05
	Combined TEs	42.31	6.99	0.51	12.55	0	3.48	58.69

Table 6. Classification and genomic proportion of transposable elements (TEs) in the *R. ventralis* genome.

Type		Copy	Average length (bp)	Total length (bp)	% of genome
miRNA		2,278	88.67	201,982	0.02
tRNA		24,632	76.76	1,890,729	0.18
rRNA	rRNA	27,533	161.50	4,446,573	0.43
	18S	211	1,858.20	392,081	0.04
	28S	210	4,479.49	940,692	0.09
	5S	27,112	114.85	3,113,800	0.30
snRNA	snRNA	2,317	149.06	345,377	0.03
	CD-box	211	178.02	37,563	0.00
	HACA-box	54	155.46	8,395	0.00
	splicing	2,033	146.25	297,321	0.03
	scaRNA	7	207	1,449	0.00

Table 7. Summary of non-coding RNA in the *R. ventralis* genome. Note: miRNA, microRNA; tRNA, transfer RNA; rRNA, ribosomal RNA; snRNA, small nuclear RNA.

Method	Gene set	Number	Average length (bp)				Average exon per gene
			gene	CDS	exon	intron	
De novo	Augustus	28,276	17,424.53	1,504.13	187.45	2,266.52	8.02
	Genscan	24,590	28,585.70	1,656.78	188.52	3,457.58	8.79
Homolog	*Danio rerio*	46,891	13,479.02	1,180.48	204.06	2,570.17	5.79
	Labeo rohita	49,403	11,821.67	1,065.19	194.87	2,408.48	5.47
	Paracanthobrama guichenoti	65,517	8,097.59	747.89	178.66	2,306.76	4.19
	Puntius tetrazona	46,550	10,410.89	915.52	184.97	2,404.12	4.95
RNA-seq		14,087	20,746.23	1,603.48	290.30	1,879.96	10.43
Maker		22,506	23,701	20,547.38	1,663.42	247.10	9.67
PASA update		23,246	21,435.18	1,689.45	258.01	2,109.53	9.87

Table 8. Gene annotation of *R. ventralis* genome via three methods. Note: data sources for other species see Table 9.

(93.70 Gb) was aligned to the primary assembly using Juicer v3[21], enabling identification of chromosome-associated contigs. Chromosome-level scaffolds were generated using the 3D-DNA pipeline with manual adjustment in Juicebox[21,22]. The final assembly comprised 25 chromosomes, which totaled 1017.1 Mb in length and incorporated 478 contigs (Fig. 2A). The assembly achieved a contig N50 of 18.1 Mb and a Hi-C anchoring rate of 97.9% (Table 2). The Hi-C anchoring rate was 97.9%, confirming the high continuity of the assembly (Fig. 2B). Assembly quality

Latin name	Family	Genus	Reference genome fna file link	Reference genome annotation gff file link
Rhinogobio ventralis	Gobionidae	*Rhinogobio*	The present study	The present study
Paracanthobrama guichenoti	Gobionidae	*Paracanthobrama*	https://download.cncb.ac.cn/gwh/Animals/Paracanthobrama_guichenoti_Paracanthobrama_guichenoti_genome_assembly_GWHALNJ00000000/GWHALNJ00000000.genome.fasta.gz	https://figshare.com/ndownloader/files/29047344
Puntigrus tetrazona	Cyprinidae	*Puntigrus*	https://download.cncb.ac.cn/gwh/Animals/Puntigrus_tetrazona_the_genome_of_Puntius_tetrazoa_GWHACFJ00000000/GWHACFJ00000000.genome.fasta.gz	https://figshare.com/ndownloader/files/29047347
Labeo rohita	Cyprinidae	*Labeo*	https://api.ncbi.nlm.nih.gov/datasets/v1/genome/download?filename=GCF_022985175.1.zip&ncbi_phid=939B8A24906E3AF500002A9383C05D53.1.m_3.032	https://api.ncbi.nlm.nih.gov/datasets/v1/genome/download?filename=GCF_022985175.1.zip&ncbi_phid=939B8A24906E3AF500002A9383C05D53.1.m_3.083
Danio rerio	Danionidae	Labeo	https://api.ncbi.nlm.nih.gov/datasets/v1/genome/download?filename=GCF_000002035.6.zip&ncbi_phid=939B8A24906E3AF500005891E061DE0D.1.m_3.025	https://api.ncbi.nlm.nih.gov/datasets/v1/genome/download?filename=GCF_000002035.6.zip&ncbi_phid=939B8A24906E3AF500005891E061DE0D.1.m_3.077

