



OPEN

DATA DESCRIPTOR

A chromosomal-level genome assembly of *Phoxinus grumi* (Cypriniformes: Leuciscidae)

Jia Wang^{1,2}✉, Hongxiong Chang^{1,2}, Ping Yang¹, Xin Wang¹, Xinyang Li¹, Yuqing He¹, Minghui Gao¹ & Wei Guo¹✉

The Turpan minnow (*Phoxinus grumi*) is a small endemic fish species inhabiting the extreme environment of the Turpan Basin in Xinjiang, China, holding significant value for evolutionary and conservation biology research. However, the absence of a high-quality reference genome has severely constrained studies on its adaptive evolution and conservation genetics, in stark contrast to the available chromosome-level genomes of its congeners, such as *Phoxinus phoxinus*. A total of 240.38 Gb of sequencing data was generated in this study, comprising 44.12 Gb (53.35 ×) of PacBio HiFi reads, 50.36 Gb (60.90 ×) of Illumina reads, 120.59 Gb (133.95 ×) of Hi-C data and 25.31 Gb of RNA sequencing data, which enabled the successful assembly of a chromosome-level genome for *P. grumi*. The assembled genome has a total size of 900.41 Mb, with 97.58% of the sequences anchored onto 25 chromosomes. The contig N50 and scaffold N50 reached 17.52 Mb and 34.99 Mb, respectively. BUSCO assessment indicated a genome completeness of 98.1%. We predicted a total of 24,224 protein-coding genes, of which 90.8% were functionally annotated. This high-quality reference genome will serve as a key genetic resource for in-depth exploration of the environmental adaptation mechanisms and species conservation of *P. grumi*.

Background & Summary

The genus *Phoxinus* comprises a diverse assemblage of leuciscines widely distributed across Eurasia and North America, as inferred from traditional morphological taxonomic studies¹. Members of this genus occupy diverse freshwater habitats and have long been recognized for their considerable taxonomic complexity². However, advances in molecular systematics, particularly phylogenetic analysis based on multi-locus and whole mitochondrial genome datasets, have demonstrated that the traditionally circumscribed *Phoxinus* lacks monophyly, which precludes its recognition as a valid Holarctic clade^{3,4}. Accordingly, its systematic position has undergone substantial revisions over the past decade. Notably, the American fishes formerly included in *Phoxinus* (e.g., “*P. erythrogaster*”) have been reclassified into the genus *Chrosomus*⁵. Concurrently, the taxonomic status of East Asian representatives of *Phoxinus* has been reassessed. For example, the species previously referred to as “*Phoxinus lagowskii*” is now formally annotated as “*Rhynchocypris lagowskii*” in public sequence repositories such as GenBank^{6,7}. This nomenclatural adjustment reflects the broader trend of refining leuciscine systematics under the framework of molecular phylogenetics. Collectively, these findings highlight that *Phoxinus* has a remarkably intricate phylogenetic history and likely harbors substantial cryptic species diversity that remains to be fully characterized.

The Turpan minnow (*P. grumi*) belongs to the genus *Phoxinus* in the subfamily Phoxininae, family Leuciscidae. It is distributed in the Turpan Basin of Xinjiang, China, and is generally considered to be a species endemic to China⁸. It exhibits a unique ecological niche differentiation among fishes of the genus *Phoxinus*. However, there is still controversy regarding the genus-level classification of this species within a modern molecular systematics framework. Therefore, the phylogenetic analysis based on the mitochondrial cytochrome b (*cytb*) gene conducted in this study reveals the complex evolutionary pattern within Eurasian leuciscines (Fig. 1). The results show that *P. grumi* clustered with *Phoxinus steindachneri*, forming a sister clade with moderate bootstrap support, and was nested within the Eurasian *Phoxinus* clade. Notably, species currently

¹Xinjiang Key Laboratory for Ecological Adaptation and Evolution of Extreme Environment Organisms, College of Life Sciences, Xinjiang Agricultural University, Urumqi, 830052, China. ²These authors contributed equally: Jia Wang, Hongxiong Chang. ✉e-mail: wangjia365@xjau.edu.cn; guowei612@xjau.edu.cn

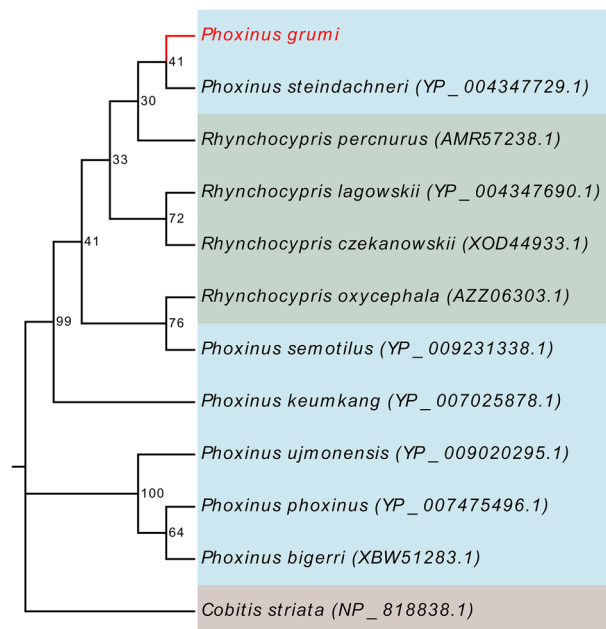


Fig. 1 Phylogenetic tree inferred from mitochondrial cytochrome b (*cytb*) gene sequences, showing the evolutionary relationships of *Phoxinus grumi* and closely related leuciscine fishes. The tree was constructed using the Maximum likelihood (ML) and Neighbor-Joining (NJ) methods, with bootstrap support values (≥ 0.90) labeled on the corresponding branches. Each species is followed by its GenBank accession number in parentheses. *Cobitis striata* was used as the outgroup to root the tree.

