scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A chromosome level genome assembly of Homatula variegata from the Yangtze River basin

Yong Tang ✉, Qiaoxing Wu, Yusu Wang, Shaoqi Jiang, Lan Liu & Lingjin Xian ✉

Homatula variegata **is a small benthic loach from the upper Yangtze and adjacent basins with aquaculture and ornamental value but no reference genome. We present a near telomere-to-telomere (T2T) chromosome-level assembly built from PacBio HiFi, Oxford Nanopore ultra-long, Illumina short reads, and Hi-C. The 641.26-Mb genome resolves 24 chromosomes as single contigs (contig N50, 24.40 Mb). Hi-C confirms chromosome-length scaffolding; we detect 24 putative centromeres and 20 terminal telomeric tracts, with 22 chromosomes gap-free and two containing one gap. Annotation identifies 24,479 protein-coding genes, 93% functionally assigned, and 27.13% repetitive content dominated by DNA transposons. Quality assessments show high completeness (BUSCO, 96.48% complete) and base-level accuracy consistent with k-mer and read-mapping metrics. To our knowledge this is the first near T2T-level reference for any loach (Cobitoidei), filling a key gap in Cypriniformes genomics. This resource will enable comparative and population genomics, illuminate adaptation to montane stream habitats, and support selective breeding, conservation, and aquaculture of this native species.**

Background & Summary

Homatula variegata (syn. *Paracobitis variegatus*) is a small benthic fish belonging to the order Cypriniformes and subfamily Nemacheilinae. It is distributed across multiple freshwater systems in China, including the main and tributary streams of the upper Yangtze River in Sichuan and Chongqing, southern Shaanxi, the Bailong River in Gansu, and the Jinsha and Nanpan Jiang rivers in Yunnan[1,2]. *H. variegata* inhabits mid-elevation montane streams with pebble substrates and moderate flow. It reaches a maximum total length of approximately 13.9 cm and feeds primarily on benthic invertebrates, detritus, and small fishes, consistent with the ecological traits of Nemacheilinae loaches[3]. Owing to its palatable flesh, high nutritional content, and vibrant coloration, the species holds economic and ornamental value, particularly as a native ornamental fish with aquaculture potential[4].

Research to date has focused on its biology[3] and genetic structure[5,6]. However, no chromosome-scale nuclear genome has been reported for *H. variegata*. The absence of a high-quality reference genome limits investigations into its evolutionary divergence, local adaptation, population genomics, and the genetic basis of ecologically and economically relevant traits. T2T genome assemblies – gap-free chromosome sequences spanning from one telomere to the other – have recently become attainable thanks to innovations in long-read sequencing and assembly algorithms[7]. This complete coverage is crucial for capturing genomic regions that were previously unresolved in draft genomes, such as highly repetitive telomeric and centromeric DNA and segmental duplications, thereby revealing novel genes and structural variations that underpin evolution and important[8]. High-quality T2T references have now been reported for a few fish species (e.g., common carp/koi and Asian seabass) – demonstrating the feasibility and value of gap-free assemblies in non-model organisms[9,10]. However, no such resource existed for loaches (suborder Cobitoidei), leaving a significant knowledge gap in this lineage. Addressing this gap is a primary goal of our study, as a T2T reference genome for *H. variegata* would provide a foundation to explore its genomic diversity, adaptive evolution in montane stream environments, and to inform selective breeding, conservation, and aquaculture efforts for this ornamental species.

In this study, we present the first near T2T nuclear genome assembly of *H. variegata*, generated using a hybrid approach combining PacBio HiFi long reads, Oxford Nanopore ultra-long reads, and Hi-C chromatin conformation capture. A total of 24 chromosomes were assembled, with a genome size of 641.26 Mb and an N50 value

School of Modern Agriculture, Leshan Vocational and Technical College, Leshan, Sichuan Province, 614000, China. ✉e-mail: ty20042028@163.com; xljin01@163.com

