scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Telomere-to-telomere gap-free genome assembly of the *Opsariichthys evolans* (Cypriniformes: Cyprinidae)

Pan Wang[1,2,4], Xinyue Wang[2,3,4], Denghua Yin[1,2], Jie Liu[1,2], Min Jiang[1,2] & Kai Liu[1,2,3] ✉

Opsariichthys evolans **is a stream-dwelling fish species endemic to China, with its primary distribution encompassing southeastern China and Taiwan. Initially classified under the genus *Zacco*, it was later reclassified into the genus *Opsariichthys* based primarily on mitochondrial DNA evidence. However, this taxonomic revision remains partially inconclusive due to the absence of whole-genome data. Therefore, we assembled a telomere-to-telomere, gap-free genome assembly of *O. evolans*, consisting of 39 chromosomes with one contiguous sequence per chromosome. The assembly had a total size of 886.9 Mb and a contig N50 of 25.44 Mb. The presence of the telomere repeat was clearly confirmed in the genome. BUSCO assessment confirmed 99.34% genome completeness. Collinearity analysis revealed high synteny between *O. evolans*, *O. bidens* and *Zacco platypus*. Genomic comparisons revealed key candidate genes and related biological pathways potentially responsible for color patterning and hydrodynamic adaptation. The complete *O. evolans* genome provides insights into its genome structure and function, and supports the taxonomic reclassification between the genera *Opsariichthys* and *Zacco*.**

Background & Summary

Opsariichthys evolans represents a characteristic stream-dwelling species, currently classified within the Xenocyprididae family's Opsariichthyinae subfamily under modern taxonomic systems[1] (Fig. 1A). Its natural distribution primarily includes flowing freshwater ecosystems throughout southeastern China and Taiwan. During the breeding season, *O. evolans* exhibits pronounced sexual dimorphism: Males display vivid coloration, with clearly visible lateral body stripes, and develop distinct granular pearl organs on their lips and periorbital regions[2]. A key diagnostic feature of this species is the presence of a blackish-purple snout in adult males during their active life phase[3]. As a keystone species in stream ecosystems, *O. evolans* plays a crucial ecological role in maintaining aquatic food web dynamics, serving as an indicator of water quality, and contributing to biodiversity conservation. However, recent ecological surveys have revealed concerning population trends, including habitat fragmentation and marked declines in abundance across portions of its range, primarily attributable to anthropogenic disturbances such as habitat degradation, aquatic pollution, and competition with invasive alien species[4]. Despite its ecological significance, research attention on *O. evolans* remains limited compared to commercially valuable species *Opsariichthys bidens*, with existing studies primarily focused on taxonomic classification and species identification[5–7].

The widespread minnow tribe Opsariichthine, a significant endemic group of Xenocyprididae fishes in East Asia, has long attracted scholarly attention due to the high morphological similarity among its members[8–10]. As a representative species, *O. evolans* epitomizes the central paradox in opsariichthine systematics: apparent morphological homogeneity likely results from convergent evolution, whereas molecular phylogenies uncover substantially more intricate evolutionary relationships[11,12]. Initially described as *Zacco evolans* by Jordan and Evermann[13] in 1902, traditional taxonomy relied on several key features: the absence of distinct jaw notches,

[1]Wuxi Fisheries College, Nanjing Agricultural University, Wuxi, 214081, China. [2]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. [3]School of Ecology and Environment, Anhui Normal University, Wuhu, China. [4]These authors contributed equally: Pan Wang, Xinyue Wang. ✉e-mail: liuk@ffrc.cn

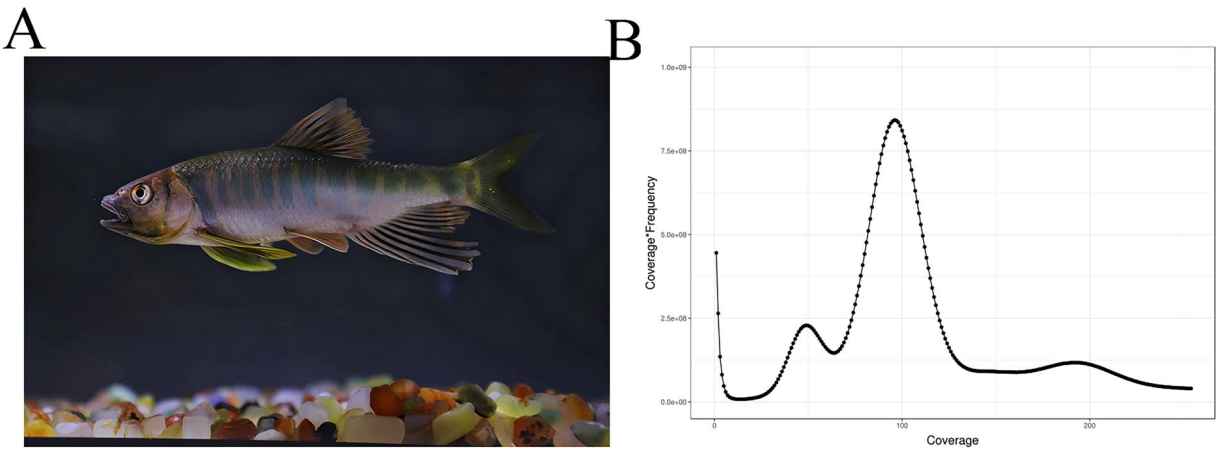

Fig. 1 Morphological characteristics of *O. evolans* (**A**); Distribution of k-mer depth and frequency at 17-mer for *O. evolans* genome (**B**).