assigned to the genus *Rhynchocypris*, including *R. percunurus*, *R. lagowskii*, *R. czezanowski*, and *R. oxycephala*, did not form a monophyletic group distinct from *Phoxinus*, but were interspersed among *Phoxinus* lineages. This topology indicates a lack of clear phylogenetic separation between these two nominal genera based on mitochondrial data, consistent with previous molecular phylogenetic studies suggesting taxonomic complexity and ongoing revisions within Eurasian leuciscine fishes. These results further highlight the necessity of incorporating genome-scale data to resolve the phylogenetic relationships and taxonomic boundaries among East Asian *Phoxinus*–*Rhynchocypris* lineages.

In addition, its habitat is characterized by typical inland arid zone features: high temperatures, intense evaporation, high salinity-alkalinity, and strong ultraviolet radiation. The Turpan Basin constitutes a closed hydrological unit with a highly vulnerable aquatic ecosystem, which has recently faced increasing pressure from both climate change and human activities⁹. Excessive water extraction for agricultural irrigation, declining water tables due to groundwater over-exploitation, and habitat fragmentation caused by hydraulic engineering projects^{10,11} directly threaten the survival of this species. Furthermore, the introduction of non-native fish species poses additional risks, potentially disrupting local ecological balance through niche competition and predation. Although the species is listed as a Class II protected animal in Xinjiang, significant knowledge gaps remain regarding its population dynamics, genetic background, and adaptation mechanisms, highlighting the urgency for a high-quality reference genome. Meanwhile, we also note that existing studies have conducted genome analyses of the Eurasian minnow (*Phoxinus phoxinus*) and some closely related groups^{12,13}. However, the high-quality genome data for *P. grumi* has not been reported to date. This data gap limits the ability to clarify phylogenetic positions and evolutionary relationships with Eurasian closely related taxa, population genetic structure, and environmental adaptation mechanisms at the molecular level. Given that the ecosystem in which this species lives is facing the dual pressures of climate change and human activities, constructing its chromosome-level genome is not only fundamental to clarifying the systematic evolution of leuciscine fishes, but also provides crucial molecular genetic evidence for the protection and management of this unique and endangered species. Future studies should integrate ecological surveys, comparative genomics, and environmental genomics to fully elucidate the evolutionary history and adaptation mechanisms of this endemic species, thereby providing a theoretical foundation and technical support for conserving aquatic organisms in arid regions.

In this study, we successfully constructed a chromosome-level genome assembly for *P. grumi* by integrating approximately 44.12 Gb of PacBio HiFi reads, 50.36 Gb of Illumina reads, 120.59 Gb of Hi-C sequencing data and 25.31 Gb of RNA sequencing data. The final assembled genome size is 900.41 Mb, with 97.58% of the sequence anchored to 25 chromosomes. The completion of this high-quality reference genome provides crucial data for clarifying the genetic basis of adaptive evolution in extreme environments and establishes a solid genomic foundation for the conservation, population recovery, and management of this endangered endemic fish species.

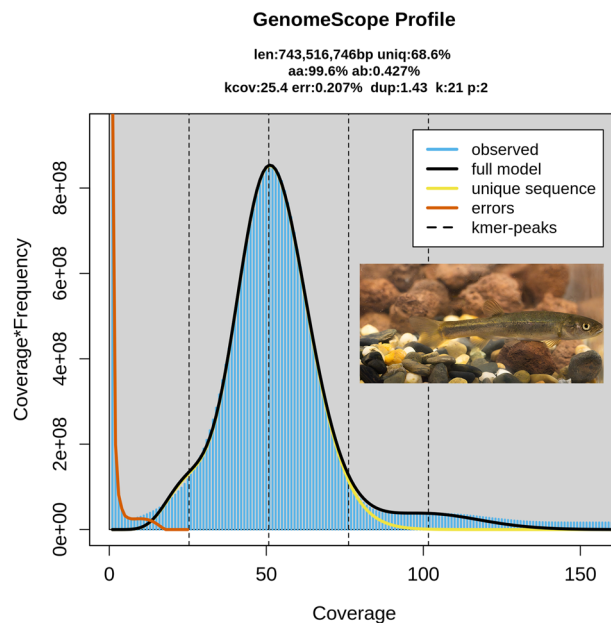


Fig. 2 K-mer distribution profile of *P. grumi* genome. The 21-mer coverage frequency plot, generated with GenomeScope, shows observed k-mers (blue), unique sequences (orange), heterozygous k-mers (yellow), and homozygous k-mers (gray). The analysis estimated a genome size of 743.52 Mb, with 68.6% unique sequences and ~31.4% repetitive sequences. The inset shows a field-collected *P. grumi* individual in a laboratory aquarium.