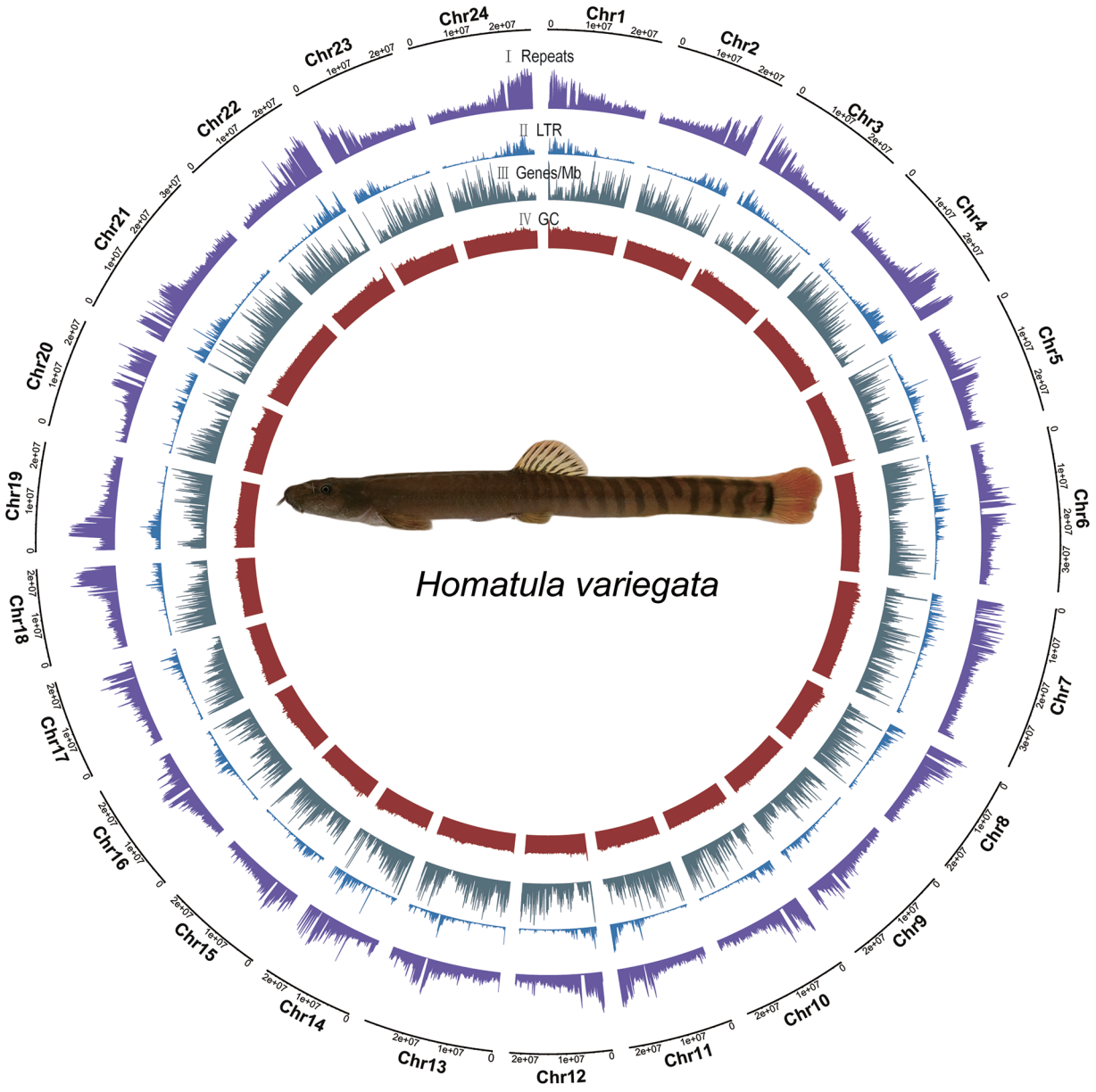

Fig. 1 Circos plot summarizing major genomic features of the assembled genome. From the outside inward: the 24 chromosomes (Chr1-Chr24) arranged clockwise; (I) repeat coverage per 100-kb window (purple, value = fraction of window overlapped by any TE); (II) LTR coverage per 100-kb window (blue); (III) gene density (grey-blue, number of genes per Mb) calculated from gene models; and (IV) GC content (red) computed as (G + C)/(A + T + G + C) within each 100-kb window.

of 24.40 Mb (Fig. 1). This assembly fills a critical gap in genomic resources for Nemacheilinae and represents the first near T2T genome for any species within the suborder of Cobitoidei. The genome provides a valuable reference for advancing genetic improvement, resource management, and aquaculture development of *H. variegata* as a native ornamental and economic species.

Methods

Ethics statement. All procedures involving *H. variegata* complied with institutional and national regulations. The protocol was approved by Science and Technology Ethics (Review) Committee of Leshan Vocational and Technical College (approval ID: 202403).

Sample collection. A live adult *H. variegata* (≥15 g) male was captured by licensed fishers from the Ya'an reach of the Qingyi River (upper Yangtze, China) on 21 August 2024. The fish was rinsed with pre-chilled PBS (4 °C), surface-dried with sterile gauze, and kept on ice. Venous blood (≥0.5 mL) was drawn from the caudal vein using a sterile 1 mL syringe, transferred into pre-chilled EDTA tubes, gently inverted to mix, and immediately snap-frozen in liquid nitrogen. The same blood sample was used for Oxford Nanopore Technologies (ONT)

