scientific data

Article in Press

Check for updates

OPEN

DATA DESCRIPTOR

Near telomere-to-telomere and gap-free reference genome assembly of the WuLi carp (*Cyprinus carpio* var. *Quanzhounensis*)

Chongqing Wang[1], Xiaowei Xu[1], Xu Huang[1], Xidan Xu[1], Kun Zhang[1], Yue Zhou[1], Jinhai Bai[1], Zhengkun Liu[1], Yuchen Jiang[1], Siyang Li[1], Enkui Hu[1], Wanjing Peng[1], Ling Xiong[1], Zeyang Li[1], Qian Xiao[1], Yuhan Yang[1] & Qinbo Qin[1,2,3] ✉

Cyprinus carpio* var. *Quanzhounensis **(WuLi carp), a native fish from Quanzhou, Guangxi, China, is well-adapted to mountainous rice-fish farming systems, yet genomic studies on the strain remain scarce. This study presents the first near telomere-to-telomere (T2T) genome and gapless assembly of *C. carpio* var. *Quanzhounensis*, generated using PacBio HiFi sequencing, ONT ultra-long reads, and Hi-C data. The final assembly, spanning 1.54 Gb across 68 contigs, features a contig N50 of 29.52 Mb, with 99.37% of sequences anchored to 50 chromosomes. A total of 49,781 protein-coding genes were annotated, and transposable elements account for 42.96% of the genome. The high-quality assembly provides a valuable genomic resource for conservation, germplasm evaluation, and future comparative studies of WuLi carp and related cyprinid fishes.**

Background & Summary

Cyprinus carpio var. *Quanzhounensis*, commonly known as WuLi carp or Hehua carp, is an indigenous freshwater species native to Quanzhou, Guangxi, China[1]. With a long history in aquaculture, this strain thrives in mountainous paddy field environments where it feeds on rice flowers, contributing to its unique ecological niche. The WuLi carp is recognized for its robust physique, dark purplish-brown coloration, and rapid growth[2]. It is particularly valued for its delicate, high-protein meat, minimal intermuscular bones, and lack of muddy odor, making it a prized species both economically and culturally[3]. However, the industry faces significant challenges, including genetic degradation, inconsistent fry quality, frequent disease outbreaks, and stunted growth[4]. These issues are largely attributed to widespread inbreeding and the absence of systematic breeding management.

Existing research on WuLi carp has largely centered on paddy-field aquaculture, nutritional quality, stress tolerance, and selective breeding practices[5–7]. These studies have provided useful biological and production-level information, but genomic resources for this local carp strain are still limited. The lack of a strain-specific, chromosome-scale reference genome has restricted reliable analysis of structural variation, repetitive regions, gene copy differences, and molecular markers in WuLi carp.

Although chromosome-level and complete genome assemblies are already available for common carp, the common carp reference cannot fully represent the genomic features of a locally cultured strain. WuLi carp has been maintained under long-term paddy-field farming and local breeding conditions, and a strain-matched assembly is therefore needed to improve read mapping, recover strain-specific sequences, characterize telomeric and repetitive regions, and provide a more accurate reference for future comparative and population genomic studies.

To improve genomic resources for this locally important carp strain, we generated a chromosome-level near telomere-to-telomere assembly of WuLi carp by integrating PacBio HiFi reads, ONT ultra-long reads, and Hi-C

[1]Engineering Research Center of Polyploid Fish Reproduction and Breeding of the State Education, Ministry, College of Life Sciences, Hunan Normal University, Changsha, 410081, China. [2]Yuelushan Laboratory, Changsha, 410128, China. [3]Hunan Research Center of the Basic Discipline Developmental Biology (2025JC2006), College of Life Sciences, Hunan Normal University, Changsha, 410081, China. ✉e-mail: qqb@hunnu.edu.cn

Sequencing technology	Bases (Gb)	Reads number	Mean read length (bp)	GC Content (%)
HiFi	127.16	5,290,873	24,033.96	36.91
ONT	38.60	665,624	57,991	37.20
Hi-C	147.63	984,252,510	150	38.99
RNA-seq	7.41	24,786,496	150	48.93
Genome survey short reads	162.56	1,083,727,670	150	37.15

Table 1. Summary of sequencing data generated for the WuLi carp genome assembly.

data. This assembly provides a strain-specific reference for future studies of genetic diversity, germplasm conservation, and comparative genomics in paddy field carp.

Methods

Sample collection. A female WuLi carp (*Cyprinus carpio* var. *Quanzhounensis*), one year old, was sourced from the Engineering Research Center of Polyploid Fish Reproduction and Breeding, located in Changsha, Hunan Province, China. Tissue samples, including those from the muscle, eye, brain, liver, heart, and kidney, were collected for genomic and transcriptomic analysis. These samples were immediately snap-frozen in liquid nitrogen and stored at -80 °C. The experimental procedures followed the ethical guidelines set by Hunan Normal University (Approval No. 380; 2025). All methods are reported in accordance with ARRIVE guidelines (https://arriveguidelines.org).

