scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of Amur Minnow (*Rhynchocypris lagowskii*)

Xiaohui Bai[1], Hanlin Li[1], Yaoxin Wang[1], Xinhui Zhang[2], Jingjing Li[1], Xin Luo[1], Zhang Luo[3], Lin Ma[1], Shuang Hao[1], Chen Wu[1] & Wei Zhang[1]✉

Amur minnow (*Rhynchocypris lagowskii*), an economically important freshwater fish, inhabits tributaries of the Heilongjiang River to Yangtze River in China. Thus far, no genome data of this species have been reported. In this study, we constructed chromosome-level genome assembly for *R. lagowskii* by integration of MGI, PacBio HiFi and Hi-C sequencing technologies. The assembled genome was 1,041.22 Mb in size, with the scaffold N50 of 32.63 Mb, and 99.04% of the sequences (1,007.75 Mb) were anchored to 25 chromosomes. A total of 32,425 protein-coding genes were predicted in the final assembly, of which 91.09% were functionally annotated. The BUSCO score of genome assembly and protein annotation were 96.32% and 97.53%, respectively, indicating a high quality of genome assembly. The high-quality genome serves as a critical resource for in-depth investigations on comparative genomics, population genetics, molecular breeding, and functional exploration of this economically important fish.

Background & Summary

Belonging to the order Cypriniformes and the family Cyprinidae, Amur minnow (*Rhynchocypris lagowskii*) is among the most economically important small-sized freshwater fish species. It is distributed in waters such as tributaries in the middle and upper reaches of the Heilongjiang River, Liaohe River, Yellow River, Huaihe River, Haihe River, and Yangtze River in China[1]. The flesh of *R. lagowskii* is tender and delicious, with high nutritional and economic value[2,3]. It has been selected as one of the top ten characteristic aquatic germplasm resources by the Ministry of Agriculture and Rural Affairs of China. Furthermore, it serves as the principal bait organism for the rare and endemic cold-water fish species in China, *Brachymystax tsinlingensis*[4]. In recent years, the natural population of *R. lagowskii* has been progressively declining due to environmental pollution and overfishing, leading to a market supply shortage[3].

R. lagowskii exhibits a broad geographical distribution, ranging from the low-latitude Yangtze River system to the high-latitude cold Heilongjiang River system. Comparative transcriptome analyses have revealed that genes in *R. lagowskii*, which are influenced by environmental factors such as temperature, dissolved oxygen, and pH, have undergone positive selection[5]. This process has facilitated the development of adaptability to varying temperatures throughout evolution, thereby endowing this species with a wide temperature tolerance range[6]. The artificial breeding and aquaculture of *R. lagowskii* have been successfully carried out in Northeast China, achieving good economic benefits[7,8]. However, most of the fry cultured are domesticated and bred from wild parents without systematic selection. Research on *R. lagowskii* primarily concentrates on aspects such as morphological characteristics and classification[5], artificial breeding and aquaculture techniques[7,8], analysis of muscle nutritional components[3], karyotype[9], population genetics[10], and comparative transcriptome research[5]. However, reports on the genome of *R. lagowskii* have been lacking, which seriously hampers research on genetic selection at the molecular level. Therefore, conducting genome sequencing and assembly of *R. lagowskii* is of great significance for the protection of its germplasm resources, breeding of improved varieties, and evolutionary research.

In this study, we successfully generated a chromosome-level genome assembly for the Amur minnow through the integration of PacBio HiFi, and Hi-C sequencing technologies. Subsequently, the completeness and accuracy of the assembled genome were comprehensively assessed. High-quality genome assembly at the

[1]Tianjin Fisheries Research Institute, Tianjin, 300221, China. [2]Laboratory of Aquatic Genomics, College of Life Sciences and Oceanography, Shenzhen University, Shenzhen, 518057, China. [3]Tianjin Key Lab of Aqua-Ecology and Aquaculture, College of Fisheries, Tianjin Agricultural University, Tianjin, 300384, China. ✉e-mail: zhang-wei-316@163.com