Methods

Sample collection, DNA and RNA extraction. All experiments were approved by Animal Welfare and Ethics Committee of Xinjiang Agricultural University. A mature female *P. grumi* (Fig. 2) with a body weight of 14.50 g and a body length of 11.15 cm was collected from Dacao Lake in Turpan (Xinjiang, China). Tissue was extracted from the individual, rapidly frozen in liquid nitrogen for one hour, and subsequently stored at -80°C for further processing. Genomic DNA was extracted from fresh muscle tissue using a modified SDS method for whole-genome sequencing, which included Illumina short-read sequencing, PacBio HiFi long-read sequencing, and Hi-C sequencing for genome assembly. Meanwhile, in order to obtain comprehensive transcript sequences to assist in gene annotation, total RNA was extracted from different tissues, including muscle, caudal fin, liver, kidney, heart, intestine, ovary and eye, which were mixed in equal amounts to construct a strain-specific transcriptome sequencing library. Extraction was carried out in accordance with the manufacturer's instructions using TRNzol Unified Reagent (TIANGEN BIOTECH CO., LTD, Beijing, China). The integrity and potential contamination of DNA and RNA samples were assessed through electrophoresis on 1% agarose gels. Purity evaluation was conducted by measuring the OD 260/280 ratio using a Nanodrop spectrophotometer (NanoDrop, USA) and the Qubit DNA Assay Kit on a Qubit 3.0 Fluorometer (Invitrogen, USA) following the manufacturer's instructions.

Library construction and sequencing. For short-read sequencing, genomic DNA was fragmented to 350 bp by sonication, subjected to end repair, and then used for library construction with the NEB Next Ultra™ DNA Library Prep Kit for Illumina (NEB, USA). Finally, sequencing was performed on the Illumina NovaSeq X Plus platform (Illumina, USA) to generate 150 bp paired-end reads. For long-read sequencing, 20 kb HiFi SMRTbell libraries were constructed using the SMRTbell Prep Kit 3.0 (PacBio) and subsequently subjected to PacBio HiFi sequencing on the PacBio Revio platform (Pacific Biosciences, USA) in circular consensus sequencing (CCS) mode¹⁴. For Hi-C library construction¹⁵, muscle tissues of *P. grumi* were first cross-linked with the cell cross-linking agent paraformaldehyde and then digested using the restriction enzyme MboI. The DNA ends were labeled with biotin-14-dCTP, followed by blunt-end ligation of cross-linked fragments. Proteins at the ligation sites were digested to reverse the protein-DNA cross-linking, after which genomic DNA was extracted and randomly sheared using a Covaris sonicator. The sheared DNA was end-repaired using a mixed enzyme system consisting of DNA polymerase I, T4 polynucleotide kinase, and Klenow fragment. Finally, the Hi-C sequencing library was amplified by PCR (12–14 cycles) and sequenced on the Illumina PE150 platform. Library construction and sequencing were performed by Novogene (Beijing, China). The sequencing efforts in this study yielded a total of 215.07 Gb of high-quality data, comprising 44.12 Gb ($53.35\times$ coverage) of PacBio HiFi reads, 50.36 Gb ($60.90\times$ coverage) of Illumina short reads, and 120.59 Gb ($133.95\times$ coverage) of Hi-C data (Table 1). In addition, to facilitate genome annotation, RNA-seq libraries were constructed by mixing RNA samples from different tissues in equal volumes using the Fast RNA-seq Lib Prep Kit V2 (ABclonal Biotechnology Co., Ltd., Wuhan, China) according to the manufacturer's instructions. All libraries were constructed according to standard procedures and subjected to paired-end sequencing on the Illumina NovaSeq X Plus (Illumina, USA) sequencing platform. This resulted in a total of 25.31 Gb of transcriptome data, which was used to support genome annotation (Table 1).

Libraries	Insert size	Clean data (Gb)	Sequence coverage (X)
Illumina reads	350 bp	50.36	60.90
PacBio reads	15 kb	44.12	53.35
Hi-C	350 bp	120.59	133.95
Internal organs RNAseq	220 bp	9.92	—
Muscle RNAseq	220 bp	7.45	—
Caudal fin and eyes RNAseq	220 bp	7.94	—

Table 1. Sequencing Data Statistics for Genome Assembly.

Type	Contigs	Scaffolds
Total (bp)	900,351,248	900,412,448
Max (bp)	40,160,965	52,083,149
N50 (bp)	17,521,639	34,986,944
N90 (bp)	439,850	26,733,659
N50 number	18	12
N90 number	195	23

Table 2. Genome assembly statistic of contigs and scaffolds.