the markedly elongated anal fin in mature males, and the clear blue-green vertical stripes along the flanks. After that, this species was frequently confused with the *Zacco platypus* and was even classified as a subspecies of *Z. platypus*[14–16]. It was not until 2005 that Chen and Chang[2] clearly assigned it to the genus *Opsariichthys*, renaming it *O. evolans*. They identified key diagnostic traits distinguishing it from *Z. platypus*: (1) distinct, separate vertical stripes and granular breeding tubercles in *O. evolans*, versus (2) fused stripes forming patches and plate-like tubercles in *Z. platypus*. These characteristics later became widely accepted as critical taxonomic criteria for differentiating *Opsariichthys* from *Zacco*[17,18]. Subsequent molecular phylogenetic studies have consistently validated the species-level distinction of *O. evolans*[11,12,17]. Analyses based on mitochondrial *Cyt b* and complete mitochondrial genomes indicate that the genetic divergence between *O. evolans* and *Z. platypus* meets the criteria for sister-species differentiation[2,5,11,12]. Nevertheless, phylogenetic relationships remain contentious, mitochondrial genomic data support its placement within *Opsariichthys*, whereas recombination activating gene 1 (RAG1) analyses suggest a nested relationship with *Zacco*[16]. Therefore, higher-resolution molecular data are required to validate the genetic basis of these key morphological traits and enhance the accuracy of taxonomic classification.

The whole genome data of *O. bidens* and *Z. platypus* have been published, while the whole genome data of other species in the same genus have not yet been reported. To provide more reliable data for the classification research of Opsariichthine, this study successfully constructed a high-quality chromosome-level reference genome of *O. evolans* (886.9 Mb, 39 chromosomes) based on PacBio and Hi-C technologies, with a BUSCO completeness of 99.34%, which is superior to the existing genomes of *O. bidens* and *Z. platypus*. To address potential taxonomic biases from mitochondrial and nuclear gene markers and to identify key genes and pathways involved in body stripe pigmentation, we performed comparative genomic analyses.This genome assembly not only fills a critical gap in genomic data for congeneric species, but also establishes a high-resolution molecular framework for both taxonomic revision and environmental adaptation studies in opsariichthine fishes.

Methods

Sample collection and ethics statement. In May 2023, field sampling was conducted along the Jingde segment of the Huishui River in Anhui Province, China (geographic coordinates: 38.377254°N, 118.428048°E). The collected male specimen, measuring 122.77 mm in body length and weighing 18.4 g, served as the biological material for this study. Following standard protocols, we aseptically dissected 9 tissue types (including muscle, liver, eye, brain, spleen, gonads, gills, heart and kidney) from the individual, which were immediately preserved in liquid nitrogen for subsequent molecular analyses. DNA extraction was specifically performed on muscle and liver tissues to facilitate whole genome assembly, while RNA extracted from 9 tissue samples were used for genome annotation.

The experimental samples were obtained during a routine survey under the Monitoring of Aquatic Resources in Key Waters of Anhui Province Project (ZF2022-18-0399), 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 Laboratory animal—Guideline for ethical review of animal welfare (GB/T35892-2018) issued by the Technical Committee for Laboratory Animal Sciences of the Standardisation Administration of China (SAC/TC281).

Library construction and sequencing. For whole-genome shotgun (WGS) short-read sequencing, high molecular weight genomic DNA (gDNA) was isolated from muscle tissue using the CTAB (Cetyltrimethylammonium Bromide) method. DNA quality and quantity were assessed using three methods: (i) a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, USA), (ii) gel electrophoresis, and (iii) a Qubit fluorometer (Invitrogen, USA). Libraries were prepared by MGIEasy Universal DNA Library Prep Kit V1.0 (MGI, China) following the standard protocol. Briefly, 1 μg of gDNA was randomly fragmented using the

Library type	Tissue	Raw data (Gb)	Clean data (Gb)	Average read length (bp)
DNBSEQ T7	Muscle	91.55	87.65	150
PacBio HiFi	Muscle	49.19	49.19	20000
ONT	Muscle	58.39	58.39	34386
Hi-C	Muscle	110.23	103.19	150

Table 1. Statistics of sequencing data for *O. evolans* genome assembly and annotation.

Covaris Adaptive Focused Acoustics® (AFA) (Covaris, Brighton, UK). The DNA fragments were size-selected (300–350 bp) with MGIEasy DNA Clean Beads (MGI, China) and used to construct a single strand circle DNA (ssCir DNA) library. The qualified libraries were then subjected to paired-end sequencing on the DNBSEQ-T7 platform. A total of 87.65 Gb of clean data was obtained by removing adapters and low-quality reads using fastp (v0.20.0)[19] with default parameters (Table 1). The genome size (818 Mb) and heterozygosity (0.82%) of *O. evolans* were estimated using k-mer analysis (k = 17) in GCE (v1.0.0) software[20] with DNBSEQ short reads (Fig. 1B).

DNA processing was performed using the same method as described above. For PacBio sequencing, the purified DNA was processed using SMRTbell Prep Kit 3.0 (Pacific Biosciences, USA) to construct SMRTbell libraries. The library preparation involved DNA shearing, end repair, and hairpin adapter ligation to generate circularized templates compatible with single-molecule real-time (SMRT) sequencing. The libraries were then sequenced in Circular Consensus Sequencing (CCS) mode on the PacBio Revio platform (https://www.pacb.com/). PacBio HiFi reads were processed through SMRT Link v11.0.0 (https://www.pacb.com/support/software-downloads) with the following filtering parameters "-minPasses3-minPredictedAccuracy 0.99-minLength 500". Finally, we generated 49.19 Gb of the PacBio HiFi long reads with N50 of 20.18 kb and an average length of 20.19 kb using the CCS algorithm (Fig. S1; Table S1).