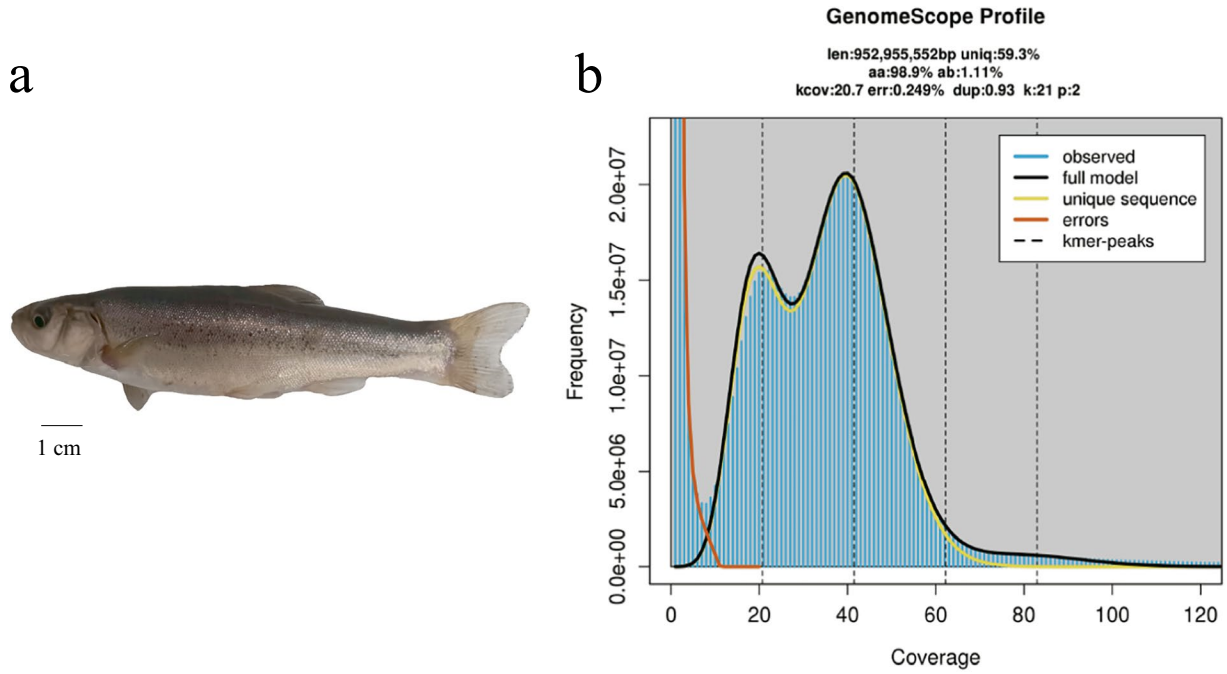

Fig. 1 Genome assembly of *R. lagowskii*. (**a**) The photo of male *R. lagowskii*; (**b**) The K-mer (21-mer) distribution curves for estimation of the genome size of *R. lagowskii*.

chromosome level would be helpful for future evolutionary and functional research, as well as molecular breeding and conservation of Amur minnow for improvement of its economic and ecological values.

Methods

Sample collection. Adult male Amur minnow (The body length and weight were 12.5 cm and 13.3 g, respectively) (Fig. 1a) was collected from Tianjin Tianxiang Aquatic Products Co., Ltd Ninghe District, Tianjin City, China. Muscle and skin tissues were pooled for whole genome sequencing, including MGI short-read, PacBio HiFi long-read, and Hi-C sequencing technologies. Meanwhile, Muscle, spleen, kidney, heart, and testis tissues were also collected for transcriptome sequencing. These samples were frozen in liquid nitrogen and stored at −80 °C before usage.

All animal experiments were approved by the Animal Welfare Committee of the Tianjin Fisheries Research Institute (Tianjin, China).

DNA extraction and genome sequencing. High molecular weight genomic DNA (gDNA) was prepared via a modifed SDS-based method[11], followed by purification using a Grandomics Genomic kit in accordance with the manufacturer's standard operating procedure. For quality and quantity assessment of the extracted DNA, a NanoDrop™ One UV-Vis spectrophotometer (Thermo Fisher Scientific, USA), a Qubit® 4.0 Fluorometer (Invitrogen, USA), and 1% agarose gel electrophoresis were employed, respectively.

Subsequently, the gDNA was randomly fragmented to construct a library with an average size of 200–400 bp by using MGIEasy universal DNA library Prep Kit V1.0 (MGI, Shenzhen, China) for subsequent sequencing on a DNBSEQ-T7RS platform (MGI). A total of 55.04 Gb of paired-end raw reads (150 bp in length) were generated. These reads were filtered with fastp v0.23.4 software[12] (parameter: -n 0 -f 5 -F 5 -t 5 -T 5 -q 20) to remove low-quality reads and adaptor sequences. Ultimately, approximately 49.92 Gb of clean reads were obtained (Table 1) for estimation of the genome size and error correction.

For the PacBio HiFi long-read sequencing, gDNA was utilized to construct long-read libraries with the SMRTbell® prep kit 3.0, strictly following PacBio's standard protocol. These libraries were then sequenced on a PacBio Revio System (Pacific Biosciences, Menlo Park, CA, USA). In total, 79.53 Gb of HiFi data were generated, featuring an average read length of 18,418 bp and an N50 size of 19,121 bp (Table 1).