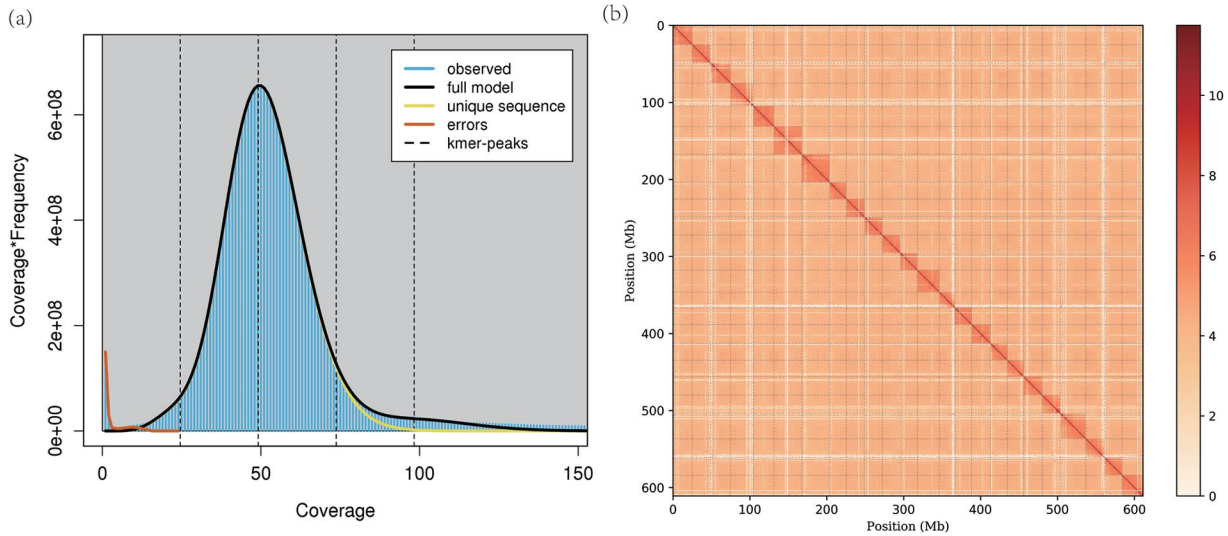

Fig. 2 Genome profiling and chromosome-scale scaffolding of *H. variegata*. (**a**) 21-mer frequency distribution of Illumina short reads with the GenomeScope2 fit, estimating a genome size of ~641.3 Mb, heterozygosity of ~0.20%, and repeat content of ~36.53%. (**b**) KR-normalized Hi-C contact heatmap (DpnII-based; Juicer/3D-DNA) of the final assembly, revealing 24 chromosome-length linkage groups with strong on-diagonal signal and minimal inter-chromosomal noise, consistent with correct scaffolding.

ultra-long read, PacBio HiFi, Illumina whole-genome shotgun (Illumina WGS) and high-throughput chromosome conformation capture (Hi-C) library preparations.

Nucleic-acid extraction and QC. *Genomic DNA from whole blood (for Illumina/ONT/PacBio/Hi-C).* High-molecular-weight genomic DNA (gDNA) was extracted from EDTA-anticoagulated whole blood using the Kurabo QuickGene DNA whole blood kit S (DB-S, 40321300101) following the manufacturer's instructions. Concentration was measured by Qubit Fluorometer (Invitrogen, Qubit 3.0/4.0; dsDNA HS Assay), purity by NanoDrop 2000/One (Thermo), and integrity by 0.8% agarose gel electrophoresis (Tanon EPS600; HE-120 tank). Library preparation proceeded only when total mass $\geq$6 µg (per PacBio single-library requirement), concentration $\geq$ 50 ng/µL, OD260/280 = 1.7–2.2, OD260/230 = 1.7–2.5, and HMW smear $\geq$ 23 kb with minimal degradation.

Total RNA from whole blood (for RNA-seq). Total RNA was isolated from EDTA blood using TRIzol, followed by column cleanup (Qiagen RNeasy) with on-column DNase I. Quantity (NanoDrop; Qubit RNA HS), integrity (PerkinElmer LabChip GX Touch HT or Agilent Bioanalyzer) and purity were assessed. Libraries were constructed only if RIN $\geq$ 7 and 28S/18S $\geq$ 1.8.

Library construction and sequencing. *Illumina short-insert WGS.* A paired-end library with approximately 350 bp insert size was constructed from genomic DNA using the VAHTS™ Fg DNA Library Prep Kit for Illumina (ND617-02, Vazyme/Yeasen), following enzymatic fragmentation, end repair, A-tailing, adaptor ligation, bead-based size selection, and limited-cycle PCR. The library was quantified with Qubit, and the fragment size distribution evaluated by Qsep-400 (peak 430–530 bp, mean 420–580 bp, single-peak profile). Sequencing was performed on an Illumina NovaSeq X Plus platform using NovaSeq™ X Series 25B reagent kits (PE150). Adapter sequences were: 5′-AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC-3′ and 5′-AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT-3′.

Raw reads were filtered using *fastp* v0.23.4 (with default settings)[11], and read quality summaries generated by FastQC v0.12.1 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). A total of 39.41 Gb of raw paired-end data were generated, corresponding to an estimated sequencing depth of approximately 61.5 × for the ~641.3 Mb genome. Base-level quality metrics were excellent, with Q20 $\geq$ 99.92% and Q30 $\geq$ 99.30%, and GC content measured at 39.96%. Post-filtering, clean reads were used to estimate genome properties: genome size $\approx$ 641.3 Mb, heterozygosity $\approx$ 0.20%, and repeat content $\approx$ 36.53% (Fig. 2a).