Chromosome-level Genome survey and assembly. To ensure the quality of data analysis, Illumina raw reads were subjected to quality control using fastp v0.23.1¹⁶ with default parameters. The quality control steps included: (1) removal of adapter-contaminated reads; (2) exclusion of reads containing over 10% undetermined bases (N); and (3) filtering out paired-end reads where low-quality bases (Phred score < 5) constituted more than 20% of either read length. This procedure generated high-quality clean reads for subsequent analyses. Meanwhile, k-mer analysis was performed using Illumina clean reads to estimate genome size, heterozygosity, and repetitive sequence content. First, the frequency of 17-mers was calculated with Jellyfish v2.3.0¹⁷, followed by analysis using GenomeScope v2.0¹⁸. The estimated genome size was 743.52 Mb, with a unique sequence proportion of 68.6% and a repetitive sequence proportion of approximately 31.4% (Fig. 2). The PacBio HiFi long reads were assembled using Hifiasm v0.19.8¹⁹ with default parameters, yielding 812 contigs, with the largest contig measuring 40.16 Mb in size. During the construction of the chromosome-level genome assembly, Hi-C reads were first subjected to quality filtering and then aligned to the assembled genome using Juicer v1.6²⁰. The contigs were subsequently anchored and ordered into chromosomes following the standard pipeline of 3D-DNA v.180922²¹. Finally, the assembly was manually corrected and optimized for chromosomal boundaries and scaffolding errors using Juicebox v1.11.08²². To ensure the accuracy of the assembly results, BLAST was used to align sequences against the NCBI nucleotide database for contaminant screening and removal of exogenous sequences.

The final chromosome-level assembly of *P. grumi* spans 900.41 Mb, consisting of 222 scaffolds with a scaffold N50 of 34.99 Mb and a contig N50 of 17.52 Mb (Table 2). The assembled genome size (900.41 Mb) is larger than the k-mer based estimate (743.52 Mb), which is often observed in genomes with high repeat content and heterozygosity, and indicates a more complete assembly of repetitive regions. Notably, 97.58% of the assembled sequences (878.65 Mb) were successfully anchored to 25 chromosomes, and the Hi-C heatmap clearly illustrates the interaction patterns among these 25 chromosomes (Fig. 3A). Additionally, Circos²³ was employed to visualize the 25 chromosomes, overall transposable element density, gene density, GC content, and chromosomal synteny (Fig. 3B). The lengths of the individual chromosomes ranged from 25.45 Mb to 52.08 Mb (Table 3). Furthermore, the completeness of the genome assembly was assessed using BUSCO v5.4.3²⁴, revealing 3,574 (98.1%) complete and 40 (1.1%) fragmented BUSCOs (Table 4). To evaluate assembly accuracy, short-insert library reads were mapped to the assembled genome via BWA v0.7.171²⁵, resulting in an alignment rate of approximately 99.24% (Table 5). Sequence accuracy was quantified using Merqury v1.3²⁶ with Illumina sequencing data, yielding a quality value (QV) of 40.97, corresponding to 99.99% accuracy. Collectively, these metrics demonstrate that the genome assembly exhibits high consistency, completeness, and accuracy, thereby confirming its high-quality status.

Analysis of repetitive sequence annotation. We implemented a composite strategy combining homology-based alignment and *de novo* prediction to identify genome-wide repetitive elements for repetitive sequence annotation. For homology-based prediction, RepeatMasker v4.1.2²⁷ was employed with default parameters to extract repetitive regions using the Repbase database. For *de novo* prediction, RepeatModeler v2.0.3²⁸ was utilized to construct a *de novo* repeat library under default parameters. The Repbase and *de novo* repeat libraries were subsequently integrated, and RepeatMasker was applied to annotate repetitive elements in the *P. grumi* genome. The results showed that the *P. grumi* genome contains 49.75% repetitive sequences (Table 6).

Gene Structure and functional annotation. Genome structural annotation was performed using a combined strategy encompassing ab initio prediction, homology-based prediction, and RNA-Seq-assisted prediction,

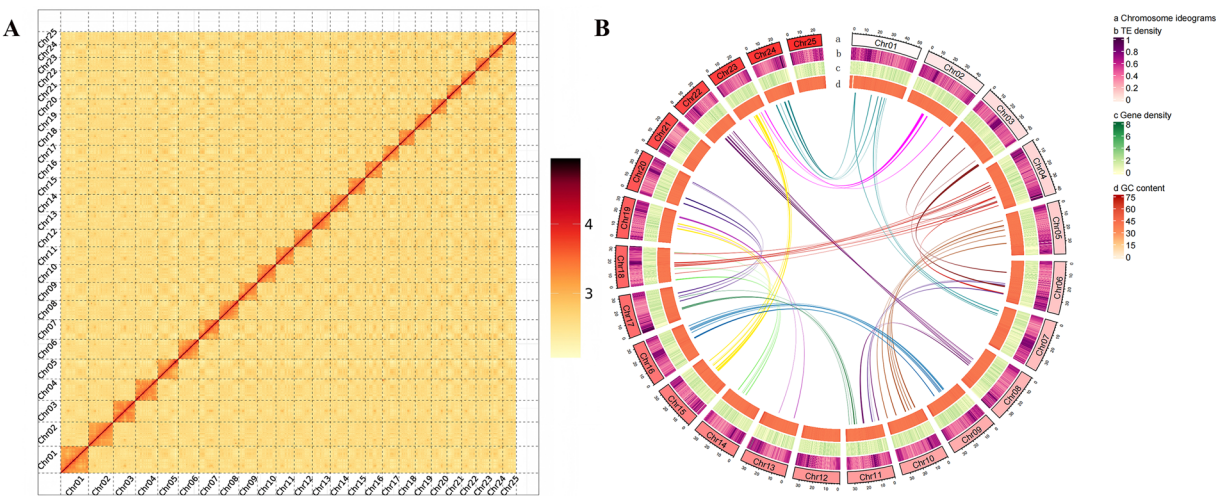


Fig. 3 Genome-wide chromosomal heatmap of Hi-C (A) and circos plot of genome (B). The rings from outer to inner depict the following: (a) chromosome length of the genome, (b) total transposable elements (TE) density, (c) gene density, (d) GC content, (e) chromosomal synteny.