Library construction and sequencing. Libraries were constructed for sequencing across multiple platforms. Genomic DNA extracted from muscle tissue served as the starting material, using the DNeasy Blood & Tissue Kit from Qiagen according to the standard protocol for animal tissues.

For PacBio SMRT sequencing, libraries were prepared following the manufacturer's guidelines and sequenced on a Sequel II platform in Circular Consensus Sequencing mode. The raw data were processed to generate high-fidelity HiFi reads, with low-quality sequences and adapters removed.

For ONT sequencing, an ultra-long DNA library was prepared and sequenced using a PromethION flow cell (Oxford Nanopore Technologies, UK). Base calling was performed using ONT's software, and only reads with a quality value (QV) $\geq$7 were retained. Adapter sequences were removed using Porechop (v0.2.4, https://github.com/rrwick/Porechop), and reads shorter than 30 kb or with an average quality score below 90% were discarded using Filtlong (v0.2.1, https://github.com/rrwick/Filtlong).

Hi-C libraries were generated from cross-linked muscle tissue using a standard protocol involving formaldehyde fixation, restriction enzyme digestion, biotin labeling, and ligation. After cross-link reversal and purification, DNA was sheared to approximately 300 bp for paired-end library construction, followed by sequencing on the DNBSEQ platform.

For RNA sequencing, total RNA was extracted from eye, brain, liver, heart, kidney, and muscle tissues using TRIzol reagent (Invitrogen). Paired-end libraries were constructed and sequenced on the DNBSEQ platform.

The sequencing produced a total of 127.16 Gb of HiFi data, 38.60 Gb of ONT data, 147.63 Gb of Hi-C data, 7.41 Gb of RNA-seq data, and 162.56 Gb genome survey short reads (Table 1).

Genome survey and assembly. For the genome survey, DNA libraries with insert sizes ranging from 300 to 400 bp were prepared. Following library preparation, the DNA was purified, quantified, and sequenced using paired-end reads on the DNBSEQ platform. Raw reads underwent quality filtering using Fastp (v0.23.4; parameters: -l 50–dedup–dup_calc_accuracy 3)[8]. K-mer frequency analysis was conducted using GCE (v1.0.0; parameters: -k 17)[9], and the genome size was estimated based on K-mer distribution with GenomeScope 2.0 (v2.0; parameters: -k 17 -p 2)[10]. GenomeScope predicted a genome size of 1.506 Gb, with a 17-mer peak depth of 91 (Fig. 1). Heterozygosity and repeat content were calculated at 0.88% and 58.38%, respectively. Ploidy analysis, performed with Smudgeplot (v0.2.3dev; parameters: -k21 -m100 -ci1 -cs1000)[11], revealed an AB-type ploidy, suggesting diploid-like k-mer behavior despite the known tetraploid ancestry of common carp (Table 2)[12].

The initial draft genome assembly was created by integrating PacBio HiFi reads with Oxford Nanopore ultra-long reads and Hi-C data. Assembly was performed using HiFiasm (v0.25.0-r726; parameters: default)[13], followed by redundancy removal with Purge Haplotigs (v1.1.3; parameters: default)[14]. Chromosome-level scaffolding was achieved using Hi-C data.

Hi-C reads were aligned to the draft genome[15], and chromosome scaffolding was completed using the 3D-DNA pipeline, which included steps like splitting, anchoring, sorting, orienting, and merging contigs or scaffolds[16]. An interaction matrix was generated using Juicer (v1.6; parameters: default)[17] and manually refined with Juicebox (v1.11.08; parameters: default)[18].

For further refinement, ultra-long ONT reads were aligned to the chromosome scaffolds using minimap2 (v2.24; parameters: ont: -ax map-ont ccs: -ax map-hifi)[19] to generate consensus sequences. These sequences were aligned with telomere regions using blastn (v2.11.0+; parameters: -outfmt 7), and sequences with $\geq$90% coverage replaced telomeric regions based on their alignment. Gaps between contigs were filled using TGS-GapCloser (v1.2.0; parameters:–min_nread 10)[20], leveraging ultra-long ONT reads for gap extension. Additional polishing was done with Pilon (v1.23; parameters:–fix all -changes)[21] to correct errors in the genome, resulting in a near telomere-to-telomere assembly of WuLi carp. Telomeric repeats were identified using an in-house script developed by Wuhan OneMore Technology Co., Ltd. The script scanned the whole genome for the motifs "TTAGGG"