ONT ultra-long read sequencing. High-molecular-weight (HMW) gDNA was used to construct ONT ultra-long libraries using the ONT Ultra-long DNA Sequencing Kit V14 (SQK-ULK114), following the manufacturer's protocol: transposase tagmentation (room temperature 5 min; 75 °C 5 min), adaptor ligation (room temperature 45 min), DNA precipitation and resuspension in 80 µL Elution Buffer with $\geq$ 24 h of dissolution. Libraries were sequenced on a PromethION 48 using R10.4.1 flow cells (FLO-PRO114M), with runs controlled by MinKNOW software for 72 h.

Basecalling was performed using the MinKNOW-integrated basecaller, or Dorado v0.5.3 (SUP mode) when offline. Adapter trimming utilized Porechop v0.2.4 (--discard_middle) (https://github.com/rrwick/Porechop), followed by filtering with Filtlong v0.2.1 (https://github.com/rrwick/Filtlong), and QC summaries via NanoPlot v1.41.6[12].

A total of 26.45 Gb of raw ONT reads were generated. After adapter trimming, removal of short and low-quality reads, 22.49 Gb of clean data remained, with metrics including N50 read length of 100 kb, N90 of 39 kb, mean read length of 78 kb, maximum read length of 1,120 kb, and an average per-read quality score of Q12. These ultra-long reads were essential for spanning complex repeat regions and achieving T2T assembly completeness, in keeping with established long-read assembly practice.

PacBio HiFi (CCS). SMRTbell libraries were prepared using SMRTbell Prep Kit 3.0 (PacBio, 102-182-700), purified with AMPure PB and SMRTbell cleanup beads, and sequenced on PacBio Revio using REVIO SPRQ Polymerase Kit (103-496-900), REVIO SPRQ Sequencing Plate (103-504-900) and REVIO SMRT Cell Tray (102-202-200). CCS generation used Pacbio SMRT Link v13.0 (ccs; --min-passes 3 --min-rq 0.99).

A targeted data output of 42.0 Gb was planned for the project. Sequencing of the single sample yielded ~40.82 Gb of high-quality CCS data, encompassing approximately 2.68 million reads. The reads had a mean length of 15.25 kb, an N50 of 15.34 kb, and a maximum read length of 61.19 kb. The CCS N50 surpassed the QC threshold of ≥10 kb, confirming its suitability for *de novo* assembly. Moreover, sampling of ~10% CCS reads against the NCBI NT database showed 18.46 mapped reads (13.01% to *Cyprinus carpio*, 2.20% to *Danio rerio*). We emphasize that these best-hit species labels are influenced by database representation and conserved regions, and are not interpreted as evidence of contamination or species identity. To confirm sample identity, we extracted the assembly contig that contains the mitochondrial COI (*cox1*) region and searched it against NCBI nt. The top hit was Homatula laxiclathra (query coverage = 100%, identity = 98.97%), supporting assignment to genus Homatula (Nemacheilidae).

Hi-C libraries. Hi-C libraries were constructed from the blood sample using a DpnII-based protocol: formaldehyde crosslinking, cell lysis, DpnII digestion, end-repair with biotin-dNTPs, proximity ligation, crosslink reversal, purification, fragmentation to ~300–700 bp, and streptavidin bead capture. Libraries were quantified by Qubit and sized by Agilent 2100/Qsep-400, then sequenced on NovaSeq X Plus (PE150).

Adapters and low-quality bases were trimmed with fastp v0.23.4[11]. Reads were aligned to the draft assembly with BWA-MEM2 v2.2.1 (−5 -SP)[13], valid pairs identified with HiC-Pro v3.1.0[14], and.hic matrices produced with Juicer v1.7.6[15] for visualization.

The Hi-C dataset comprised 69.79 Gb of clean data, with Q30 ≥ 91.8%, meeting project quality benchmarks. Library complexity was high, evidenced by > 47.2% of reads spanning valid restriction enzyme junctions, consistent with typical high-quality Hi-C libraries. These data provided robust chromosomal contact information, essential for accurate scaffolding and correction of assembly structure (Fig. 2b).

Genome profiling. Genome size, heterozygosity and repeat content were estimated from Illumina clean reads using Jellyfish v2.3.0[16] and GenomeScope2[17]. K-mer analysis (k = 21) based on a 350-bp Illumina library estimated a haploid genome size of ~641.3 Mb, repeat content of ~36.53%, and heterozygosity of ~0.20%.