Sequences ID	Cluster Number	Sequences Length(bp)
Chr01	25	52,080,749
Chr02	29	47,744,929
Chr03	27	42,796,627
Chr04	21	42,648,283
Chr05	26	39,728,595
Chr06	19	39,381,637
Chr07	29	39,025,000
Chr08	28	38,050,252
Chr09	22	36,137,236
Chr10	24	35,775,943
Chr11	29	35,267,778
Chr12	29	34,984,144
Chr13	36	34,796,975
Chr14	17	34,645,868
Chr15	26	33,238,693
Chr16	38	32,851,566
Chr17	23	32,151,497
Chr18	22	31,621,491
Chr19	27	31,049,451
Chr20	30	30,391,843
Chr21	20	28,433,924
Chr22	27	28,102,351
Chr23	23	26,731,459
Chr24	26	25,506,610
Chr25	14	25,447,021
Total	637	878,651,122
Mount rate	97.58%	

Table 3. Statistics on the cluster number and length of chromosome of Hi-C assisted assembly.

with the aim of annotating gene models. For ab initio-based gene prediction, Augustus v3.5^{29–31} and SNAP v2013.11.2928³² were integrated into our automated gene prediction pipeline.

For homology-based prediction, protein sequences were downloaded from NCBI, with the utilized protein sequences derived from four reference species: *Cyprinus carpio*, *Pimephales promelas*, *Danio rerio*, *Phoxinus phoxinus*. Protein sequences were aligned to the genome using TblastN v2.2.26³³ with an E-value of 1e-5. Subsequently, the matched proteins were aligned to their homologous genomic sequences, and GeneWise v2.4.1³⁴ was employed to generate accurate spliced alignments, and statistical analysis of gene structure

Type	Genome Assembly		Protein-coding gene models	
	Number	Rate (%)	Number	Rate (%)
Complete BUSCOs (C)	3574	98.1	3303	90.8
Complete and single-copy BUSCOs (S)	3518	96.6	3206	88.1
Complete and duplicated BUSCOs (D)	56	1.5	97	2.7
Fragmented BUSCOs (F)	40	1.1	136	3.7
Missing BUSCOs (M)	26	0.8	201	5.5

Table 4. BUSCO evaluation of *P. grumi* genome.

Type	Catagory	Percentage
Reads	Mapping rate	99.24%
	Average sequencing depth	42.58
	Coverage	99.80%
Genome	Coverage at least 4X	99.25%
	Coverage at least 10X	97.07%
	Coverage at least 20X	90.99%

Table 5. Statistics of Genomic Read Coverage from Sequencing Datasets.

Type	Number	Length(bp)	Percentage(%)
SINE	13,227	6,811,737	0.76
LINE	189,502	69,132,842	7.68
LTR	353,754	108,019,463	12.00
DNA	1,187,167	213,187,585	23.68
Unknown	25,132	2,827,591	0.31
Total	—	447,955,691	49.75

Table 6. Statistics of classification results of repetitive sequences.

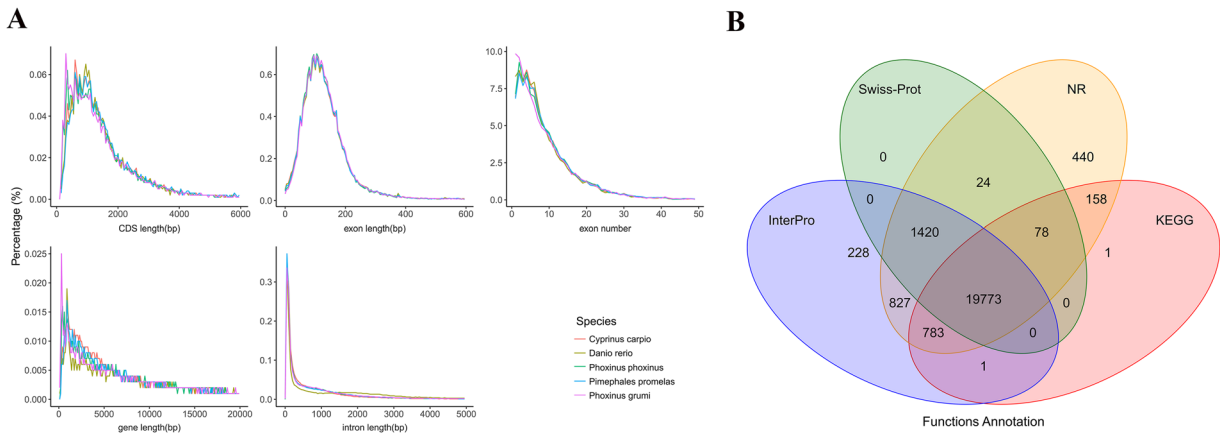


Fig. 4 Structural feature comparisons and functional annotation of *P. grumi* and related species. **(A)** Distribution of structural features, including CDS length, gene length, intron length, exon length, and exon number across *P. grumi* and four other teleost species. **(B)** Venn diagram showing the overlap of gene function annotations across four databases: Swiss-Prot, NR, KEGG and InterPro.