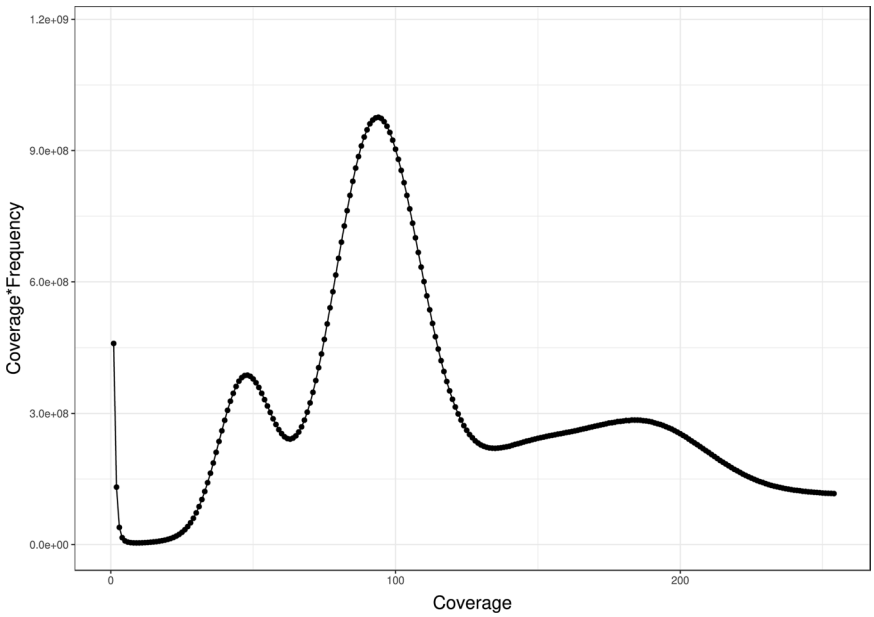


Fig. 1 The K-mer frequency distribution for the WuLi carp genome. The x-axis indicates the coverage of the K-mer, the y-axis represents the k-mer frequency for a given depth.

Peak	Kmers [number]	Kmers [proportion]	Summit B / (A + B)	Summit A + B
AB	56,377,703	0.49	0.49	86.75
AABB	41,147,766	0.36	0.49	182.38
AAB	16,639,250	0.15	0.32	134.56

Table 2. Smudgeplot-derived statistics used in assessing the ploidy level.

and "AATCCC". For each chromosome, the terminal 50 kb at both ends were examined, and a region containing at least four tandem copies of the repeat was counted as a telomeric signal. The number of chromosome ends meeting these criteria was then recorded.

The final assembly spans 1.54 Gb across 50 chromosomes, with a scaffold N50 of 30.35 Mb, a contig N50 of 29.52 Mb, and an anchoring rate (the proportion of the total assembled sequence length that was assigned to chromosome-scale scaffolds using Hi-C data) of 99.37% (Table 3). High-quality chromosome assignments were validated by the Hi-C interaction heatmap (Fig. 2). A total of 34 chromosomes showed telomeric repeats at both ends by searching for the telomeric repeat motif across the genome (Table 4). These telomeric sequences were annotated and visualized along with their genomic positions (Fig. 3). These results demonstrate that the WuLi carp genome achieves contiguity and completeness comparable to the other two carp genomes[22,23].

Repeats annotation. Repetitive elements were identified using a combination of de novo and homology-based approaches. Tandem repeats were detected with TRF (v4.09; default)[24]. For homology-based identification, RepeatMasker (v4.0.9; default)[25] was applied, utilizing the RepBase library (http://www.girin-st.org/repbase). To construct a denovo repeat library for WuLi carp, RepeatModeler (open-1.0.11; parameters: default)[26] was used, followed by additional de novo predictions. Long terminal repeats (LTRs) were identified using LTR_FINDER_parallel (v1.0.7; parameters: default)[27].

The results from TRF, RepeatMasker, RepeatProteinMask, and de novo predictions were integrated, with redundant elements removed. This analysis revealed that DNA transposons and transposable elements (TEs) constitute approximately 24.87% and 42.96% of the WuLi carp genome, respectively (Fig. 4). More specifically, repetitive DNAs, long interspersed nuclear elements (LINEs), short interspersed nuclear elements (SINEs), and long terminal repeats (LTRs) accounted for 24.87%, 5.38%, 0.49%, and 6.50% of the genome, respectively (Table 5). The repeat proportion observed in WuLi carp is broadly comparable with values reported for other common carp, although exact estimates vary among species and studies because of differences in genome assembly continuity, repeat library construction, and annotation pipelines[22,23].