High-quality gDNA was extracted from muscle tissue using the SDS method, followed by quality and quantity assessment via Pulsed Field Gel Electrophoresis Systems (Bio-Rad, USA), Qubit fluorometer (Invitrogen, USA) and NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, USA). For Oxford Nanopore Technologies (ONT) ultralong sequencing, we constructed sequencing libraries for Oxford Nanopore platforms (https://nanoporetech.com) using the SQK-LSK110 ligation kit (ONT, UK) with extracted muscle DNA. The purified libraries were loaded onto pre-primed R9.4 Spot-On Flow Cells and sequenced on a PromethION device (ONT, Oxford, UK) with 48-h runs. Approximately 58.39 Gb of data was obtained after performed using the Oxford Nanopore GUPPY software (v0.3.0)[21] (Table 1; Table S2, Fig. S2).

For Hi-C sequencing, 1 g of liver tissue was crosslinked with 40 ml of 2% formaldehyde for 15 minutes at room temperature to facilitate DNA-protein crosslinking. DNA was digested with the 4-cutter restriction enzyme (400 units of MboI) and marked ends with biotin-14-dCTP. Proximity ligation of the cross-linked fragments was then performed with T4 DNA ligase to capture interactions between adjacent chromatin regions. After de-crosslinking and purification, DNA was sheared to 300–500 bp segments. Biotinylated DNA was captured, followed by end repair, A-tailing, adapter ligation, PCR cycle optimization and purification for library construction. After quality control, libraries were pooled based on effective concentration and sequencing requirements for short-read sequencing on the DNBSEQ-T7 platform. Following quality control with fastp software, a total of 110.23 Gb of 150 bp paired-end reads were obtained (Table 1).

Gap-free genome assembly and telomere identification. Combining the HiFi reads and ONT reads ultralong sequencing, the initial contigs were assembled with the Hifiasm (v0.19.6) (https://github.com/chhylp123/hifiasm). After removing redundant contigs using the Purge_Haplotigs algorithms (v1.0.4)[22], we obtained an initial contigs assembly of 917.3 Mb, featuring a largest contig size of 44.42 Mb and a contig N50 length of 22.05 Mb (Table S3). The Hi-C sequencing data were processed using HiCUP pipeline (v0.7.2)[23] to generate non-redundant contig interaction matrices, followed by chromosomal anchoring of contigs through the 3D-DNA pipeline (https://github.com/theaidenlab/3d-dna). The scaffolds were subjected to a meticulous manual assessment and refinement process using Juicebox Assembly Tools (v1.11.08)[24] to address any chromosome inversions or translocations[25]. Using the Hi-C data (Table S4), the preliminary assembly was refined and anchored to 39 chromosomes, achieving an anchoring rate of 98.35% (Fig. 2A, Table S5).

All ONT ultra-long reads were alignd to the chromosome using minimap2 (v2-2.24)[26]. Uniquely mapped reads within the 100 bp regions at both chromosome ends were collected and reassembled into consensus sequences using medaka_consensus (https://github.com/nanoporetech/medaka); then perform blastn (https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/), alignment of these consensus sequences to both ends of each chromosome, replacing the original telomeric sequences with aligned sequences (coverage $\geq$ 90%) based on their positional matches. Using TGS-GapCloser (v1.2.0)[27] to fill gaps between contigs by leveraging the coverage relationship between ONT ultra-long reads and the already assembled contigs, thereby extending the contigs. Pilon (v1.23)[28] was used to correct the errors of the extended and gap-filled genome using short read sequencing. The gap-free level was reached after three rounds of gap filling.

Then, 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 within the 50 kb region at each end of the chromosome. A total of 51 telomeres were annotated across the 39 chromosomes (Fig. 2B, Table S6). The final assembly yielded a 0.89 Gb telomere-to-telomere (T2T), gap-free level genome of *O. evolans*, with each chromosome represented by a single contig (Fig. S3, Table 2).