characteristics between *P. grumi* and closely related species was conducted, including statistics on gene length, coding sequence (CDS), intron length, and exon length. These genomic elements were found to be comparable across the examined species (Fig. 4A).

In the RNA-seq-assisted analysis, sequencing reads were initially aligned to the reference genome FASTA file using Hisat2 v2.2.1³⁵ with default parameters to identify exon regions and splice sites. The resulting alignment files were subsequently utilized as input for StringTie v2.2.1³⁶ to perform reference-based transcript assembly under default settings.

Through integration of gene predictions from the above three methods using EVM v1.1.1³⁷ and further manual curation, a total of 24,224 protein-coding genes were identified in the *P. grumi* genome, with an average

Gene structure annotation	Number
Number of protein coding genes	24,224
Average transcript length(bp)	16,241.00
Average CDS length(bp)	1,619.55
Average exons per gene	9.43
Average exon length(bp)	171.83
Average intron length(bp)	1,735.45
Gene function annotation	Number(Percent)
NR	23,503(97.02%)
Swissprot	21,295(87.91%)
KEGG	20,794(85.84%)
InterPro	23,032(95.08%)
Pfam	20,079(82.89%)
GO	19,527(80.61%)
Annotated	23,733(97.97%)
Unannotated	491(2.03%)

Table 7. Gene Structure and functional annotation statistics.

Type	Copy number	Average-length(bp)	Total-length(bp)	% of genome
miRNA	3,105	124.601	386,885	0.043
tRNA	30,999	74.601	2,312,565	0.257
rRNA	14,629	121.989	1,784,574	0.198
18S	96	538.562	51,702	0.006
28S	175	393.440	68,852	0.008
5.8S	23	145.609	3,349	0.000
5S	14,335	115.847	1,660,671	0.184
snRNA	2,578	148.036	381,638	0.042
CD-box	373	161.638	60,291	0.007
HACA-box	93	153.140	14,242	0.002
splicing	2,046	144.686	296,027	0.033
scaRNA	59	181.339	10,699	0.001
Unknown	7	54.143	379	0.000
miRNA	3,105	124.601	386,885	0.043
Total	71,623	2,482.232	7,418,759	0.824

Table 8. Statistical results of non-coding RNA.

transcript length of 16,241.00 bp. The average lengths of CDS, exons and introns were 1,619.55 bp, 171.83 bp, and 1,735.45 bp, respectively (Table 7). For functional annotation, protein sequences were aligned against the Swiss-Prot³⁸ database using BLASTP v2.2.26³⁹ with an E-value of 1e-5, and functional annotations were assigned based on the best hit results. Protein motifs and domains were annotated using InterProScan v5.39-91.0⁴⁰ by aligning against public databases including ProDom^{41,42}, PRINTS⁴³, Pfam⁴⁴, SMART⁴⁵, PANTHER⁴⁶, and PROSITE⁴⁷. The Gene Ontology (GO) identifier for each gene was assigned based on the corresponding InterPro entry. Meanwhile, protein functions were predicted based on the Swissprot database and the screening of matching results in the NR database using DIAMOND v0.8.22⁴⁸. Additionally, we mapped the gene set to the KEGG pathway database. The final functional annotation showed that 23,733 genes of *P. grumi* were functionally annotated, accounting for 97.97% of the total predicted protein-coding genes (Table 7), with 19,773 genes annotated in common (Fig. 4B). The completeness of the annotated gene set was assessed with BUSCO v5.4.6²⁴ against the actinopterygii_odb10 dataset, revealing 90.8% complete BUSCO genes, indicating high completeness of the gene annotation (Table 4).

Annotation of non-coding RNA genes. The identification of tRNA genes was performed using the tRNAscan-SE v1.4⁴⁹. Given the high conservation of rRNA sequences, we selected rRNA sequences from closely related species as references and predicted rRNA sequences using the BLAST tool. The remaining non-coding RNAs (including miRNAs and snRNAs) were identified using Infernal v1.1.4⁵⁰ with default parameters by searching against the Rfam database. Ultimately, we identified 71,623 non-coding RNAs with an average length of 2482.23 bp and a total length of 7,418,759 bp, accounting for approximately 0.82% of the total genome length (Table 8).