Gene prediction and functional annotation. Gene annotation for the WuLi carp genome was performed using both structural and functional annotation methods. Structural annotation focused on identifying gene locations and structures through homology-based, *de novo*, and RNA-sequencing-assisted approaches, while functional annotation aimed to characterize the biological roles and metabolic pathways associated with the predicted gene products.

Genome assembly statistics	WuLi carp	Common carp	*Cyprinus carpio var. Xiangxi*
Total length (Gb)	1.54	1.6	1.54
Assembly level	near T2T	chromosome	chromosome
Number of scaffolds	50	50	50
N50 length (scaffold) (Mb)	30.35	30.40	30.26
Number of contigs	68	/	216
N50 length (contig) (Mb)	29.52	30.4	28.12
Number of chromosomes	50	50	50
Anchoring rate (%)	99.37	/	99.38
GC Content (%)	37.20	37	/
BUSCO (%)	99.37	/	97.40
Gene prediction	49,781	/	46,362

Table 3. Comparison of assembly statistics among WuLi carp, common carp (ASM5088196v1) and *Cyprinus carpio var. Xiangxi* genomes.

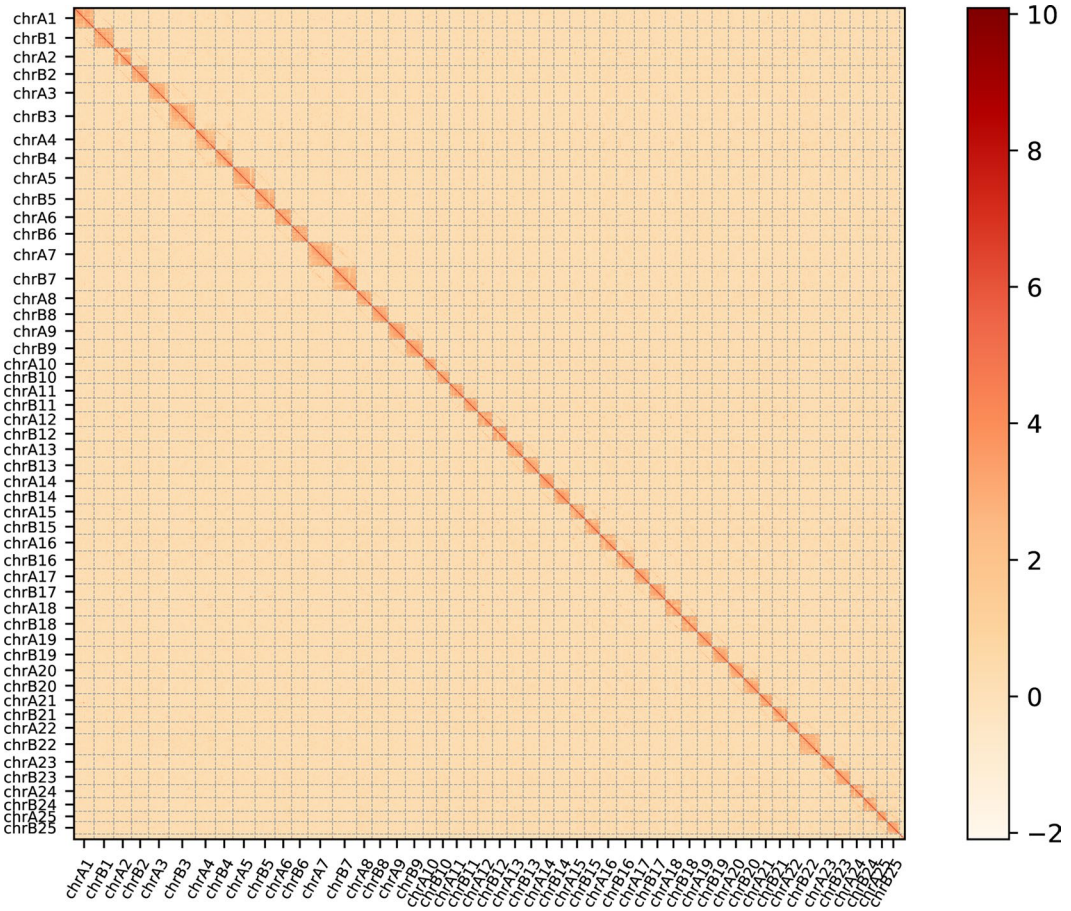


Fig. 2 Hi-C interaction heatmap of the WuLi carp genome. Genomic coordinates are arranged along both axes, with interaction frequency represented by color gradient from yellow (weak) to red (strong). Each of the 50 distinct squares corresponds to an individual chromosome.