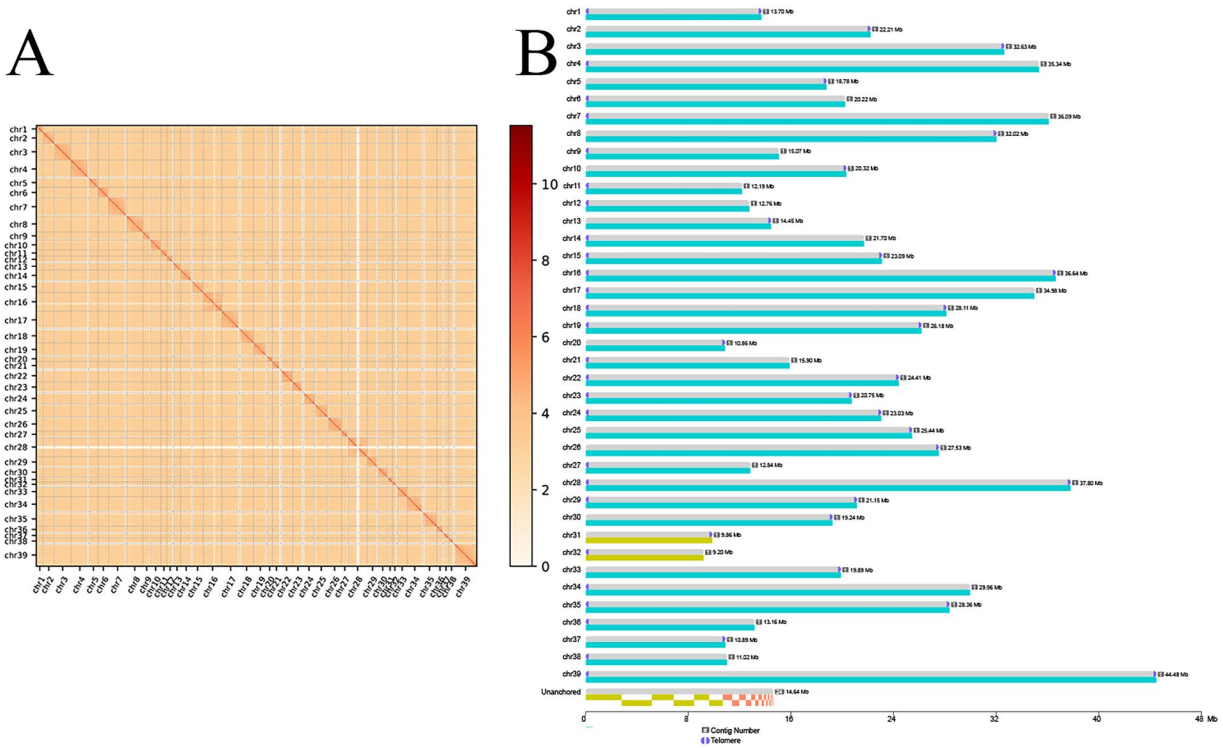

Fig. 2 Heatmap of interactions within and among chromosomes based on Hi-C data (**A**); Chromosome-level genome display for *O. evolans* (**B**).

	Contig Length (bp)	Contig Number
N90	12,760,410	33
N80	18,777,144	26
N70	20,317,145	22
N60	22,206,782	18
N50	25,440,736	14
Total length	886,877,176	39
Number(>=100 bp)	—	63
Number(>=2 kb)	—	63
Mean length (bp)	14,077,415	—
Max length (bp)	44,478,744	—

Table 2. Genomic features of *O. evolans* assembly.

Genomic structure and functional annotation. Through combined homology-based, de novo sequencing approaches, we identified 398 Mb of repetitive sequences representing 44.91% of the whole genome. The repetitive sequences consisted of DNA transposons (242 Mb; 27.34%), LINEs (32 Mb; 3.65%), SINEs (6 Mb; 0.70%), and LTR retrotransposons (77 Mb; 8.73%) (Table 3; Fig. 3A). The annotation was performed using the following integrated approaches: RepBase transposable elements (TEs) and TE proteins were obtained by annotating the genome using RepeatMasker (v4.0.9)[29] and RepeatProteinMask (v4.1.0) (http://www.repeatmasker.org) respectively based on the RepBase database[30]; De novo prediction was performed by constructing a non-redundant repeat library using RepeatModeler (v1.0.11)[31] and LTR-FINDER_parallel (v1.0.7)[32] followed by genome annotation with RepeatMasker; Combined TEs represent the final non-redundant repetitive sequences generated by integrating and deduplicating the results from these three methods (Table S7).

Subsequently, Protein-coding genes within the *O. evolans* genome were predicted using a combination of three distinct approaches: homologous methods, ab initio and RNA sequencing methods. To annotate homologous sequences, we aligned the *O. evolans* genome against reference protein sequences from four species (*Opsariichthys bidens*[33], *Zacco platypus*[34], *Danio rerio*[35], and *Hypophthalmichthys molitrix*[36]) using TBlastN (E-value $\leq$ 1e-5)[37], followed by precise splicing annotation with miniprot (v0.11-r23)[38] and liftoff (v1.6.3)[39]. Additionally, ab initio gene prediction was conducted using Augustus (v3.3.2)[40] and Genscan[41], guided by the organism's taxonomic classification. Total RNA was extracted from 9 tissues using TRIzol reagent (Tiangen,

Methods		DNA	LINE	SINE	LTR	Satellite	Unknown	Total
RepBase TEs	Length (bp)	160,371,500	26,322,705	4,672,698	35,993,527	11,625,107	4,044,207	225,151,338
	Percentage (%)	18.08	2.97	0.53	4.06	1.31	0.46	25.39
TE Proteins	Length (bp)	16,884,760	13,870,562	0	14,627,739	0	2,205	45,363,356
	Percentage (%)	1.9	1.56	0	1.65	0	0	5.11
De novo	Length (bp)	148,282,232	14,185,004	2,564,178	55,308,875	2,629,173	54,493,393	270,940,213
	Percentage (%)	16.72	1.6	0.29	6.24	0.3	6.14	30.55
Combined TEs	Length (bp)	242,448,719	32,333,261	6,178,237	77,380,689	13,572,430	57,477,058	398,336,441
	Percentage (%)	27.34	3.65	0.7	8.73	1.53	6.48	44.91

Table 3. Statistics of repeated sequence classification for *O. evolans* genome.

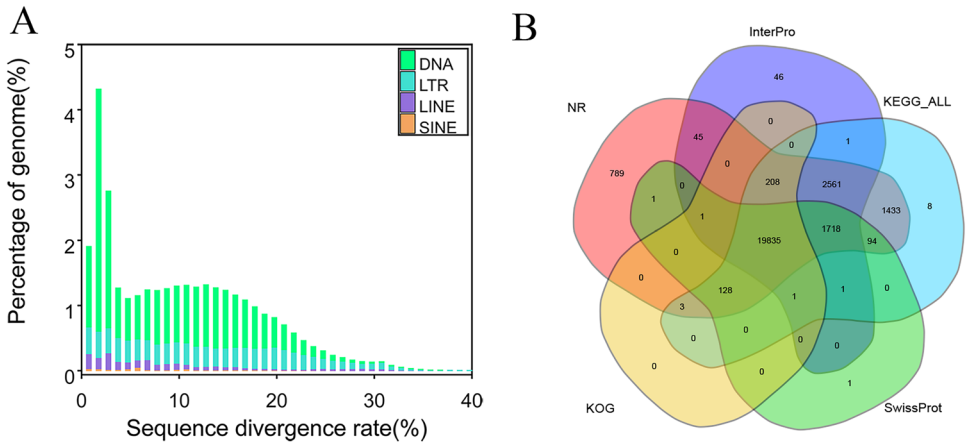

Fig. 3 Distribution of divergence of four TE sequences predicted by De novo method (**A**); The venn diagram of annotation results of five databases for *O. evolans* genome (**B**).