For the high-throughput chromosome conformation capture (Hi-C) sequencing, one Hi-C library was constructed using a GrandOmics Hi-C kit according to the manufacturer's protocol (GrandOmics, Wuhan, China). Briefly, gDNA was first cross-linked with a 4% formaldehyde solution to stabilize the chromatin structure. Subsequently, the DNA was digested with the restriction enzyme MboI to introduce specific cleavage sites. The resulting DNA fragments were then labeled with biotin-14-dCTP, enabling the integration of a detectable marker. These labeled DNA fragments were ligated with T4 DNA ligase to facilitate subsequent enrichment steps. Following ligation, the DNA was further digested to obtain fragments within the size range of 200 to 600 bp. These fragments were then sequenced on a MGI DNBSEQ-T7 platform, with the PE150 module employed to generate paired-end reads. A total of 94.82 Gb of raw reads were generated. Subsequently, splice

Library type	Raw data (Gb)	Clean data (Gb)	Read N50/Length(bp)	Coverage
MGI	55.04	49.92	150	57.81
PacBio HiFi	/	79.53	19,124	83.53
Hi-C	94.82	93.77	150	99.6
RNA-seq				
Muscle	8.02	7.76	150	
Kidney	6.72	6.60	150	
Testis	8.38	8.26	150	
Heart	9.53	9.39	150	
Spleen	8.65	8.58	150	

Table 1. Sequencing data of the *R. lagowskii* genome. *For the PacBio HiFi sequencing, this number is read N50.

sequences and low-quality reads were filtered using fastp v0.23.4 software[12]. Ultimately, 93.77 Gb of Hi-C clean data were retained for chromosome assembly (Table 1).

RNA extraction and transcriptome sequencing (RNA-seq). Total RNA was extracted from five tissues (heart, spleen, kidney, testis, and muscle) using Trizol reagent (Invitrogen, CA, USA), followed by purification with a Qiagen RNeasy Mini Kit (Qiagen, Germantown, MD, USA). The purity and integrity of RNA were evaluated using a NanoDrop™ One UV-Vis spectrophotometer (Thermo Fisher Scientific, USA) and a Qubit® 4.0 Fluorometer (Invitrogen, USA), respectively. RNA contamination was assessed by 1% agarose gelelectrophoresis. Subsequently, RNA was used to construct an Illumina cDNA library in accordance with the manufacturer's instructions. The constructed library was then sequenced on a DNBSEQ-T7RS platform (MGI). Approximately 8.0 Gb of transcriptome data was generated for each tissue (for more detailed information, refer to Table 1), which was utilized to aid in gene annotations.

Genome size and heterozygosity Estimation. To characterize the genome of *R. lagowskii*, the k-mer analysis was conducted using Illumina DNA data prior to genome assembly, aiming to estimate the genome size and heterozygosity. Briefly, high-quality reads were subjected to 21-mer frequency distribution analysis with jellyfish v2.2.10[13]. Subsequently, a generated histogram served as an input file for GenomeScope v2.014[14] to predict genetic characteristics. K-mer analysis results revealed that the estimated genome size of *R. lagowskii* was 952.95 Mb, with a heterozygosity rate of 1.11% (Fig. 1b).

Genome assembly assessment. *De novo genome assembly.* Contigs were assembled using Hifasm v0.16.0[15] (default parameters) based on HiFi long-read length data. The results showed that the length of this primary genome assembly was 1077.26 Mb, and the N50 size was 32.63 Mb. Subsequently, to optimize the assembly, quality-controlled MGI short-read data was used with NextPolish v1.2.4[16] (default parameters) to correct the initial results. After redundancy removal, the genome size reached 1041.22 Mb, while the N50 remained 32.63 Mb. Based on this primary genome assembly, Hi-C short reads were then utilized to construct chromosomes for *R. lagowskii*.

Construction of chromosomes. First, the Hi-C clean reads were aligned to the assembled contigs using bowtie2 v2.3.2[17] (--very-sensitive -L 30--score-min L −0.6 −0.2--end-to-end--reorder). Subsequently, the HiC-Pro v3.1.0[18] pipeline was deployed to identify valid ligation products; only valid paired reads were retained for further analyses. Based on these valid reads, the primary assembly underwent orientation, ordering, and clustering onto chromosomes using Juicer v1.5[19] and 3D-DNA v3.0 software[20] (parameters: -m haploid -r 2 -c 25). Juicebox v1.11.08[19] was employed to visualize before manually adjusting the candidate assemblies. Finally, a total of 25 chromosomes were delineated, achieving a chromosome assembly rate of 99.04% (Fig. 2a). The genome size after Hi-C-assisted assembly was counted as 1017.47 Mb, with a Scaffold N50 of 39,226,643 bp and a Contig N50 of 29,202,791 bp (Table 2). A chromosomal circle diagram was generated based on the 25 constructed chromosomes (Fig. 2b) using Circos v0.69-3[21].