For structural annotation, three strategies were combined: homology-based prediction, *de novo* prediction, and RNA-sequencing-assisted prediction. Homology-based predictions were conducted using miniprot (v0.11-r234; parameters: --gff-only -O 11 -E 1 -F 23 -C 1 -B 5 -G 200000 -j 1)[28] and Liftoff (v1.6.3; parameters: -a 0.5 -s 0.5)[29] to align the WuLi carp genome sequence with protein sequences from closely related species, including *Cyprinus carpio, Carassius auratus*, *Carassius gibelio*, *Carassius auratus* red var., and autotetraploid *Carassius auratus*. De novo predictions were performed using AUGUSTUS (v3.3.2; default parameters)[30] and Genscan (v1.0; default parameters)[31]. RNA-seq data were mapped to the WuLi carp genome using HISAT2 (v2.1.0; parameters:–dta)[32], and transcripts were assembled and quantified with StringTie (v1.3.5; default parameters)[33]. Protein-coding genes were then predicted using TransDecoder (v5.5.0; default parameters). The predictions

Name	Length (Mb)	Number of gaps	Number of telomeres	Name	Length (Mb)	Number of gaps	Number of telomeres
chrA1	37,604,534	0	2	chrB1	36,941,188	0	1
chrA2	33,074,279	0	2	chrB2	31,885,149	0	2
chrA3	37,976,603	0	2	chrB3	49,352,945	0	2
chrA4	37,505,275	0	2	chrB4	32,953,188	0	2
chrA5	41,300,498	0	2	chrB5	37,399,261	0	1
chrA6	30,909,454	0	0	chrB6	30,953,430	0	1
chrA7	45,879,859	0	2	chrB7	45,251,876	0	2
chrA8	28,577,267	0	2	chrB8	30,823,843	0	2
chrA9	32,250,276	0	2	chrB9	33,382,518	0	2
chrA10	24,876,355	0	2	chrB10	24,242,493	0	2
chrA11	27,257,556	0	1	chrB11	25,769,430	0	2
chrA12	27,511,735	0	2	chrB12	27,656,847	0	2
chrA13	30,037,027	0	2	chrB13	30,136,999	0	1
chrA14	27,586,354	0	2	chrB14	29,340,815	0	1
chrA15	27,930,709	0	1	chrB15	28,680,639	0	2
chrA16	31,131,537	0	1	chrB16	33,445,123	0	0
chrA17	28,180,574	0	2	chrB17	29,647,648	0	2
chrA18	30,365,939	0	2	chrB18	29,665,496	0	2
chrA19	27,306,516	0	2	chrB19	30,496,115	0	2
chrA20	28,676,192	0	2	chrB20	29,190,326	0	1
chrA21	24,983,710	0	2	chrB21	28,317,297	0	2
chrA22	22,467,107	0	2	chrB22	39,315,013	0	1
chrA23	26,901,886	0	2	chrB23	28,942,893	0	2
chrA24	24,775,422	0	1	chrB24	25,273,478	0	1
chrA25	19,467,193	0	1	chrB25	23,407,569	0	0

Table 4. Chromosome assembly statistics.

from these methods were integrated into a high-quality, non-redundant gene set using MAKER2 (v2.31.10; default parameters)[34].

For functional annotation, protein sequences derived from structural annotation were compared against several databases, including GO[35], KEGG[36], Swiss-Prot[37], TrEMBL[37], NR[38], KOG (https://ftp.ncbi.nih.gov/pub/COG/KOG/) and AnimalTFDB[39]. This comparison was conducted using DIAMOND (v2.1.11.165; parameters:–evalue 1e-05–max-hsps 1)[40], providing insights into protein functions, metabolic pathways, and additional characteristics. To further identify conserved sequences, motifs, and structural domains, the Pfam[41] and InterPro[42] databases were analyzed using InterProScan (v5.74-105.0; parameters:–seqtype p–formats TSV–goterms–pathways -dp)[43]. Pathway annotation was performed using KOBAS (v3.0; parameters: -t blastout:tab -s ko)[44] against the KEGG database.

A total of 49,781 protein-coding genes were predicted, with an average gene length of 14,682.63 bp, coding sequence (CDS) length of 1,660.25 bp, 9.24 exons per gene, and an average exon length of 240.35 bp (Table 6). To further evaluate the evidence supporting the final gene models, the 49,781 predicted protein-coding genes were compared with gene models derived from de novo prediction, homology-based prediction, and RNA-seq-derived transcript evidence using CDS overlap thresholds of ≥20%, ≥50%, and ≥80%. It should be noted that this analysis evaluates overlap with assembled transcript sequences, not direct read-level mapping support. Although RNA-seq-only prediction identified 24,191 genes, this value represents genes recovered solely from transcript-based prediction and does not reflect the full extent of transcriptomic support for the final integrated gene set. At the ≥50% CDS overlap threshold, 31,093 genes (62.46%) were supported by RNA-seq-derived transcript evidence. Under the more stringent ≥80% threshold, 29,439 genes (59.14%) still showed RNA-seq-derived support. In addition, 22,689 genes (45.58%) were supported simultaneously by *de novo*, homology-based, and RNA-seq-derived evidence at the ≥80% threshold (Table 7). These results indicate that the integrated gene set is supported by multiple lines of evidence, while genes without RNA-seq support may represent lowly expressed, tissue-specific, developmental-stage-specific, or condition-specific genes not captured by the six tissues sampled in this study. Of the predicted genes, 98.07% (48,819 genes) and 98.22% (61,387 mRNA transcripts) were successfully annotated (Fig. 5 and Table 8).