De novo assembly and polishing. Primary contigs were assembled with hifiasm v0.19.8[18] using HiFi reads as the primary input and ONT ultra-long reads as an auxiliary input in HiFi + UL hybrid mode (--ul) with otherwise default parameters. Long-read mappings used minimap2 v2.26[19]. Haplotypic duplications were removed using purge_dups v1.2.6[20] to generate a collapsed assembly for downstream polishing and scaffolding. Two rounds of short-read polishing were performed with NextPolish v1.4.1 (Illumina)[21]. To monitor exogenous contamination during assembly, we randomly sampled ~10% of the PacbBio CCS/HiFi reads ($n = 267{,}000$) and queried both the read subset and assembly contigs against the NCBI nt database using BLASTN v2.13.0[22] (-evalue 1e-25 -max_target_seqs. 5). Taxonomic matches were inspected at higher ranks to identify obvious non-target signals and any contigs showing strong matches to clearly unrelated taxa, which would trigger re-evaluation prior to scaffolding. The final assembly spans 641.26 Mb and is anchored to 24 chromosomes (Fig. 1), with a contig N50 of 24.40 Mb.

Chromosome scaffolding and manual curation. Hi-C valid pairs were used to scaffold contigs with 3D-DNA[23] (release 180922; run-asm-pipeline.sh -m diploid -r 0), followed by manual curation in Juicebox Assembly Tools v1.11.08[24]. As an orthogonal check, YAHS v1.2.2[25] (default parameters) was also run; only concordant joins were retained. Residual gaps were closed with TGS-GapCloser v1.2.1[26] using both HiFi and ONT reads.

Telomere and centromere features. Telomeric (TTAGGG)n arrays were identified with tidk v0.2.0 (find-telomeres)[27] at scaffold ends. Tandem Repeats Finder v4.09.1[28] annotated satellite-rich regions. Putative centromeres were inferred from (i) long tandem arrays and (ii) local insulation in Hi-C maps using cooltools v0.5.4[29]. Ultimately, we resolved all 24 chromosomes as single contigs and identified 24 centromeres and 20 telomeric tracts (9 upstream, 11 downstream); two chromosomes are capped at both ends, while six lack detectable telomere arrays (Table 1).

Repeat annotation. A *de novo* repeat library was generated with RepeatModeler2 v2.0.4[30] (including LTRharvest/LTR_retriever v2.9.0[31] and MITE-Hunter v11-2011[32]), classified with TEclass v2.1.0[33], merged with Repbase Update[34], and applied with RepeatMasker v4.1.6[35]. Tandem repeats were added from TRF v4.10.0[28]; redundant consensus entries were collapsed with cd-hit-est v4.8.1 (-c 0.9)[36]. In total, 686,404 repeats span 174.00 Mb, accounting for 27.13% of the genome. DNA transposons dominate (14.65%)—notably CACTA (3.70%), ClassII/Unknown (3.31%), Helitron (2.17%), hAT (1.73%) and Tc1–Mariner (1.43%). Retrotransposons comprise 12.45%, mainly LINEs (4.24%) and LTR elements such as Gypsy (2.19%) and LTR/Unknown (3.82%). (Table 2).

Chr	Contigs	Contigs Length (bp)	Gaps	Centromere		Telomere			
				Start Pos	End Pos	Upstream Start (bp)	Upstream End (bp)	Downstream Start (bp)	Downstream End (bp)
LG01	1	25,034,042	0	4,614,664	5,615,333	—	—	25,031,349	25,034,042
LG02	1	24,138,896	0	17,439,705	18,101,404	1	6018	—	—
LG03	1	25,875,418	0	5,441,745	6,561,908	—	—	25,866,822	25,875,418
LG04	1	29,470,937	0	24,338,219	26,702,069	1	6968	29,468,311	29,470,878
LG05	1	26,971,872	0	12,994,250	13,803,122	—	—	26,962,164	26,971,872
LG06	1	36,137,583	0	16,570,802	18,324,963	—	—	—	—
LG07	1	36,069,280	1	11,540,992	12,045,756	—	—	36,061,267	36,069,271
LG08	1	21,907,463	0	5,135,691	5,475,430	—	—	—	—
LG09	1	23,528,384	0	15,620,176	16,633,023	1	3448	—	—
LG10	1	23,324,279	0	4,017,339	5,564,194	—	—	23,317,309	23,324,195
LG11	1	23,035,761	1	7,265,086	7,631,375	1	1596	—	—
LG12	1	22,421,582	0	4,554,621	5,838,688	—	—	22,420,314	22,421,582
LG13	1	28,663,172	0	18,937,775	19,909,336	—	—	—	—
LG14	1	20,247,748	0	16,194,167	17,061,444	1	7031	—	—
LG15	1	21,829,513	0	6,022,519	6,920,656	1	3444	21,824,442	21,829,513
LG16	1	24,398,077	0	13,451,687	14,392,030	—	—	24,391,981	24,398,077
LG17	1	21,963,828	0	21,999,339	24,420,426	—	—	21,960,072	21,963,828
LG18	1	21,216,350	0	16,442,062	17,086,228	—	—	—	—
LG19	1	24,020,927	0	3,399,193	4,236,042	1	9188	—	—
LG20	1	24,051,685	0	16,269,960	17,924,280	—	—	—	—
LG21	1	32,735,689	0	6,006,300	7,811,279	—	—	—	—
LG22	1	24,441,762	0	16,291,854	16,872,664	1	9647	—	—
LG23	1	22,760,638	0	3,688,513	4,658,942	—	—	22,756,623	22,760,638
LG24	1	26,950,966	0	20,043,526	20,865,722	9	1697	—	—

Table 1. Per-chromosome contiguity and coordinates of centromeric and telomeric tracts in the *H. variegata* assembly.