Gene structure prediction and functional annotation. *Repetitive sequence annotation.* To detect repetitive sequences in the genome of *R. lagowskii*, both de novo and homology-based strategies were employed. For SSR (Simple Sequence Repeat) sequences: GMATA software[22] was used for analysis. It was found that there were 146,209 SSR sequences in the genome, accounting for 0.17% of the total genome length (Table 3). Regarding tandem repeats: TRF software[23], with default parameters, was applied. There were 75,722 tandem repeat sequences in the genome, with a total length of 14,169,939 bp. In total, the length of tandem repeats in the genome was 14,245,661 bp, accounting for 1.40% of the entire genome (Table 3). RepeatMasker v4.0.6[24] was utilized to predict the TE repeats based on the finally constructed repeat database. The results indicated that TE repeats accounted for 45.75%, and the total length of repetitive sequences was 547,037,662 bp, corresponding to a proportion of 53.76% (Table 3).

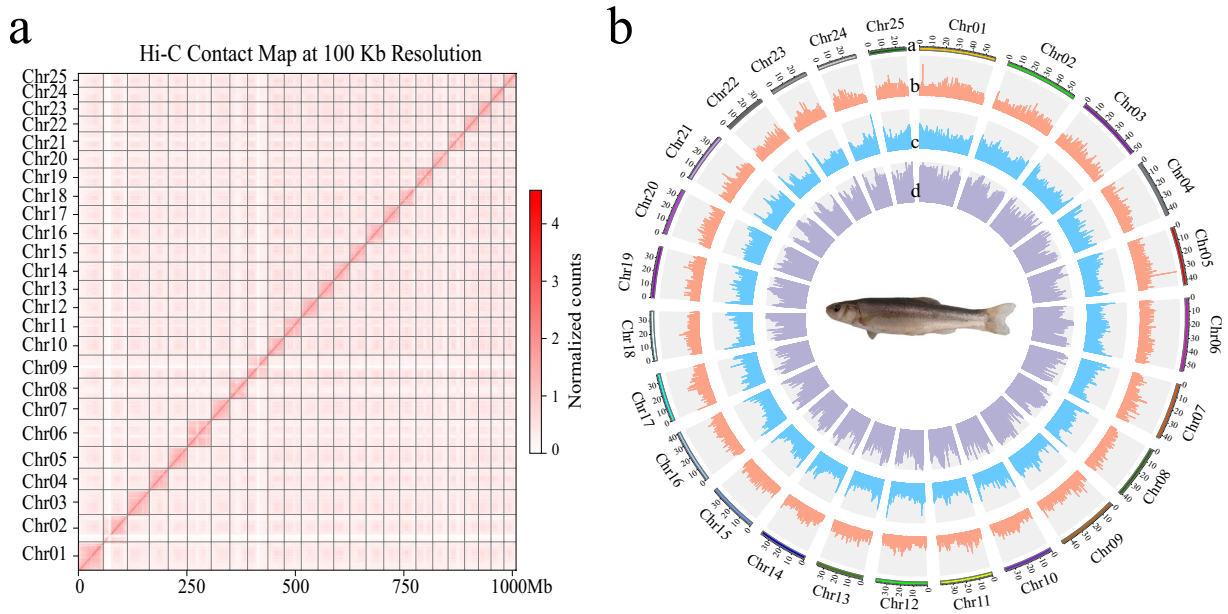

Fig. 2 Characteristics of the *R. lagowskii* genome. (**a**) Genome-wide analysis of chromatin interactions at 100-kb resolution in genome; The color bar shows the contact density from low (white) to high (red). (**b**) A Circos plot of the main genome features. From outside to inside include the 25 chromosomes, gene density, GC content and repetitive sequences density. The window size for density calculation is 100 kb.

Category	*Rhynchocypris lagowskii*
Genome survey (Mb)	952.95
Genome length (bp)	1,017,470,858
Longest scafold (bp)	57,596,106
Number of scaffolds	178
Contig N50 (bp)	29,202,791
Scaffold N50 (bp)	39,226,643
GC content	39.53%
Anchor ratio	99.04%
Number of chromosomes	25
Chromosome length (bp)	1,007,753,864

Table 2. Statistics of the *R. lagowskii* genome assembly.