Non-coding RNAs were predicted using BLASTN (v2.11.0+; parameters: -evalue 1e-5)[45] for rRNAs, tRNAscan-SE (v1.3.1; default parameters)[46] for tRNAs, and Infernal (v1.3.3) against Rfam (v14.8; parameters: cmscan–rfam–nohmmonly)[47] for miRNAs and snRNAs. A total of 4,196 miRNAs, 20,088 tRNAs, 8,226 rRNAs, and 3,853 snRNAs were identified (Table 9). These numbers are generally within the range reported for closely related cyprinid genomes, although variation among species may be influenced by genome continuity, annotation methods, and the stringency of non-coding RNA prediction[22,23].

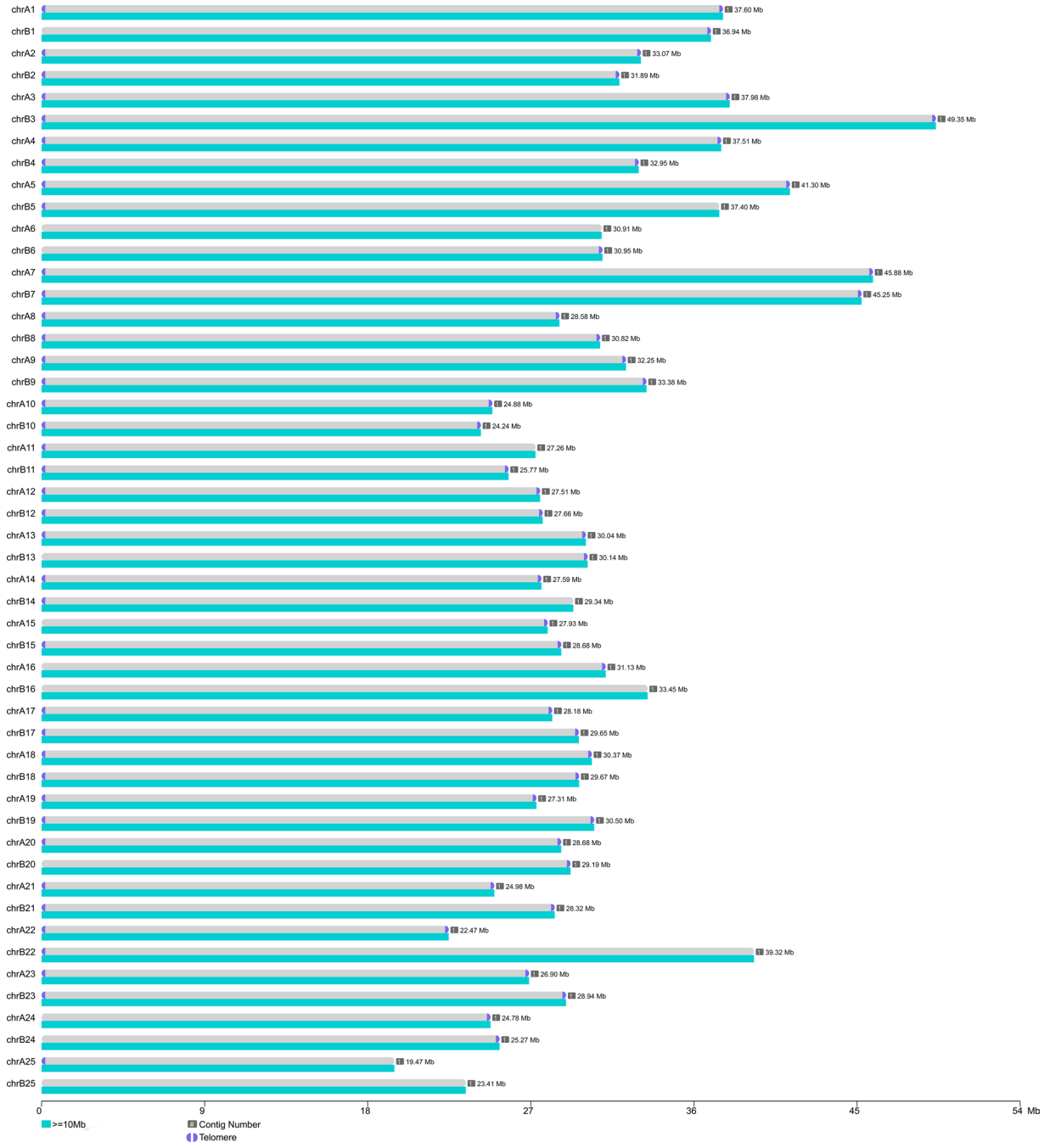


Fig. 3 Overview of the near T2T gapless reference genome of WuLi carp. The plot summarizes chromosome-scale scaffold lengths and telomeric repeat distribution across the 50 chromosomes. Chromosomes with telomeric signals detected at both ends are distinguished from chromosomes with one or no terminal telomeric signal. The internal figure legend defines the colors and symbols used for chromosomes, telomeric repeats, and chromosome-end categories.

Data Records

All raw sequencing reads generated in this study, including PacBio HiFi, ONT ultra-long, Hi-C, genome survey, and RNA-seq data, have been deposited in the NCBI Sequence Read Archive under accession number SRP38622118[48]. The complete near telomere-to-telomere WuLi carp genome assembly is deposited in GenBank under the accession number GCA_060414195.1[49]. Annotated coding and protein sequences are accessible on Figshare via https://doi.org/10.6084/m9.figshare.30814787[50].