Gene and transcript annotation. RNA-seq reads (blood only) were QC'd with fastp v0.23.4[11], aligned with HISAT2 v2.2.1[37] (--rna-strandness RF -k 10), and assembled with StringTie v2.2.1[38]. Homology-based predictions used GeMoMa v1.9[39] with teleost references. Ab initio models were predicted by AUGUSTUS v3.5.0[40]. EVidenceModeler v1.1.1[41] integrated evidence with weights set to RNA-seq = 50, homology = 50, and ab initio = 0.3, prioritizing transcript and homology support while retaining ab initio predictions as low-weight complementary evidence. However, because RNA-seq data were available only from blood, tissue-specific transcripts may be underrepresented; multi-tissue RNA-seq will be incorporated in future annotation updates. Functional annotation used DIAMOND v2.1.9[42] (NR; --evalue 1e-5 --max-target-seqs1), InterProScan v5.61-93.0[43], eggNOG-mapper v2.1.12[44], and KofamScan v1.3.0[45]. In total, 24,479 protein-coding genes (93%) were functionally annotated in at least one database. Coverage by resource was: TrEMBL 24,065 (91.43%), nr 23,475 (89.18%), Pfam 22,304 (84.74%), GO 21,453 (81.5%), KEGG 21,109 (80.2%), eggNOG 20,981 (79.71%), Swiss-Prot 16,319 (62.0%), and KOG 15,904 (60.42%) (Table 3).

Annotation of non-coding RNA genes. Non-coding RNAs were annotated with tRNAscan-SE v2.0.12[46], Infernal v1.1.4 (Rfam 14.9)[47] and barrnap v0.9 (https://github.com/tseemann/barrnap). We annotated major ncRNA classes totaling ~4.67 Mb (~0.729% of the genome). The set comprises tRNAs (26,565 copies; 2.00 Mb; 0.3123%), rRNAs (13,293; 2.58 Mb; 0.4026%)—dominated by 5S rRNA (12,706 copies; 1.45 Mb; 0.2266%) with additional 18S (179; 0.0514%) and 28S (232; 0.1204%) and 5.8S (176; 0.0042%)—snRNAs (512; 76.1 kb; 0.01186%) mainly spliceosomal (454; 68.1 kb), and miRNAs (174; 14.1 kb; 0.00221%) (Table 4).

Data Records

Raw sequencing data are available on the NCBI SRA under BioProject PRJNA1307247[48]. The dataset comprises five experiments: P.variegatus_HiFi (PACBIO_SMRT Revio; SRX30176691), P.variegatus_rnaseq (Illumina NovaSeq X Plus; SRX30158545), P.variegatus_hi-c (Illumina NovaSeq X Plus; SRX30158544), P.variegatus_illumina (Illumina NovaSeq X Plus; SRX30158543), and P.variegatus_ONT (Oxford Nanopore PromethION; SRX30158541).

The primary genome assembly (P. variegatus v1.0) is available at NCBI GenBank under the accession GCA_052674685.1[49]. The GenBank release includes: genomic sequences in fasta format (GCA_052674685.1_ASM5267468v1_genomic.fna.gz) and GenBank flat file format of the genomic sequences in the assembly (GCA_052674685.1_ASM5267468v1_genomic.gbff.gz).

Type	Number	Length	Rate(%)
ClassI:Retroelement	205,837	79,849,924	12
ClassI/DIRS	8,148	7,768,409	1
ClassI/LINE	71,968	27,158,612	4
ClassI/LTR/Copia	1,777	1,021,405	0
ClassI/LTR/ERV	11,069	3,535,719	1
ClassI/LTR/Gypsy	27,252	14,021,526	2
ClassI/LTR/Ngaro	1,026	150,849	0
ClassI/LTR/Pao	749	268,739	0
ClassI/LTR/Unknown	74,309	24,486,763	4
ClassI/SINE	9,539	1,437,902	0
ClassII:DNA transposon	479,723	93,944,983	15
ClassII/Academ	1,397	347,841	0
ClassII/CACTA	104,647	23,714,903	4
ClassII/Crypton	8,202	1,802,914	0
ClassII/Dada	2,111	427,108	0
ClassII/Ginger	623	63,744	0
ClassII/Helitron	110,973	13,916,333	2
ClassII/IS3EU	5,604	1,148,677	0
ClassII/Kolobok	11,724	2,513,427	0
ClassII/Maverick	1,001	217,839	0
ClassII/Merlin	2,695	368,658	0
ClassII/Mutator	2,686	587,678	0
ClassII/P	1,964	250,415	0
ClassII/PIF-Harbinger	17,940	4,344,538	1
ClassII/PiggyBac	2,471	449,735	0
ClassII/Sola	3,876	869,822	0
ClassII/Tc1-Mariner	23,962	9,166,941	1
ClassII/Unknown	110,581	21,256,798	3
ClassII/Zator	2,659	533,889	0
ClassII/Zisupton	5,110	843,086	0
ClassII/hAT	59,497	11,120,637	2
Unknown	832	202,241	0
srpRNA	12	3,555	0
Total	686,404	174,000,703	27