Type			Length (bp)	Number of elements	% of Genome
Dispersed repeats	DNA transposons		238,960,543	1,940,323	23.49
	Retroelements	LINE	69,933,699	277,272	6.87
		LTR	95,540,395	418,241	9.39
		SINE	7,382,481	62,363	0.73
Tandem Repeats	Tandem repeat		14,169,939	75,722	1.39
	SSR		1,745,459	146,209	0.17
Simple repeats			8,739,801	43,626	0.86
Satellites			11,445,092	75,962	1.12
Other			553	166,935	0.02
Unknown			45,005,250	284,353	4.42
Total Repeats			547,037,662	3,560,290	53.76

Table 3. Classification of repetitive sequences in genome.

Protein-coding genes and functional annotations. Three methods were combined to annotate protein-coding genes, including *de novo* gene prediction, homology-based annotation, and transcriptome-based annotation. Based on the transcript sequences, the RNA-seq data from five tissues were assembled into contigs with Trinity

Method	Software/Species	Number	Average length (bp)				Average exon per gene
			gene	CDS	exon	intro	
De novo	Augustus	29,276	18,160.65	1,625.11	170.33	1,936.08	9.54
	Glimmer	63,927	14,637.28	798.57	153.18	3,284.51	5.21
Homolog	*D. rerio*	84,558	36,134.41	1,581.99	205.45	5,157.03	7.7
	L. waleckii	79,545	24,888.46	1,268.22	208.19	4,639.13	6.09
	C. carpio	101,356	31,572.28	1,530.21	188.05	4,209.07	8.14
	C. auratus	135,786	33,683.09	1,575.64	192.46	4,467.57	8.19
	C. idella	62,157	36,833.18	1,644.96	192.36	4,659.67	8.55
RNA-seq	PASA	24,870	18,477.53	2,770.65	279.95	1,765.4	9.9
Integrated	EVM	32,425	17,600.44	1,601.75	189.51	2,146.86	8.45

Table 4. Summary of the predicted gene structures using three methods.

Category	Number	Average length (bp)
Gene	32,425	17,600.44
Exons	8.45 (per gene)	189.51
Intron	7.45 (per gene)	2,146.86
Database	**Number**	**Percentage (%)**
Swissprot	24,105	74.34
KEGG	17,335	53.36
KOG	16,979	52.36
GO	17,632	54.38
NR	29,006	89.46
Annotated	29,535	91.09

Table 5. Functional annotation of predicted protein-coding genes.

v2.5.1[25]. Subsequently, gene structures were delineated using PASA v2.3.3[26]. A total of 24,870 genes were predicted, and the corresponding training model was generated for *de novo* prediction (Table 4).

GeMoMa v1.6.1[27] was applied for the homology-based prediction. We aligned homology proteins from five other fish species, including zebrafish (*Danio rerio*), Amur ide (*Leuciscus waleckii*), carp (*Cyprinus carpio*), crucian carp (*Carassius auratus*) and Grass carp (*Ctenopharyngodon idella*) (downloaded from the NCBI). All predictions from these homologous species were integrated to obtain the structural information of the corresponding predicted genes, resulting in the prediction of 42,301 genes. Based on the genes predicted from the transcriptome, 3,000 reliable genes were selected and trained with AUGUSTUS[28] to obtain the prediction model of the species. Subsequently, leveraging this training model, AUGUSTUS was used to predict the gene structure, yielding 29,276 genes (Table 4). Using Evidence Modeler (EVM)[29], we integrated the gene prediction results from PASA, GeMoMa, and *ab initio* methods following predefined weights (default hierarchy: PASA $\geq$ GeMoMa $\geq$ ab initio), to obtain the initial genomic gene set. Genes containing transposable elements and coding errors were filtered out using TransposonPSI[30]. A total of 32,425 genes were predicted from the genome, with an average gene length of 17,600.44 bp, an average CDS length of 1,601.75 bp, an average exon length of 189.51 bp and an average intron length of 2,146.86 bp (Table 4).

To predict the functions of these annotated genes, Blastp v2.2.26[31] (with parameters: -evalue 1e-5, -max_target_seqs 1)was used. The deduced protein sequences were queried against with five public databases, including the SwissProt[32], Gene Ontology (GO)[33], NCBI Non-Redundant Protein Sequence (NR), Kyoto Encyclopedia of Genes and Genomes (KEGG)[34] and Eukaryotic Orthologous Groups (KOG)[35] databases. Ultimately, 29,535 (91.09%) genes were successfully annotated with at least one database (Table 5 and Fig. 3).