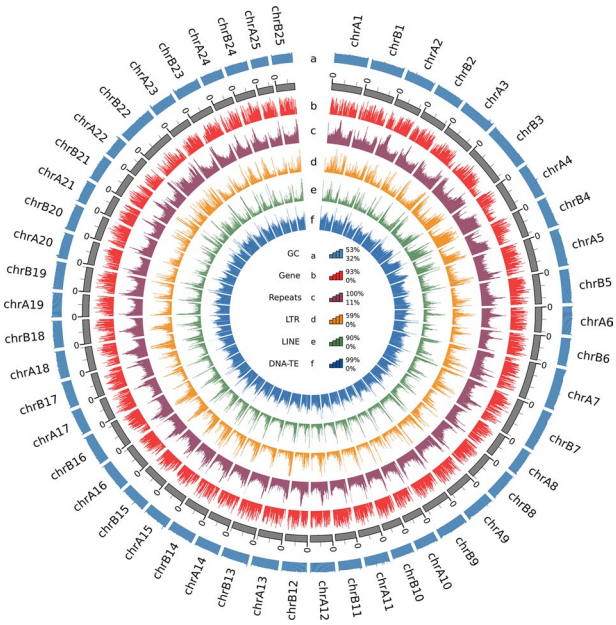


Fig. 4 Genomic landscape across the chromosome-level assembly of the WuLi carp. All features were calculated using a sliding window of approximately 500 kb. From outer to inner rings, the Circos plot displays (a) GC content (range: 32%-53%); (b) gene density (range: 0%-93%); (c) repeat density (range: 11%-100%); (d) LTR retroelement density (range: 0%-59%); (e) LINE density (range: 0%-90%); and (f) DNA transposon density (range: 0%-99%).

Type	RepBase TEs		TE Proteins		Denovo		Combined TEs	
	Length (bp)	% in Genome	Length (bp)	% in Genome	Length (bp)	% in Genome	Length (bp)	% in Genome
DNA	254,462,515	16.45	24,688,237	1.60	234,408,038	15.15	384,782,728	24.87
LINE	66,395,980	4.29	44,964,222	2.91	49,885,955	3.22	83,216,536	5.38
SINE	6,177,090	0.40	0	0.00	2,529,149	0.16	7,606,091	0.49
LTR	59,453,537	3.84	31,825,247	2.06	69,296,928	4.48	100,599,316	6.50
Satellite	53,367,520	3.45	0	0.00	5,276,428	0.34	57,652,645	3.73
Simple_repeat	0	0.00	0	0.00	525,816	0.03	525,816	0.03
Other	44,785	0.00	0	0.00	0	0.00	44,785	0.00
Unknown	4,120,453	0.27	6,216	0.00	71,275,857	4.61	73,825,681	4.77
Total	417,733,920	27.00	101,442,943	6.56	426,505,867	27.57	664,544,275	42.96

Table 5. Analysis of categorized repeat sequences in the WuLi carp genome.