Table 9. Data sources for gene annotation in the *R. ventralis* genome analysis.

was further validated by mapping HiFi reads (98.33% coverage; 99.99% mapping rate) and BUSCO analysis (96.6% complete) using minimap2 (v2.21)[23], with consensus QV scores 48.22 (Table 3, Table 4) using Merqury[24].

Repeat annotation. The identification of repetitive elements was performed as follows: tandem repeats were detected using Tandem Repeats Finder (TRF, v4.09.1)[25]. For transposable elements (TEs), a de novo repeat library was built with LTR_FINDER (v1.0.7)[26] and RepeatModeler (v2.0.1)[27]. This library was integrated with the Repbase TE library[28,29] for comprehensive TE annotation using RepeatMasker (v4.1.2)[30]. RepeatProteinMask (v4.1.2) searched TE protein sequences against WU- BLASTX. Overall, repetitive repeats occupied 63.44% of the genome (Table 5), with DNA transposons representing 42.31%, LTR elements 12.55%, LINE 6.99%, and SINE 0.51% (Table 6), indicating a high proportion of mobile elements contributing to genome complexity.

Non-coding RNA annotation. tRNAs were predicted using tRNAscan-SE (v2.0.9) algorithms[31], rRNAs with RNAmmer (v1.2)[32], and other small RNAs (miRNAs, snoRNAs, snRNAs) using Infernal (v1.1.2)[33] with default parameters against the Rfam (v14.6) database[34] (Table 7). tRNAs and rRNAs accounted for 0.18% and 0.30% of the genome, respectively, while other ncRNAs represented $<0.05\%$, playing essential roles in transcriptional regulation and RNA processing.

Gene prediction and functional annotation. Protein-coding genes were identified through an integrative strategy combining ab initio gene prediction (Augustus v. 3.3.3[35–37], Genscan v1.0[38]), homology-based gene prediction (Exonerate v2.2.0[39]) and RNA-Seq-guided gene prediction (HISAT2 v.2.2.1, Stringtie v.2.1.7, Trinity v2.8.5). The three datasets were integrated with MAKER v3.00[40] and refined using PASA v2.4.1 to annotate UTRs and alternative isoforms. Annotation resulted in 23,246 predicted protein-coding genes, exhibiting average gene, exon, and intron lengths of 21,435 bp, 258 bp, and 2,110 bp, respectively (Table 8). Functional annotation was performed against NR, KEGG[41], GO[42], TrEMBL[43], Swiss-Prot, and InterPro databases using Diamond BLASTP (v2.0.7)[44] and InterProScan (v5.50–84.0)[45] in conjunction with the InterPro[46] database, with 98.6% of predicted genes successfully assigned functional roles (Table 10).

Data Records

The Sequencing reads for *R. ventralis* have been deposited in the NCBI Sequence Read Archive (SRA) under the following accession numbers: SRR28266255for Hi-C data[47], SRR28266256 for T7 data[48], SRR28266257 for PacBio data[49], and SRR28266254 for RNA-seq data[1]. Genome assembly and annotation results for *R. ventralis* were deposited in NCBI GenBank (https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_039654625.1(2024))[50] and Figshare (https://doi.org/10.6084/m9.figshare.29135633) database[51], respectively.