Table 2. Repeat landscape of the *H. variegata* genome.

Anno Database	Annotated Number	Annotated Ratio (%)
GO_Annotation	21,453	81.50
KEGG_Annotation	21,109	80.20
KOG_Annotation	15,904	60.42
Pfam_Annotation	22,304	84.74
Swissprot_Annotation	16,319	62.00
TrEMBL_Annotation	24,065	91.43
eggNOG_Annotation	20,981	79.71
nr_Annotation	23,475	89.18
All_Annotated	24,479	93.00

Table 3. Functional annotation summary of *H. variegata* protein-coding genes across databases.

Technical Validation

Read mapping and coverage. Illumina reads were aligned with BWA-MEM2 v2.2.1[13]; PacBio HiFi/ONT reads with minimap2 v2.26[19]. RNA-seq reads were aligned with HISAT2 v2.2.1[37]. Metrics were computed with samtools v1.20[50] and Picard v3.1.1 (https://broadinstitute.github.io/picard/).

QUAST v5.3.0[51] summarized a 641.26 Mb assembly with a contig N50 of 24.40 Mb. Each of the 24 chromosomes is represented by a single contig; 22 are gap-free, whereas LG07 and LG11 each contain one gap (Table 1). Illumina short reads mapped at 99.81% (proper pairs 95.07%) with a mean depth of 57 × and coverage of 99.93%, 99.79%, 99.56% and 98.59% at ≥ 1 × /5 × /10 × /20 × , respectively. PacBio HiFi reads mapped at 99.97%, with

Type		Copy	Average length (bp)	Total length (bp)	% of genome
miRNA		174	81.28	14,142	0.0022
tRNA		26,565	75.38	2,002,430	0.3123
rRNA	rRNA	13,293	194.23	2,581,865	0.4026
	18S	179	1840.87	329,516	0.0514
	28S	232	3328.81	772,285	0.1204
	5S	12,706	114.38	1,453,280	0.2266
	5.8S	176	152.18	26,784	0.0042
snRNA		512	148.58	76,075	0.0119
	splicing	454	149.93	68,066	0.0106
	HACA-box	25	145.32	3,633	0.0006
	CD-box	32	126.06	4,034	0.0006
	scaRNA	1	342	342	0.0001

Table 4. Statistics of the non-coding RNA annotations.

a mean depth of 63× and coverage of 99.98%, 99.73%, 99.49% and 98.19% at $\geq 1\times/5\times/10\times/20\times$. Merqury v1.3.1[52] indicated high k-mer completeness and base-level consensus quality, consistent with the mapping statistics. Gene-space completeness was high by BUSCO v5.6.1[53] (actinopterygii_odb10; --augustus --long): 96.48% complete (94.97% single-copy; 1.51% duplicated), 0.25% fragmented, and 3.27% missing (3,640 total). Although the BUSCO completeness is high, future re-annotation incorporating multi-tissue transcriptomes and updated pipelines may further improve completeness. CEGMA[54] further recovered 457/458 (99.78%) core eukaryotic genes and 246/248 (99.19%) highly conserved CEGs.

KR-normalized Hi-C matrices inspected in Juicebox[24] showed 24 chromosome-length interaction blocks with strong on-diagonal signal and minimal off-diagonal noise, consistent with correct scaffolding. Contact probability P(s) curves and insulation profiles computed with cooltools v0.5.4[29] were smooth and free of abrupt discontinuities, with local insulation minima coinciding with putative centromeric regions, indicating the absence of major structural artifacts.

Data availability

All raw sequencing reads (Illumina WGS, Oxford Nanopore, PacBio HiFi, and Hi-C), the final genome assembly (FASTA), and functional/structural annotations (GFF3/FASTA) are available under NCBI BioProject PRJNA1307247[48]. The assembly is deposited at GenBank under accession GCA_052674685.1[49].

Code availability

All software and versions are listed above. No custom code was used for this study.