Methods/Tools		Gene number	Average gene length (bp)	Average exons per gene	Average length (bp)		
					CDS	Exon	Intron
De novo	Genscan	33,623	17,261	7.68	1,463	190.48	2,365
	Augustus	25,068	12,807	8.11	1,459	179.93	1,597
Homolog	*Z. platypus*	39,186	14,654	8.87	1,602	180.57	1,658
	O. bidens	40,551	10,699	7.2	1,462	203.16	1,490
	H. molitrix	33,638	15,729	8.32	1,622	194.93	1,927
	D. rerio	63,970	22,205	11.14	2,058	184.84	1,988
Liftoff	*Z. platypus*	23,740	15,103	8.92	1,579	177.11	1,708
	O. bidens	29,301	11,277	8.62	1,540	178.7	1,278
Trans ORF	*RNAseq*	16,273	23,148	12.02	1,891	324.2	1,746
BUSCO		3,651	13,229	10.23	1,630	159.38	1,257
MAKER		27,222	18,042	9.62	1,526	299.41	1,759
HiFAP		29,670	15,398	8.81	1,575	267	1,669

Table 4. Gene prediction of assembly for *O. evolans* genome.

Germany) under cryogenic conditions, following the manufacturer's instructions. For RNA-sequencing-based prediction, transcriptomic data derived from 9 pooled tissue types were aligned with assembled genome using HISAT2 (v2.1.0)[42], and predicted transcript structures were detected using StringTie (v1.3.5)[43]. All gene models were merged to eliminate redundancy using MAKER2 (v2.31.10)[44] and HiFAP (https://www.onemore-tech.com/) with default settings. Finally, we identified 29,670 genes, with an average CDS lengths of 1575 bp, average exon and intron lengths were 267 bp and 1669 bp, respectively (Table 4).

We annotated protein-coding genes through sequence homology searches using diamond (v2.0.1424)[45] with an e-value cut-off of 1e-5. The analysis revealed high annotation coverage across multiple databases: KEGG[46] (90.51%), SwissProt[47] (77.73%), TrEMBL (http://www.uniprot.org) (92.27%), KOG[48] (72.07%), and NR[49] (92.64%) (Fig. 3B). Notably, comprehensive functional annotations were achieved for all predicted genes in *O. evolans*, with 26,885 genes successfully annotated (92.9% coverage) (Table 5).

	Gene		mRNA	
	Number	Percent (%)	Number	Percent (%)
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
Total	26,885	90.61	39,453	92.90

Table 5. The Functional annotation of *O. evolans* genome.

Type	Copy	Average length (bp)	Total length (bp)	% of genome
miRNA	1,696	88	149,898	0.016902
tRNA	8,757	75	660,322	0.074455
rRNA	5,835	121	708,439	0.079880
18S	12	1,542	18,508	0.002087
28S	0	0	0	0.000000
5.8S	12	150	1,799	0.000203
5S	5,811	118	688,132	0.077590
snRNA	1,537	149	228,371	0.025750
CD-box	227	147	33,279	0.003752
HACA-box	73	152	11,068	0.001248
splicing	1,227	148	181,647	0.020482
scaRNA	10	238	2,377	0.000268

Table 6. Statistics of non-coding RNA annotation for *O. evolans* genome.

To complete the non-coding RNA prediction and annotation, we used INFERNAL based on the rfam (v14.8)[50] and miRBase[51] to predict ribosomal RNA (rRNA), microRNA (miRNA) and small nuclear RNA (snRNA) genes. Additionally, tRNAscan-SE (v1.3.1) software[52] was applied to identify Transfer RNA (tRNA) genes. In total, we identified four types of noncoding RNAs in the *O. evolans* genome: 1,696 miRNAs, 8,757 tRNAs, 5,835 rRNAs, and 1,537 snRNAs (Table 6).

Data Records

The sequencing dataset and genome assembly of *O. evolans* have been deposited in the NCBI SRA database under project number PRJNA1282795[53], including: Hi-C data (SRX29433841)[54]; DNBSEQ-T7 genome sequencing data (SRX29433842)[55]; PacBio Revio genome sequencing data (SRX29433843)[56]; oxford_nanopore genome sequencing data (SRX29433844)[57]; transcriptomic data (SRX29433845)[58]. The complete genome assembly and its annotation have been deposited in Figshare (https://doi.org/10.6084/m9.figshare.29399834)[59]. The sequencing data are available in the European Nucleotide Archive (ENA) under accession number PRJEB100826. (https://www.ebi.ac.uk/ena/browser/view/PRJEB100826)[60].