Data overview. The genome of *R. ventralis* was sequenced using multiple platforms including DNBSEQ-T7, PacBio HiFi and Hi-C. A total of 128.05 Gb of short-read data provided 123.2-fold genome coverage, with high quality (Q20: 98.8%, Q30: 95.7%) and a GC content of 38.9%. 37.4 Gb of PacBio HiFi reads with >99% accuracy and an N50 of 14 kb, along with 93.70 Gb of Hi-C data, were used to assemble a final genome of 1,039.3 Mb across 478 contigs. The genome assembly included 25 chromosomes, with an average chromosome length of 40.7 Mb and a Hi-C anchoring rate of 97.9%. The assembly quality was confirmed with 98.33% coverage using HiFi reads and 99.99% mapping rate, achieving 96.6% completeness in BUSCO analysis. Repetitive sequences constitute 63.44% of the genome, with transposable elements accounting for 58.69%, predominantly DNA transposons (42.31%). Gene prediction identified 23,246 protein-coding genes, with an average gene length of 21,435 bp and 98.6% of genes successfully annotated. Non-coding RNAs such as tRNA (0.18%) and rRNA (0.30%) represent a

Database	Number	Percent (%)
InterPro	19,979	86.0
GO	18,449	79.4
KEGG_ALL	22,673	97.5
KEGG_KO	15,804	68.0
Swissprot	20,756	89.3
TrEMBL	22,461	96.6
NR	22,897	98.5
Annotated	22,914	98.6

Table 10. Protein-coding gene prediction for *R. ventralis* genome.