Received: 25 September 2025; Accepted: 21 January 2026;
Published online: 26 January 2026

References

1. Li, W., Pu, Y. & Tian, H. Spatial and temporal distribution characteristics and optimum habitat conditions of *Paracobitis variegatus* in Heishui River. *Journal of Fishery Sciences of China* **30**, 515–524 (2023).
2. Mauice, K. Subspecific differentiation of *Paracobit variegatus* with comments on its zoogeography. *Zoological Research* **15**, 58–67 (1994).
3. Zhou, Y. Preliminary study on the biology of *Paramisgurnus rubripes* in the middle reaches of Qingyi River, Sichuan Agricultural University (2007).
4. Ma, B. S. *et al*. Length–weight and length–length relationships of four native fish species from the Yalong River, China. *Journal of Applied Ichthyology* **33**, 839–841 (2017).
5. Guo, Z. Sequencing of mitochondrial genome of *Paragonimus rubripes* and phylogenetic analysis of *Cyprinus carpio*, Shaanxi Normal University. (2012).
6. Liu, C. Z., Wei, G. H., Hu, J. H. & Liu, X. Y. Complete mitochondrial genome of *Paracobitis variegates* and its phylogenetic analysis. *Mitochondrial DNA Part A* **27**, 2421–2422 (2016).
7. Liu, F. *et al*. The telomere-to-telomere gapless genome of grass carp provides insights for genetic improvement. *GigaScience* **14**, giaf059 (2025).
8. Nurk, S. *et al*. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
9. Yuan, J. *et al*. A telomere-to-telomere genome assembly of koi carp (*Cyprinus carpio*) using long reads and Hi-C technology. *GigaScience* **14**, giaf087 (2025).
10. Zhang, X., Chen, J., Zhou, W., Wen, J. & Shi, Q. A telomere-to-telomere gap-free genome assembly of the protandrous hermaphrodite Asian seabass (*Lates calcarifer*). *Scientific data* **12**, 1457 (2025).
11. Shifu, C. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *Imeta* **2**, e107 (2023).
12. De Coster, W. & Rademakers, R. NanoPack2: population-scale evaluation of long-read sequencing data. *Bioinformatics* **39**, btad311 (2023).
13. Jung, Y. & Han, D. BWA-MEME: BWA-MEM emulated with a machine learning approach. *Bioinformatics* **38**, 2404–2413 (2022).
14. Servant, N. *et al*. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome biology* **16**, 259 (2015).
15. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell systems* **3**, 95–98 (2016).
16. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
17. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature communications* **11**, 1432 (2020).

18. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature methods* **18**, 170–175 (2021).
19. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
20. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
21. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).
22. Madden, T. The BLAST sequence analysis tool. *The NCBI handbook* **2**, 425–436 (2013).
23. Dudchenko, O. *et al*. *De novo* assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
24. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell systems* **3**, 99–101 (2016).
25. Zhou, C., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**, btac808 (2023).
26. 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**, giaa094 (2020).
27. Brown, M. R., de La Rosa, M. G. & Blaxter, M. Tidk: a toolkit to rapidly identify telomeric repeats from genomic datasets. *Bioinformatics* **41**, btaf049 (2025).
28. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* **27**, 573–580 (1999).
29. Open2CN Abdennur *et al*. Cooltools: enabling high-resolution Hi-C analysis in Python. *PLOS Computational Biology* 20: e1012067 (2024).
30. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**, 9451–9457 (2020).
31. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant physiology* **176**, 1410–1422 (2018).
32. Han, Y. & Wessler, S. R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic acids research* **38**, e199 (2010).
33. Abrusán, G., Grundmann, N., DeMester, L. & Makalowski, W. TEclass—a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* **25**, 1329–1330 (2009).
34. Jurka, J. *et al*. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research* **110**, 462–467 (2005).
35. Chen, N. Using Repeat Masker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics* **5**, 4.10. 11–14.10. 14 (2004).
36. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
37. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology* **37**, 907–915 (2019).
38. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology* **33**, 290–295 (2015).
39. Keilwagen, J., Hartung, F. & Grau, J. GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Gene prediction: Methods and protocols* Springer, 161–177 (2019).
40. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).
41. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology* **9**, R7 (2008).
42. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature methods* **12**, 59–60 (2015).
43. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
44. Cantalapiedra, C. P., Hernández-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Molecular biology and evolution* **38**, 5825–5829 (2021).
45. Aramaki, T. *et al*. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).
46. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic acids research* **49**, 9077–9096 (2021).
47. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
48. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP610030 (2025).
49. *GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_052674685.1 (2025).
50. Li, H. *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
51. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075 (2013).
52. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome biology* **21**, 245 (2020).
53. Tegenfeldt, F. *et al*. OrthoDB and BUSCO update: annotation of orthologs with wider sampling of genomes. *Nucleic acids research* **53**, D516–D522 (2025).
54. Parra, G. & Keith Bradnam, I. K. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067 (2007).

Acknowledgements

This study was funded by the Leshan Municipal Science and Technology Bureau Key Research Project (Grant No. 23NZD002) and by the Leshan Sub-center of the National Swine Industry Center.

Competing interests

The authors declare no competing interests.

Additional information

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