Data Overview

The protein sequences of 10 species (*O. evolans*; *O. bidens*; *Z. platypus*; *Danio rerio*; *Hypophthalmichthys molitrix*; *Anabarilius grahami*[61]; *Cyprinus carpio*[62]; *Labeo rohita*[63]; *Sinocyclocheilus rhinocerous*[64]) were clustered into gene families using OrthoMCL (v2.0.9)[65] with BLASTP alignments (e-value ≤ 1e-5) (https://blast.ncbi.nlm.nih.gov/Blast.cgi). Our analysis revealed 24,591 gene families, 1,074 of which were single-copy gene families. The relationships between species-specific gene families and shared gene families are visualized through clustered petal plots (Fig. 4A). Species phylogeny was reconstructed using RAxML (v8.2.12)[66] under maximum likelihood, with divergence times estimated by MCMCtree (v4.9)[67](clock = 3; model = 0) based on Timetree calibrations. The results revealed that *O. evolans* diverged later than *Z. platypus*, with an estimated divergence time between 7.9 and 19.2 million years ago (Fig. 4B). This study provides the first genome-based evolutionary analysis since their recognition as distinct species. Genome synteny among *O. evolans*, *O. bidens*, and *Z. platypus* was assessed using WGDI (v0.5.6)[68], with visualization performed via JCVI (v1.4.1445)[69]. The results showed that the genome assembly of the *O. evolans* had high collinearity with the *O. bidens* (Fig. 4C).

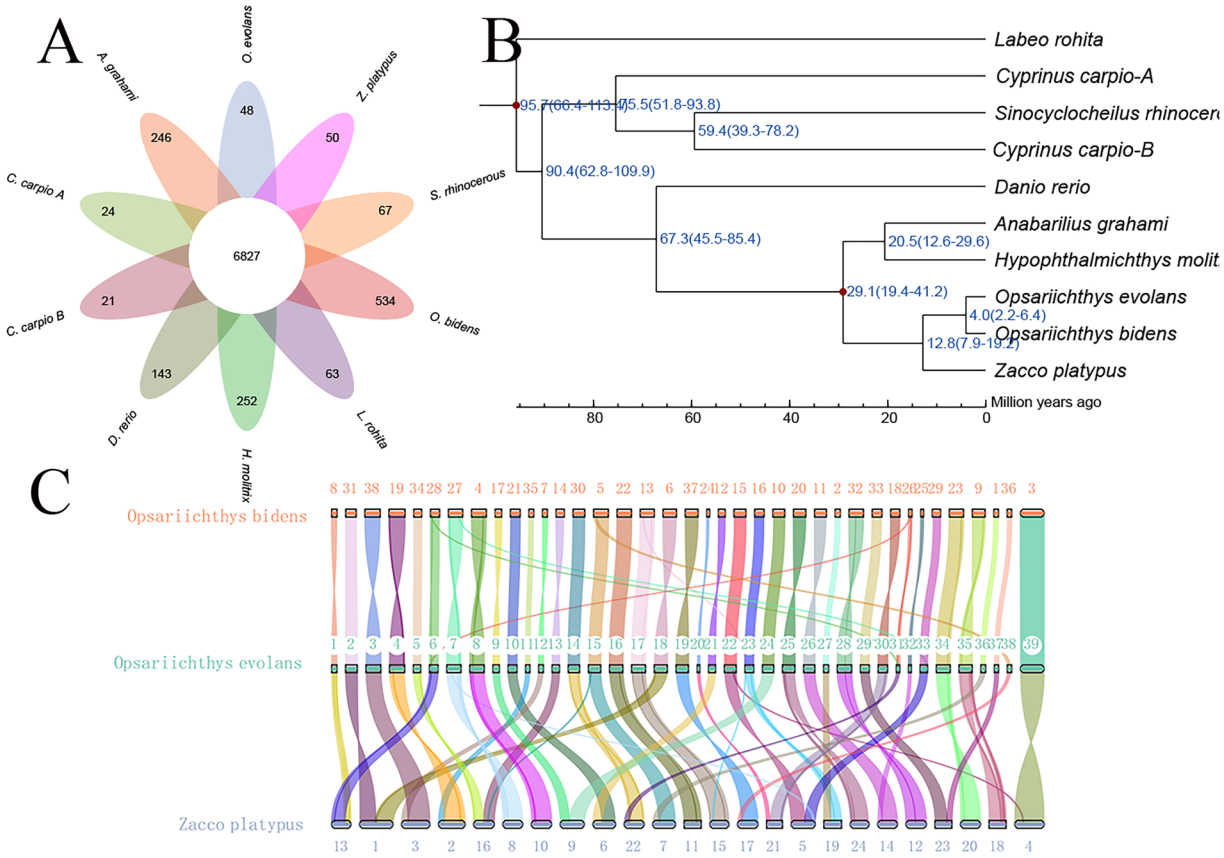

Fig. 4 Gene family clustering petal diagram (**A**), Constructing phylogenetic trees of species based on maximum likelihood method (ML TREE) (**B**) and Chromosome-level genomic synteny between *O. evolans*, *O. bidens*, and *Z. platypus* (**C**).

Software	Busco			
	Assembly		Annotation	
Type	Proteins	Percentage (%)	Proteins	Percentage (%)
Complete BUSCOs	3,616	99.34	3,595	98.76
Complete Single-Copy BUSCOs	3,589	98.6	3,549	97.5
Complete Duplicated BUSCOs	27	0.74	46	1.26
Fragmented BUSCOs	16	0.44	8	0.22
Missing BUSCOs	8	0.22	37	1.02
Total BUSCO groups searched	3,640	100.00	3,640	100.00
Database	actinopterygii_odb10		actinopterygii_odb10	

Table 7. The evaluation of genome assembly and annotation in *O. evolans*.