Data Records

The raw sequencing reads of *P. grumi* have been deposited in the NCBI Sequence Read Archive (SRA) database under the Bioproject accession number PRJNA1399684 and the BioproSample accession number SAMN54504127. The deposited data includes Illumina sequencing data (SRR3676662⁵² and SRR36766628⁵³), PacBio HiFi sequencing data (SRR36766622⁵⁶, SRR36766623⁵⁷), and RNA sequencing data (SRR36843988⁵⁸, SRR36843989⁵⁹, and SRR36843990⁶⁰). The assembled genome has been deposited at GenBank⁶¹. Meanwhile, the raw sequencing data have been also deposited in the Genome Warehouse at the National Genomics Data Center⁶² (Beijing Institute of Genomics, Chinese Academy of Sciences/China National Center for Bioinformation) with accession number PRJCA050662 (BioProject) and subSAM156170 (BioSample). Moreover, the final genome assembled and annotation files have been deposited in Figshare⁶³.

Technical Validation

The completeness of *P. grumi* genome assembly was assessed using BUSCO v5.4.3²⁴ with the actinopterygii_odb10 dataset. The genome assembly showed a BUSCO completeness of 98.1%, including 96.6% complete single-copy BUSCOs, 1.5% complete duplicated BUSCOs, 1.1% fragmented BUSCOs, and 0.8% missing BUSCOs (Table 4). To evaluate the accuracy of the genome assembly, short-insert library reads were aligned to the assembled genome using BWA v0.7.171²⁵, yielding a mapping rate of approximately 99.24%, with an average sequencing depth of 42.58× and genome coverage of 99.80% (Table 5). Base-level accuracy of the genome assembly was assessed using Merqury v1.3²⁵, which yielded a quality value (QV) of 40.97, corresponding to a sequencing accuracy of 99.99%. In addition, a total of 24,224 protein-coding genes were obtained by integrating *de novo* prediction, homology alignment, and RNA-seq prediction, among which 23,733 genes were successfully functionally annotated. Evaluation of the completeness of the assembled gene set showed that 90.8% of the complete BUSCO genes were annotated in the *P. grumi* gene set, including 88.1% complete single-copy BUSCOs, 2.7% complete duplicated BUSCOs, and 3.7% fragmented BUSCOs. These results confirm the high accuracy and reliability of the annotation. Overall, multiple methods were used to evaluate the assembled genome. The high alignment rate, identification rate of single-copy orthologous genes, and gene count collectively confirm the high quality of the *P. grumi* genome assembly.

Data availability

The raw sequencing data are available in the NCBI databases under Bioproject accession number PRJNA1399684. Additionally, the assembled genome has been deposited in GenBank. Furthermore, all datasets are available under the BioProject accession number PRJCA050662 in the Genome Warehouse (GWH) at the National Genomics Data Center (NGDC). The data are publicly accessible via the following link at <https://ngdc.cnbc.ac.cn/gwh>. Raw reads have been deposited in NGDC (Hi-C: SAMC6098139; Illumina: SAMC6098138; PacBio HiFi: SAMC6098137). The final genome assembled and annotation files have been deposited in Figshare platform via <https://doi.org/10.6084/m9.figshare.30572321.v1>.

Code availability

No custom code or scripts were utilized in this study, all commands and pipelines involved in data processing were executed in accordance with the manuals and protocols provided by the bioinformatic software employed. The specific versions of software packages and corresponding parameters implemented for each analytical step are explicitly detailed in the Methods section to ensure reproducibility.

Received: 21 October 2025; Accepted: 17 March 2026;

Published online: 23 March 2026

References

1. Zardoya, R. & Doadrio, I. Molecular evidence on the evolutionary and biogeographical patterns of European cyprinids. *J Mol Evol.* **49**, 227–237 (1999).
2. Imoto, J. M. *et al.* Phylogeny and biogeography of highly diverged freshwater fish species (Leuciscinae, Cyprinidae, Teleostei) inferred from mitochondrial genome analysis. *Gene.* **514**, 112–124 (2013).
3. Schönhuth, S. *et al.* Phylogenetic relationships and classification of the Holarctic family Leuciscidae (Cypriniformes: Cyprinoidei). *Mol Phylogenet Evol.* **127**, 781–799 (2018).
4. Palandačić, A., Witman, K. & Spikmans, F. Molecular analysis reveals multiple native and alien Phoxinus species (Leuciscidae) in the Netherlands and Belgium. *Biol Invasions.* **24**, 2273–2283 (2022).
5. Page, L. M. *et al.* Common and Scientific Names of Fishes from the United States, Canada, and Mexico (8th ed.). *Fisheries.* **48**, 497–498 (2023).
6. Zhou, Y. *et al.* Telomere-to-telomere genome assembly of *Phoxinus lagowskii*. *Sci Data.* **12**, 1025 (2025).
7. Zheng, H. *et al.* Chromosome-level genome assembly of the *Phoxinus lagowskii*. *Sci Data.* **12**, 1400 (2025).
8. Zhang, C. & Zhao, Y. *Species Diversity and Distribution of Inland Fishes in China.* (Science Press, 2016).
9. Bridle, J. R., Pedro, P. M. & Butlin, R. K. Habitat fragmentation and biodiversity: testing for the evolutionary effects of refugia. *Evolution.* **58**, 1394–1400 (2004).
10. Du, L. *et al.* Hydroclimatic Change in Turpan Basin under Climate Change. *Water.* **15**, 3422 (2025).
11. Di Giulio, M., Holderegger, R. & Tobias, S. Effects of habitat and landscape fragmentation on humans and biodiversity in densely populated landscapes. *J Environ Manage.* **90**, 2959–2968 (2009).
12. Nunn, A. D. *et al.* The genome sequence of the Eurasian minnow, *Phoxinus phoxinus* (Linnaeus, 1758). *Wellcome Open Res.* **9**, 504 (2024).
13. Oriowo, T. O. *et al.* A chromosome-level, haplotype-resolved genome assembly and annotation for the Eurasian minnow (Leuciscidae: *Phoxinus phoxinus*) provide evidence of haplotype diversity. *Gigascience.* **14**, giae116 (2025).
14. Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol.* **37**, 1155–1162 (2019).
15. van Berkum, N. L. *et al.* Hi-C: a method to study the three-dimensional architecture of genomes. *J Vis Exp.* 1869 (2010).