Annotation of non-coding RNA genes. For annotation of non-coding RNA genes, tRNA-associated genes were annotated using tRNAscan-SE v2.0[36] with default settings; rRNAs annotation was performed via RNAmmer v1.2[37]; miRNAs and snRNAs were detected with Infernal v1.1.2[38] against the Rfam v14.1[39] database, following default parameters. In total, annotations predicted 7,226 rRNAs, 17,872 tRNAs, 980 miRNAs, and 290 snRNAs (see more details in Table 6).

Synteny analysis. Based on the protein-coding sequences and gene structures, we employed JCVI v v1.0.9[40] to conduct a chromosomal synteny analysis between *R. lagowskii* and its closely related species, *Phoxinus phoxinus*[41]. Results revealed a robust collinearity relationship, with chromosomes exhibiting one-to-one correspondence (Fig. 4), indicating that the assembled genomes are complete and high-quality.

Data Records

Files of the MGI, PacBio, Hi-C and transcriptome sequencing for *R. lagowskii* were deposited at NCBI under the accession number PRJNA1229864, Raw reads are available in the Sequence Reads Archive (SRA) with the accession numbers SRX28916492 to SRX28916500[42]. This Whole Genome Shotgun project has been deposited

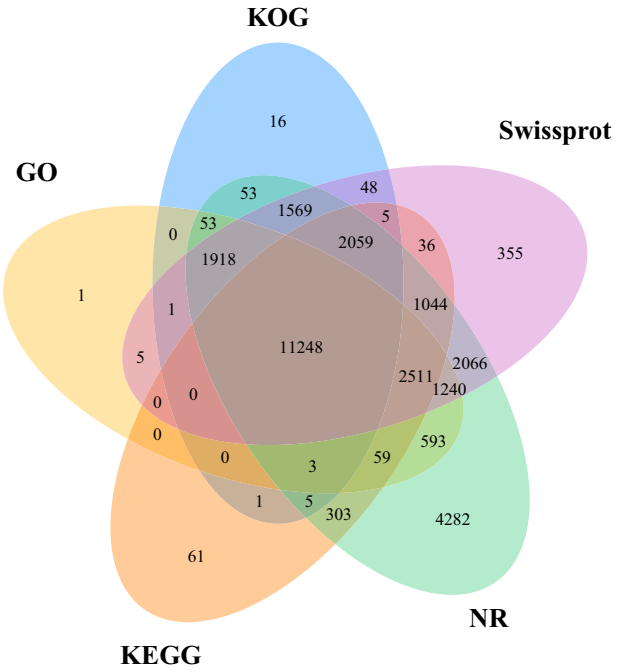

Fig. 3 Venn diagram of annotation results from the SwissProt, GO, NR, KEGG and KOG databases.

Type		Copy Number	Average length(bp)	Total Length(bp)	Percentage of sequence (%)
rRNA	18 s	25	2127.96	53,199	0.0052
	28 s	20	4551.8	91,036	0.0089
	5.8 s	2	154.00	308	0.0000
	5 s	7,179	115.45	828,782	0.0815
snRNA	snRNA	290	138.36	40,124	0.0039
	spliceosomal	1,867	157.95	294,888	0.0290
miRNA		980	89.06	87,279	0.0086
tRNA		17,872	75.23	1,344,445	0.1321
Regulatory	cis-regulatory elements	2,823	53.72	151,653	0.0149

Table 6. Statistics of the non-coding RNA annotations.

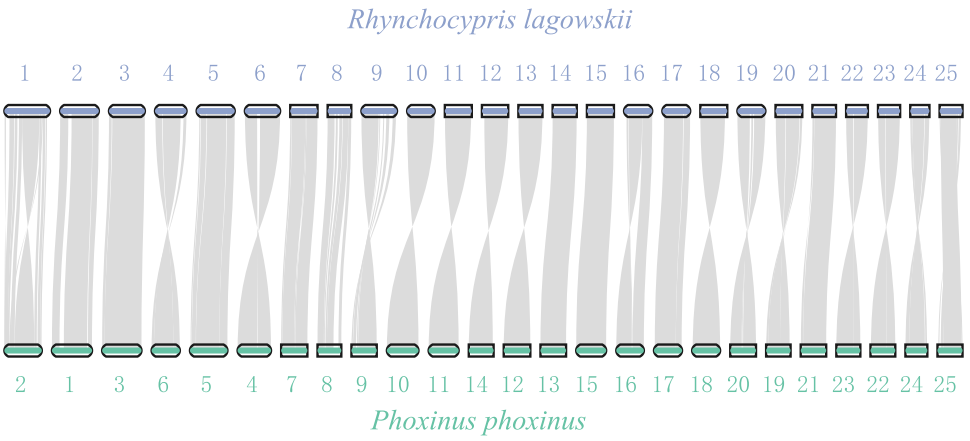

Fig. 4 Genome synteny between *R. lagowskii* and *P. phoxinus*.

at NCBI/GenBank under the accession GCA_050574305.1[43]. Furthermore, the genome assembly, predicted coding sequences and function annotation files of *R. lagowskii* were stored in Figshare (No: m9.figshare.28817660)[44].