Technical Validation

We evaluated the accuracy and completeness of the *O. evolans* genome assembly and gene annotation using BUSCO (v5.7.154)[70], identifying 3,616 (99.34%) complete conserved genes and 3,595 (98.76%) highly conserved core proteins, indicating successful genome assembly and annotation (Table 7). After mapping short reads to the genome, variants (SNPs and InDels) were detected using samtools[71], Picard (https://broadinstitute.github.io/picard/), and GATK[72], revealing low homozygous and heterozygous rates, further supporting high assembly accuracy and low heterozygosity (Table S8). To confirm the species specificity of the assembly, the genome sequences were fragmented in 10-kb windows and aligned to the NCBI NT database using BLAST, yielding a median identity of 91.74% with conspecific sequences, suggesting minimal assembly errors (Table S9). Subsequently, short and long reads were mapped to the genome using bwa[73] and minimap2[26], demonstrating high alignment rates and coverage, indicating strong consistency between the assembly and sequencing data (Table S10). Additionally, Merqury (v1.353)[74] was employed for k-mer-based quality assessment, integrating both short and long reads to generate high QV scores (Table S11, Fig. S4). High collinearity was observed between our assembly and the genomes of, *O. bidens, and Z. platypus* (Fig. 4C), further validating assembly

accuracy. Moreover, structural annotations exhibited high similarity to those of closely related species (Fig. S5). Collectively, these results demonstrate the high quality, reliability, and accuracy of the *O. evolans* genome assembly and annotation.

Data availability

The data for this study have been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB100826 (https://www.ebi.ac.uk/ena/browser/view/PRJEB100826)[60].

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: 14 July 2025; Accepted: 8 January 2026;
Published online: 22 January 2026