16. Chen, S. *et al.* fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. **34**, i884–i890 (2018).
17. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*. **27**, 764–770 (2011).
18. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun*. **11**, 1432 (2020).
19. Cheng, H. *et al.* Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*. **18**, 170–175 (2021).
20. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst*. **3**, 95–98 (2016).
21. Dudchenko, O. *et al.* De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science*. **356**, 92–95 (2017).
22. Durand, N. C. *et al.* Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst*. **3**, 99–101 (2016).
23. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res*. **19**, 1639–1645 (2009).
24. Simão, F. A. *et al.* BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. **31**, 3210–3212 (2015).
25. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*. **25**, 1754–1760 (2009).
26. Rhie, A. *et al.* Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol*. **21**, 245 (2020).
27. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics*. **25**, 4–10 (2009).
28. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA*. **117**, 9451–9457 (2020).
29. Stanke, M. *et al.* AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res*. **32**, W309–W312 (2004).
30. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Res*. **33**, W465–W467.
31. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res*. **34**, W435–W439.
32. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics*. **5**, 59 (2004).
33. Gertz, E. M. *et al.* Composition-based statistics and translated nucleotide searches:improving the TBLASTN module of BLAST. *BMC Biol*. **4**, 41 (2006).
34. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res*. **14**, 988–995 (2004).
35. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: A fast spliced aligner with low memory requirements. *Nat. Methods*. **12**, 357–360 (2015).
36. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
37. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol*. **9**, 1–22 (2008).
38. Boeckmann, B. *et al.* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*. **31**, 365–370 (2003).
39. McGinnis, S. & Madden, T. L. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Research*. **32**, W20–W25 (2004).
40. Zdobnov, E. M. & Apweiler, R. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).
41. Bru, C. *et al.* The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res*. **33**, D212–D215 (2005).
42. Corpet, F., Gouzy, J. & Kahn, D. The ProDom database of protein domain families. *Nucleic Acids Res*. **26**, 323–326 (1998).
43. Attwood, T. K. The PRINTS database: a resource for identification of protein families. *Brief Bioinform*. **3**, 252–263 (2002).
44. Mistry, J. *et al.* Pfam: Te protein families database in 2021. *Nucleic Acids Res*. **49**, D412–D419 (2021).
45. Letunic, I. & Bork, P. 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res*. **46**, D493–D496 (2018).
46. Mi, H. *et al.* The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res*. **33**, D284–D288 (2005).
47. Hulo, N. *et al.* The PROSITE database. *Nucleic Acids Res*. **34**, D227–D230 (2006).
48. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. **12**, 59–60 (2015).
49. Chan, P. P. *et al.* tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res*. **49**, 9077–9096 (2021).
50. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*. **29**, 2933–2935 (2013).
51. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR36766626> (2026).
52. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR36766627> (2026).
53. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR36766628> (2026).
54. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR36766624> (2026).
55. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR36766625> (2026).
56. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR36766622> (2026).
57. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR36766623> (2026).
58. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR36843988> (2026).
59. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR36843989> (2026).
60. NCBI Sequence Read Archive <https://identifiers.org/ncbi/insdc.sra:SRR36843990> (2026).
61. NCBI GenBank https://identifiers.org/ncbi/insdc.gca:GCA_055048795.1 (2026).
62. Members, C.-N., Partners Database resources of the National Genomics Data Center. China National Center for Bioinformation in 2024. *Nucleic Acids Res*. **52**, D18–D32 (2024).
63. Chang, H. Genome Annotation Dataset of *Phoxinus grumi*. *Figshare*. <https://doi.org/10.6084/m9.figshare.30572321.v1> (2025).

Acknowledgements

This research was supported by the Third Xinjiang Scientific Expedition Program (No. 2022xjkk1505), the Xinjiang Key Laboratory for Ecological Adaptation and Evolution of Extreme Environment Organisms (No. KFKT2402), the China Postdoctoral Science Foundation (No. 339494), and the Xinjiang Uygur Autonomous Region Tianchi Talent Introduction Program.

Author contributions

J.W. and H.C. conducted the bioinformatic analyses including genome assembly and gene annotation, and drafted the manuscript. P.Y. processed and refined the images and contributed to data analysis. W.G., X.W., X.L., Y.H. and M.G. collected the samples and performed the animal experiments. J.W. and W.G. revised and edited the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

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

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



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

© The Author(s) 2026