Type	Evaluation Methods		Results
Genome accuracy and completeness	Mapping short reads rate		99.16%
	Mapping HiFi reads rate		99.92%
	QV	HiFi reads	68.71
		Short reads	50.44
	BUSCO		96.32%
Annotation quality	Complete BUSCOs		97.53% (3,550)
	Complete and single-copy BUSCOs (S)		95.63% (3,481)
	Complete and duplicated BUSCOs (D)		1.09% (69)
	Fragmented BUSCOs (F)		0 .05% (2)
	Missing BUSCOs (M)		2.42% (88)

Table 7. Evaluation of the genome assembly and annotation.

Technical Validation

To evaluate the quality of our genome assembly, three methods were used to assess the quality of assembled genomes. First, the integrity of the genome assembly was evaluated by employing BUSCO (Benchmarking Universal Single-Copy Orthologs) version 5.0[45]. 96.32% (Single Copy Complete Genes: 94.18%, Duplicated Complete Genes: 2.14%) of complete BUSCOs in the actinopterygii_odb10 database were identified. Second, Merqury v1.3[46] was applied to estimate the base-level accuracy and completeness based on k-mer counts, resulting in a QV of 68.71 and 85.11%, respectively. Third, all the HiFi data from sequencing were back-compared to the assembled error-corrected genome by minimap2 (-ax map-hifi)[47], and the comparison rate was counted. The overall number of reads was 4,317,794, and 99.92% (4,314,505) of the reads could be compared back to the reference genome. The BUSCO completeness value calculation showed that 97.53% of the predicted protein-coding genes were complete (Table 7).

Data availability

In this study, sequencing data for *R. lagowskii*, including MGI, PacBio, Hi-C, and transcriptome sequencing, have been deposited in the NCBI BioProject database under the accession number PRJNA1229864. The raw sequencing reads are available in the NCBI Sequence Read Archive (SRA) under the accession numbers SRX28916492 to SRX28916500[42]. This Whole Genome Shotgun project has been deposited at NCBI/GenBank under the accession GCA_050574305.1[43]. Furthermore, the genome assembly, predicted coding sequences and function annotation files of *R. lagowskii* were stored in Figshare (No: m9.figshare.28817660)[44].

The workflow scripts for protein-coding sequences, circos and repeat masking have been deposited on GitHub (https://github.com/huandna/Genome-Assembly).

Code availability

The versions and parameters of bioinformatics tools applied in this study have been described in the Method section. If no parameter is provided, the default is set. The workflow scripts for protein-coding sequences, circos and repeat masking have been deposited on GitHub (https://github.com/huandna/Genome-Assembly).

Received: 30 May 2025; Accepted: 26 August 2025;
Published online: 22 October 2025