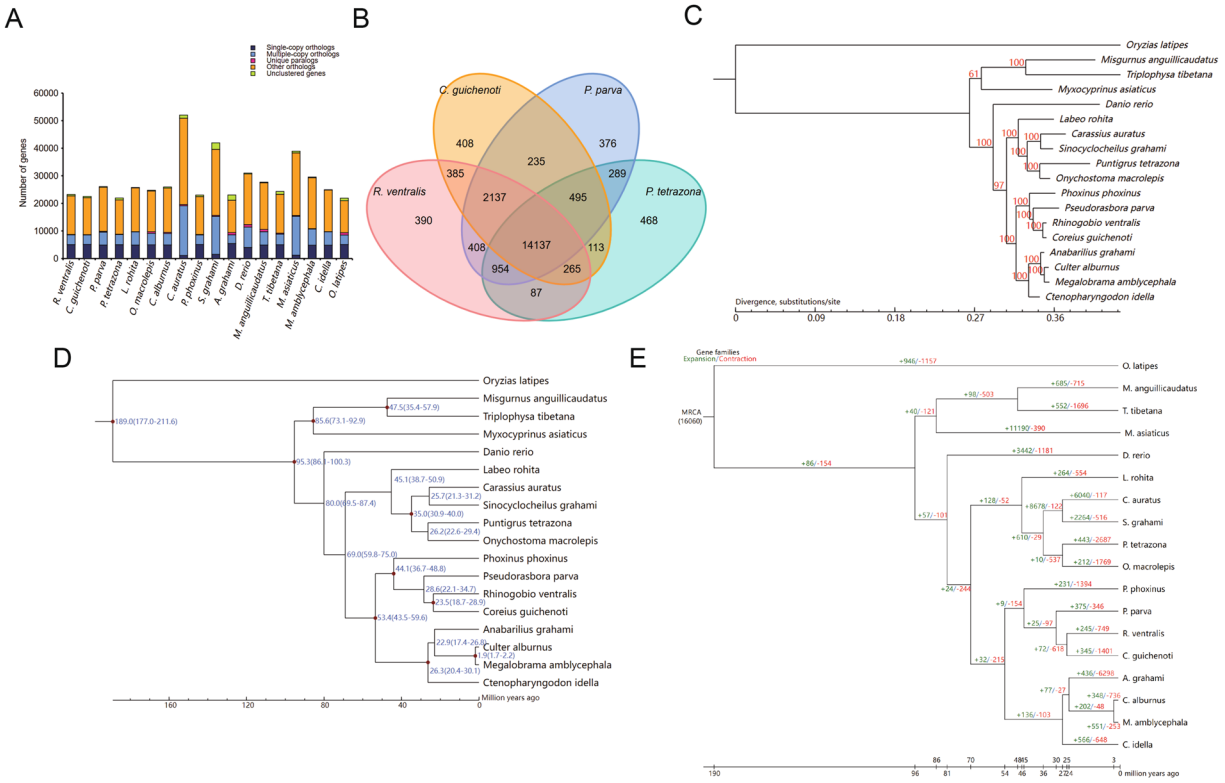

Fig. 3 Evolutionary analysis and divergence time prediction of *R. ventralis*. (**A**) Clusters of orthologous and paralogous gene families in *R. ventralis* and 17 other sequenced fish genomes. (**B**) Venn diagram representing the distribution of shared gene families among *R. ventralis* and three other fishes (*C. guichenoti*, *P. parva* and *P. tetrazona*). (**C**) Phylogenetic tree of *R. ventralis* and 17 other fish species. (**D**) Estimates of divergence time among *R. ventralis* and 17 other fish species. (**E**) Estimates of gene family expansions and contractions. The green and red numbers indicate expanded and contracted gene families, respectively. MRCA represents the most recent common ancestor.

small proportion of the genome. This comprehensive genomic dataset provides a high-quality resource for further research on *R. ventralis*.

Technical Validation

Assessment of genome quality. Genome assembly quality was elevated through a multi-tiered validation strategy. Initial alignment of PacBio long reads and Illumina paired-end reads using minimap2 (v2.21) and BWA (v0.7.17) achieved mapping rates of 99.99% and 99.41% respectively, demonstrating excellent sequence representation. Assembly completeness was further verified with BUSCO (v5.2.2) against the vertebrata_odb10 database, identifying 3,221 (96%) complete orthologs including 3,186 (95%) single-copy and 35 (1%) duplicated BUSCOs, with only 73 (2.2%) fragmented matches. The assembly exhibited minimal contamination (0.41%) and achieved a quality value (QV) of 48.22, equivalent to a base error rate of approximately 1.5×10^{-5}. Gene annotation coverage reached 98.6%, reflecting comprehensive gene prediction.

Evolutionary validation. To validate the genomic assembly of *R. ventralis*, evolutionary analysis was performed using protein data from several representative fish species. We performed gene family clustering using

Analysis	Content	Software	Version	Parameter
Genomic survey	Data quality control	SOAPnuke	v2.1.0	–lowQual = 20,–nRate = 0.005,–qualRate = 0.5
	Genome survey assessment	GCE	v1.0.2	-H 1 -m 1
Third-generation assembly	Third-generation data assembly	HiFiasm	v0.16	default
	Third-generation data alignment	minimap2	v2.11	-ax map-pb
	Second-generation data alignment	BWA	v0.7.17	default
Hi-C assisted assembly	Data alignment and filtering	Juicer	v3	default
	Clustering, constructing interaction matrix, and plotting interaction map	3D-DNA	180922	default
Annotation	Repeat sequence homology prediction	RepeatMasker	4.1.2	-a -nolow -no_is -norna -parallel 3
		RepeatProteinMask	4.1.2	-noLowSimple -pvalue 0.0001
	Self-sequence alignment	RepeatModeler	v2.0.1	-engine_db wublast
	De novo prediction of repeat sequence features	Trf	v4.09.1	2 7 7 80 10 50 2000 -d -h
	De novo annotation	AUGUSTUS	v3.3.3	default
	Homology-based annotation	blastp	v2.0.7	default
		exonerate	v2.2.0	default
	Gene annotation	MAKER	v3.00	default
		PASA	v2.4.1	default
	Non-coding RNA annotation	tRNAscan-SE	v2.0.9	default
		rnammer	v1.2	-S euk -multi -m lsu,ssu,tsu
Comparative genomics	Gene family clustering	blastp	v2.0.7	e-value > 1e-5, identity < 30%, or coverage < 50%
	Multiple sequence alignment	muscle	v3.8.31	default
	Phylogenetic tree construction	RAxML	v8.2.11	default

Table 11. Genomic data analysis workflow and software information.