References

1. Fricke, R., Eschmeyer, W. N. & van der Laan, R. *Eschmeyer's catalog of fishes: genera, species, references*. http://researcharchive.calacademy.org/research/ichthyology/catalog/fishcatmain.asp (2025).
2. Chen, I. S. & C, Y. C. *A photographic guide to the island-water fishes of Taiwan*. Vol. 1 (The Sueichan Press, Keelung, 2005).
3. Chen, I. S., Wu, J. H. & Huang, S. P. The taxonomy and phylogeny of the cyprinid genus *Opsariichthys* Bleeker (Teleostei: Cyprinidae) from Taiwan, with description of a new species. *Environ Biol Fish* **86**, 165–183, https://doi.org/10.1007/s10641-009-9499-y (2009).
4. Cao, L., Shao, W. H., Yi, W. J. & Zhang, E. A review of conservation status of freshwater fish diversity in China. *J Fish Biol* **104**, 345–364, https://doi.org/10.1111/jfb.15606 (2024).
5. Ma, G. C., Watanabe, K., Tsao, H. S. & Yu, H. T. Mitochondrial phylogeny reveals the artificial introduction of the pale chub *Zacco platypus* (Cyprinidae) in Taiwan. *Ichthyol Res* **53**, 323–329, https://doi.org/10.1007/s10228-006-0353-3 (2006).
6. Chen, I. S. & Fang, L. S. *The freshwater and estuarine fishes of Taiwan*. Vol. 1 (Nat Mus Mar Biol Aquar, Pingtung, 1999).
7. Jordan, D. S. & Hubbs, C. L. Record of fishes obtained by David Starr Jordan in Japan, 1922. *Board of Trustees of the Carnegie Institute* **2**, 93–346, https://doi.org/10.5962/p.234844 (1925).
8. Chen, I. S., Huang, S. P., Jang-Liaw, N. H., Shen, C. N. & Wu, J. H. Molecular evidence for genetic differentiation of the *Opsariichthys bidens* complex (Teleostei: Cyprinidae) in southern China around south China Sea and the validity of *Opsariichthys hainanensis*. *Raffles Bull Zool* **19**, 215–223, http://ir.sinica.edu.tw/handle/201000000A/25136 (2008a).
9. Chen, I. S., Wu, J. H. & Hsu, C. H. The taxonomy and molecular phylogeny of Candidia (Teleostei: Cyprinidae) from Taiwan, with description of a new species and comments on a new cyprinid genus. *Raffles Bull Zool* **19**, 203–214, https://doi.org/10.1007/978-90-481-3458-8_25 (2008b).
10. Huynh, T. Q. & Chen, I. S. A new species of cyprinid fish of genus *Opsariichthys* from Ky Cung-Bang Giang river basin, northern Vietnam with notes on the taxonomic status of the genus from northern Vietnam and southern China. *J Mar Sci Technol* **21**, 135–145, https://doi.org/10.6119/JMST-013-1219-12 (2013).
11. Huang, S. P., Wang, F. Y. & Wang, T. Y. Molecular phylogeny of the *Opsariichthys* group (Teleostei: Cypriniformes) based On complete mitochondrial genomes. *Zool Stud* **56**, e40, https://doi.org/10.6620/ZS.2017.56-40 (2017).
12. Wang, X., Liu, F., Yu, D. & Liu, H. Mitochondrial divergence suggests unexpected high species diversity in the opsariichthine fishes (Teleostei: Cyprinidae) and the revalidation of *Opsariichthys macrolepis*. *Ecol Evol* **9**, 2664–2677, https://doi.org/10.1002/ece3.4933 (2019).
13. Jordan, D. S. & Evermann, B. W. Note on a collection of fishes from the island of Formosa. *Proc US Nat Mus* **25**, 315–368, https://doi.org/10.5479/si.00963801.25-1289.315 (1902).
14. Chen, Y. Y. A revision of Opsariichthine cyprinid fishes. *Oceanologia et Limnologia Sinica* **3**, 293–299 (1982).
15. Oshima, M. Contribution to the study of the freshwater fishes of the island of Formosa. *Ann Carneg Mus* **13**, 169–328, https://doi.org/10.5962/p.34608 (1919).
16. Tang, K. L. *et al*. Systematics of the subfamily Danioninae (Teleostei: Cypriniformes: Cyprinidae). *Mol Phylogenet Evol* **57**, 189–214, https://doi.org/10.1016/j.ympev.2010.05.021 (2010).
17. Zhang, Y., Zhou, J. J. & Yang, J. Q. A new species of genus *Zacco* from southern China (Cypriniformes: Cyprinidae). *J Shanghai Fish Univ Uni* **32**, 544–552, https://doi.org/10.12024/jsou.20220703918 (2023).
18. Gao, J. X., Yu, D. Liu. & H. Z. Genetic diversity and phylogeography of opsariichthys macrolepis from the upper and middle yangtze river based on *cytochrome b* sequences. *Sichuan J Zool* **40**, 361-373
19. 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).
20. Liu, B. *et al*. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *Genomics* **35**, 62–67, https://doi.org/10.1016/S0925-4005(96)02015-1 (2013).
21. 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).
22. 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).
23. 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).
24. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst* **3**, 95–8, https://doi.org/10.1016/j.cels.2016.07.002 (2016).
25. 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).
26. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100, https://doi.org/10.1093/bioinformatics/bty191 (2018).
27. 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).
28. Walker, B. J. *et al*. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963, https://doi.org/10.1371/journal.pone (2014).
29. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **4**, 1–14, https://doi.org/10.1002/0471250953.bi0410s25 (2009).
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. 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).
32. 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).
33. *CNCB* https://ngdc.cncb.ac.cn/gwh/Assembly/26062/show (2022).
34. NCBI GenBank. *Whole Genome Assembly* https://www.ncbi.nlm.nih.gov/datasets/genome/GCA_034642465.1/ (2023).
35. NCBI GenBank. *Whole Genome Assembly* https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000002035.6/ (2017).
36. *Figshare* https://doi.org/10.6084/m9.figshare.12618884.v1 (2021).
37. Gertz, E. M., Yu, Y. K., Agarwala, R., Schaffer, A. A. & Altschul, S. F. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biol* **4**, 41, https://doi.org/10.1186/1741-7007-4-41 (2006).
38. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res* **14**, 988–995, https://doi.org/10.1101/gr.1865504 (2004).
39. 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).
40. 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).
41. 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).
42. 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).
43. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295, https://doi.org/10.1038/nbt.3122 (2015).
44. 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).
45. 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).
46. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27–30, https://doi.org/10.1093/nar/28.1.27 (2000).
47. UniProt, C. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res* **38**, D142–148, https://doi.org/10.1093/nar/gkp846 (2010).
48. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59, https://doi.org/10.1186/1471-2105-5-59 (2004).
49. Marchler-Bauer, A. *et al*. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res* **39**, D225–229, https://doi.org/10.1093/nar/gkq1189 (2011).
50. 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).
51. 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).
52. 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).
53. *NCBI GenBank* https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1282795 (2025).
54. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRX29433841 (2025).
55. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRX29433842 (2025).
56. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRX29433843 (2025).
57. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRX29433844 (2025).
58. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRX29433845 (2025).
59. Yin, D. H. Telomere-to-telomere gap-free genome assembly of the *Opsariichthys evolans* (Cypriniformes: Cyprinidae). *Figshare* https://doi.org/10.6084/m9.figshare.29399834 (2025).
60. Yin, D. H. Telomere-to-telomere gap-free genome assembly of the *Opsariichthys evolans* (Cypriniformes: Cyprinidae). ENA https://www.ebi.ac.uk/ena/browser/view/PRJEB100826 (2025).
61. NCBI GenBank. *Whole Genome Assembly* https://identifiers.org/ncbi/insdc.gca:GCA_003731715.1 (2018).
62. NCBI GenBank. *Whole Genome Assembly* https://identifiers.org/ncbi/insdc.gca:GCA_018340385 (2021).
63. NCBI GenBank. *Whole Genome Assembly* https://identifiers.org/ncbi/insdc.gca:GCA_022985175 (2022).
64. NCBI GenBank. *Whole Genome Assembly* https://identifiers.org/ncbi/insdc.gca:GCA_001515625 (2015).
65. Li, L., Stoeckert, C. J. Jr. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**, 2178–89, https://doi.org/10.1101/gr.1224503 (2003).
66. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313, https://doi.org/10.1093/bioinformatics/btu033 (2014).
67. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586–1591, https://doi.org/10.1093/molbev/msm088 (2007).
68. Sun, P. *et al*. WGDI: A user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. *Mol Plant* **15**, 1841–1851, https://doi.org/10.1016/j.molp.2022.10.018 (2022).
69. Tang, H. *et al*. Synteny and collinearity in plant genomes. *Science* **320**, 486–488, https://doi.org/10.1126/science.1153917 (2008).
70. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **19**, 3210–3212, https://doi.org/10.1093/bioinformatics/btv351 (2015).
71. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079, https://doi.org/10.1093/bioinformatics/btp352 (2009).
72. Brouard, J. S., Schenkel, F., Marete, A. & Bissonnette, N. The GATK joint genotyping workflow is appropriate for calling variants in RNA-seq experiments. *J Anim Sci Biotechnol* **10**, 44, https://doi.org/10.1186/s40104-019-0359-0 (2019).
73. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760, https://doi.org/10.1093/bioinformatics/btp324 (2009).
74. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**, 245, https://doi.org/10.1186/s13059-020-02134-9 (2020).

Acknowledgements

Financial support for this study was provided by the Central Public-interest Scientific Institution Basal Research Fund, CAFS (NO.2023TD11) and the Monitoring of Aquatic Resources in Key Waters of Anhui Province Project (ZF2022-18-0399). We also extend our gratitude to Wuhan Onemore-tech Co., Ltd. for their technical assistance in genome sequencing and bioinformatics analysis.

Author contributions

K.L. conceptualized research plan and obtained the funding project. P.W. and X.Y.W. completed the acquisition, analysis and interpretation of data. P.W. wrote the manuscript. K.L., D.H.Y. and L.J. revised the paper. M.J. conducted animal experiments and prepared biological samples. 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/s41597-026-06588-7.

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