References

1. Xu, Z. Y. Biological characteristics and breeding prospects of *Rhynchocypris lagowskii*. *China Fisheries*, 62-64 (2015).
2. Zu, Z. J. *et al.* The analysis of the flesh rate and nutrient composition in the muscle of *Phoxinus lagowskii* Dybowskii under artificial breeding. *J. Inner Mongolia Agric. Univ. (Nat. Sci. Edn.)* **37**, 6–13 (2016).
3. Liu, Y. H. *et al.* Comparative analysis of nutrient composition in the muscle of *Phoxinus lagowskii* Dybowskii with wild and artificial breeding. *J. Jilin Agric. Univ.* **39**, 82–89 (2017).
4. Gou, N. N. *et al.* Population characteristics and seasonal variation of *Brachymystax lenok tsinlingensis*'s Main Bait Organism *Lagowskiella lagowskii* in national nature reserves of rare aquatic wildlife of Heihe River. *Acta. Agric. Boreali-occident. Sin.* **27**, 1258–1264 (2018).
5. Xie, P. Morphological comparison and temperature adaptation of *Phoxinus* species in China. Wuhan: Huazhong Agricultural University (2019).
6. Bo, Q. K. *et al.* Effect of Temperature Acclimation on Growth and Thermal Tolerance of Juvenile *Rhynchocypris lagowskii*. *Fish. Sci.* **44**, 83–90 (2025).
7. Luo, X. N. *et al.* Artificial propagation of fat minnow *Phoxinus lagowskii* Dybowski. *Fish. Sci.* **32**, 673–675 (2013).
8. Luo, X. N. *et al.* Fry and juvenile rearing of fat minnow *Phoxinus lagowskii* Dybowski in an earthen pond. *Fish. Sci.* **33**, 186–189 (2014).
9. Zhang, S. *et al.* Karyotype of *Phoxinus lagowskii*. *J. Guangdong Ocean Univ.* **36**, 106–109 (2016).
10. Yang, P. M. *et al.* Genetic diversity and differentiation of 3 geographic populations of fat minnow *Phoxinus lagowskii* Dybowsky revealed by D-loop and Cyt b gene analyses. *Chin. J. Fish.* **30**, 7–12 (2017).
11. Futoshi, A. PCR-RFLP analysis of nuclear nontranscribed spacer for mackerel species identifcation. *J. Agric. Food. Chem.* **53**, 508–511 (2005).
12. Chen, S. *et al.* fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, 884–890 (2018).
13. Guillaume, M. & Carl, K. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
14. Ranallo-Benavidez, T. R. *et al.* GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).
15. Cheng, H. *et al.* Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
16. Hu, J. *et al.* NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **3**, 2253–2255 (2020).

17. Langmead, B. & Salzberg, S. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
18. Servant, N. *et al.* HiC-Pro: An optimized and flexible pipeline for Hi-C processing. *Genome Biol.* **16**, 259 (2015).
19. Durand, N. C. *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
20. Dostie, J. & Dekker, J. Mapping networks of physical interactions between genomic elements using 5 C technology. *Methods Enzymo.* **423**, 197–214 (2007).
21. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).
22. Wang, X. & Wang, L. GMATA: An Integrated Software Package for Genome-Scale SSR Mining, Marker Development and Viewing. *Front. Plant Sci.* **7**, 1350 (2016).
23. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
24. Bedell, J. A. *et al.* MaskerAid: a performance enhancement to RepeatMasker. *Bioinformatics* **11**, 1040 (2000).
25. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
26. Haas, B. J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 3715–3724 (2003).
27. Hoff, K. & Stanke, M. Predicting genes in single genomes with GeMoMa. *Bioinformatics* **35**, 447–449 (2019).
28. Stanke, M. *et al.* Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
29. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
30. Urasaki, N. *et al.* Draft genome sequence of bitter gourd (Momordica charantia), a vegetable and medicinal plant in tropical and subtropical regions. *DNA Res.* **24**, 51–58 (2017).
31. Altschul, S. F. *et al.* Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
32. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28**, 45–48 (2000).
33. Ashburner, M. *et al.* Gene ontology: a tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
34. Kanehisa, M. *et al.* KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).
35. Galperin, M. Y. *et al.* Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.* **43**, 261–269 (2015).
36. 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).
37. Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108 (2007).
38. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
39. Griffiths-Jones, S. *et al.* Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, 121–124 (2005).
40. Tang, H. *et al.* JCVI: A Versatile Toolkit for Comparative Genomics. *Analysis. iMeta* **3**, 211 (2024).
41. Sakai, H. *et al.* Phylogenetic and taxonomic relationships of northern far eastern phoxinin minnows, *Phoxinus* and *Rhynchocypris* (Pisces, Cyprinidae), as inferred from allozyme and mitochondrial 16S rRNA sequence analyses. *Zool. Sci.* **23**, 323–331 (2006).
42. *NCBI Sequence Read Archive.* https://identifiers.org/ncbi/insdc.sra:SRP587423 (2025).
43. Xiaohui, B. *et al. Genbank* https://identifiers.org/ncbi/insdc.gca:GCA_050574305.1 (2025).
44. Xiaohui, B. Genome assembly, predicted coding sequences and functional annotation files of *R. lagowskii*. *Figshare.* https://doi.org/10.6084/m9.figshare.28817660 (2025).
45. Simao, F. A. *et al.* BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
46. Rhie, A. *et al.* Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).
47. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016).

Acknowledgements

This work was supported by Tianjin Science and Technology Plan (no. 22ZYCGSN00820).

Author contributions

W.Z. conceived and designed the study. X.B., J.L., and Z.L. collected the samples. H.L., Y.W., X.L. and X.Z. performed data analysis. L.M., C.W. and S.H. conducted experiments for species identification. X.B. and H.L. wrote the manuscript. W.Z. revised the manuscript. All authors read and approved the final manuscript for publication.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to W.Z.

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.