OrthoFinder (v2.5.4), which identified 18,763 gene families, including 23,246 genes in the *R. ventralis* genome. The analysis revealed 202 sets of strictly single-copy orthologs, with 390 gene families specific to *R. ventralis*. Phylogenetic analysis based on these orthologs, constructed using Muscle (v3.8.31) for multiple sequence alignment and RaxML (v8.2.11) for tree building, showed that R. ventralis is closely related to *Coreius guichenoti*, diverging approximately 23.5 million years ago (Fig. 3). We employed CAFE to analyze gene family expansion and contraction, indicating that 245 gene families expanded and 749 contracted in *R. ventralis*, further supporting the genome's evolutionary accuracy.

Usage Notes

The chromosome-level genome of *R. ventralis* provides a crucial resource for various advanced research areas. It serves as a reference for transcriptomic studies, enabling the investigation of gene expression patterns and functional genomics in this species. Additionally, the genome will support future genome resequencing efforts aimed at analyzing population genetic diversity and structure, which are essential for conservation strategies. It also lays the foundation for studies exploring the relationship between genetic mutations and specific morphological traits, advancing our understanding of genotype-phenotype associations. Furthermore, this high-quality genome is invaluable for molecular breeding applications, offering a basis for identifying genetic markers that can enhance breeding programs and conservation efforts.

Data availability

All dataset from this study was publicly available under BioProject number PRJNA1083632. The raw sequencing reads have been deposited in the NCBI SRA database as follows: genome sequence (SRX23876314), genome survey (SRX23876315), Hi-C data (SRX23876316), and RNA-seq data (SRX23876317). The genome assembly is available at NCBI under accession number JBBJY000000000, and the genome annotation results are deposited in Figshare under accession number 29135633.

Code availability

The details of the software, versions and parameters used during this study for the curation and/or validation of the dataset were shown in Table 11. All instructions and sequences of actions carried out during data processing were performed in accordance with the guidelines and procedures outlined in the manual and protocols of the relevant bioinformatics software.

Received: 30 October 2025; Accepted: 23 February 2026;
Published online: 06 March 2026

References

1. Ding R. The fishes of Sichuan. Sichuan Publishing House of Science and Technology, Chengdu. (1994). (In Chinese).
2. Bao, X. *et al.* Analysis of the ratio of flesh content and nutrient and evaluation of nutritional value for *Rhinogobio ventralis*. *Journal of Hydroecology* **32**, 142–148 (2011).
3. Shao, K. *et al.* Development of sixteen novel polymorphic microsatellite markers in *Rhinogobio ventralis* (Teleostei, Cyprinidae, Gobioninae). *Conserv. Genet. Resour.* **7**, 225–227 (2015).

4. Jiang, Z. *et al*. Red List of China's Vertebrates. *Biodivers. Sci.* **24**, 500–551 (2016). (In Chinese).
5. National Forestry and Grassland Administration, PRC & Ministry of Agriculture and Rural Affair, PRC. The List of National Key Protected Wildlife in China. *Chinese. J. Wildlife* **42**, 605–640 (2021). (In Chinese).
6. Zhou, Q. & He, X. A preliminary study on biology of *Rhinogobio ventralis* in Wujiang River. *Freshw. Fish.* **5**, 11–14 (1992). (In Chinese).
7. Deng, H. & He, X. Study on the biology of *Rhinogobio ventralis* in the Changjiang River. *Journal of Southwest Agricultural University*. (5), 704-708. (In Chinese) (2005).
8. Liu, J. *et al*. Estimation on abundance of *Rhinogobio ventralis* in Yibin reach of the Yangtze River. *Resour. Environ. Yangtze Basin.* **19**, 276–280 (2010). (In Chinese).
9. Xiong, F. *et al*. Estimating population abundance and utilization of *Rhinogobio ventralis*, an endemic fish species in the upper Yangtze River. *Biodiv. Sci.* **24**, 304–312 (2016). (In Chinese).
10. Guan, M. *et al*. Artificial Propagation of *Rhinogobio ventralis*. *Fish. Sci.* **34**, 294–299 (2015). (In Chinese).
11. Zhu, Y. *et al*. Study on broodstock culture, induced spawning and artificial incubation of *Rhinogobio ventralis* (Sauvage et Dabry) in recirculating aquaculture system. *Freshw. Fish.* **48**, 101–106 (2018). (In Chinese).
12. Niu, J. *et al*. Chromosomal-scale genome assembly of the near-extinction big-head schizothorcin (*Aspiorhynchus laticeps*). *Sci. Data* **9**, 556 (2022).
13. Zhou, C. *et al*. The first chromosomal-level genome assembly and annotation of white suckerfish *Remora albescens*. *Sci Data* **11**, 523 (2024).
14. Gao, Z. *et al*. A chromosome-level genome assembly of the striped catfish (*Pangasianodon hypophthalmus*). *Genomics* **113**, 3349–3356 (2021).
15. Wang, F. *et al*. Chromosome-level assembly of *Gymnocypris eckloni* genome. *Sci. Data* **9**, 464 (2022).
16. Rao, S. S. *et al*. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
17. Li, X. *et al*. Chromosome-level genome assembly and annotation of largemouth bronze gudgeon (*Coreius guichenoti*). *Sci. Data* **12**, 76 (2025).
18. Chen, Y. *et al*. SOAPnuke: a MapReduce acceleration-supported software for integrated quality control and preprocessing of high-throughput quencing data. *Gigascience* **7**, 1–6 (2018).
19. Liu, B. *et al*. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome project. arXiv.org arXiv: 1308.2012 (2013).
20. Driguez, P. *et al*. LeafGo: Leaf to Genome, a quick workflow to produce high-quality de novo plant genomes using long-read sequencing technology. *Genome Biol.* **22**, 1–18 (2021).
21. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
22. Dudchenko, O. *et al*. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
23. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
24. Rhie, A. *et al*. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
25. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
26. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, 265–268 (2007).
27. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA* **117**, 9451–9457 (2007). (2020).
28. Jurka, J. Repbase Update: a database and an electronic journal of repetitive elements. *Trends Genet.* **16**, 418–420 (2000).
29. Jurka, J. *et al*. Repbase update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110**, 462–467 (2005).
30. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics Chapter* **4**, Unit 4.10 (2009).
31. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
32. Lagesen, K. *et al*. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108 (2007).
33. Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335–1337 (2009).
34. Kalvari, I. *et al*. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res.* **49**, 192–200 (2021).
35. Stanke, M. *et al*. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, 309–312 (2004).
36. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res.* **33**, 465–467 (2005).
37. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, 435–439 (2006).
38. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
39. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinf.* **6**, 31 (2005).
40. Cantarel, B. L. *et al*. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196 (2008).
41. Kanehisa, M. *et al*. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, 109–114 (2012).
42. Ashburner, M. *et al*. Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
43. Boeckmann, B. *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 365–370 (2003).
44. Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
45. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
46. Mitchell, A. *et al*. The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.* **43**, 213–221 (2015).
47. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR28266255 (2024).
48. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR28266254 (2024).
49. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR28266256 (2024).
50. Yangtze River Fisheries Research Institute, Chinese Academy of Fishery Sciences. Chromosome-level genome assembly of *Rhinogobio ventralis*. *GenBank* https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_039654625.1 (2024).
51. Zhu, T. *Rhinogobio ventralis* genome and annotation. *figshare* https://doi.org/10.6084/m9.figshare.29135633 (2025).

Acknowledgements

We extend our sincere gratitude to Dr. Jiajia Ni and Dr. Li Xie for their valuable contributions to language polishing and data processing in this study. This work was funded by the National Natural Science Foundation of China (Grant No. 32473165), the Financial Special fund of Chinese Ministry of Agriculture and Rural Affairs (Fisheries Resources and Environment Survey in the Key Waters of Southwest China), China Agriculture Research System of MOF and MARA (No. CARS-46), and the Central Public-Interest Scientific Institution Basal Research Fund, CAFS (No. 2023TD61).

Author contributions

Y.Z., X.W., T.Z., and X.L. conceived the study. YJ.Z. collected the samples. X.L. extracted the genomic DNA and conducted sequencing. W.Z., X.L., T.Z. and Y.Z. performed bioinformatics analysis. Y.Z., T.Z., and X.W. wrote the manuscript, all authors read and approved the final manuscript, X. L. is the lead contact for this paper.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to T.Z. or X.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.