Gene set	Number	Average gene length (bp)	Average CDS length (bp)	Average exon per gene	Average exon length (bp)	Average intron length (bp)
de novo/Genscan	50,726	19,576.36	1,599.95	8.59	186.17	2,367.25
de novo/AUGUSTUS	42,565	13,269.38	1,505.53	8.73	172.45	1,521.79
homo/*C. carpio*	91,521	21,078.00	1,957.62	11.21	174.64	1,872.78
homo/*C. auratus*	98,713	22,413.25	1,955.58	11.42	171.23	1,963.14
homo/*C.auratus* red var.	126,303	20,644.88	2,048.51	11.56	177.26	1,761.65
homo/*C. gibelio*	123,796	21,401.33	2,052.49	11.25	182.48	1,888.12
homo/Autotetraploid *C. auratus*	183,050	10,419.14	1,481.79	6.32	234.49	1,680.23
trans.orf/RNA-seq	24,191	17,872.01	1,767.70	11.37	260.63	1,437.20
BUSCO	6,049	11,879.97	1,643.02	10.20	161.12	1,112.99
MAKER	49,382	15,348.32	1,540.05	9.43	241.90	1,549.95
HiFAP	49,781	14,682.63	1,660.25	9.24	240.35	1,512.05

Table 6. Predicted protein-coding gene statistics in the WuLi carp genome.

Evidence category	≥20% overlap No.	≥20% overlap Ratio (%)	≥50% overlap No.	≥50% overlap Ratio (%)	≥80% overlap No.	≥80% overlap Ratio (%)
P(single)	0	0.00	2	0.00	150	0.30
P(more)	0	0.00	1	0.00	46	0.09
H(single)	3,554	7.14	4,355	8.75	5,442	10.93
H(more)	2,009	4.04	2,793	5.61	4,722	9.49
C(single)	29	0.06	42	0.08	95	0.19
C(more)	0	0.00	0	0.00	0	0.00
PH	11,335	22.77	11,537	23.18	9,981	20.05
PC	50	0.10	48	0.10	60	0.12
HC	1,811	3.64	2,545	5.11	6,595	13.25
PHC	30,993	62.26	28,458	57.17	22,689	45.58

Table 7. Evidence support statistics for the final WuLi carp gene set. Note: P represents evidence from de novo gene prediction, H represents evidence from homology-based prediction, and C represents RNA-seq-derived transcript evidence. "single" indicates support from one prediction source within the corresponding evidence category, whereas "more" indicates support from multiple prediction sources within the same evidence category. PH, PC, HC, and PHC indicate gene models supported by combined evidence categories. Overlap refers to the proportion of CDS overlap between the final gene model and the corresponding supporting evidence.

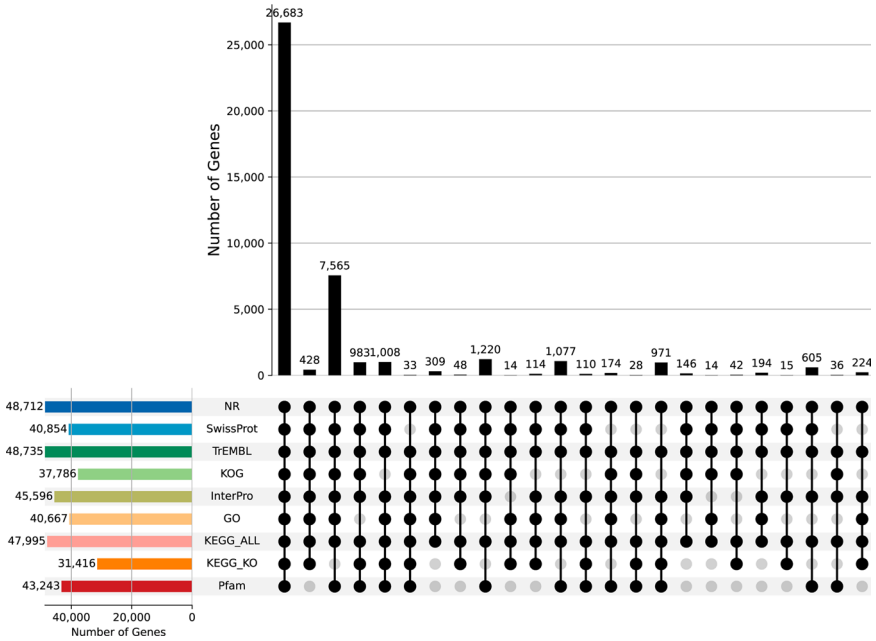


Fig. 5 An UpSet plot depicting the gene functional annotations across nine databases: NR, Swiss-Prot, TrEMBL, KOG, InterPro, GO, KEGG-ALL, KEGG-KO, and Pfam.

Technical Validation

Genome assembly and gene annotation quality assessment. To assess taxonomic consistency and possible foreign sequence contamination in the genome assembly, the assembled genome was split into 10-kb fragments and aligned against the NCBI nucleotide database (NT) using BLASTN (v2.11.0+; parameters: -evalue 0.00001 -max_hsps 1)[45]. The five genera with the largest numbers of matched fragments are listed in Table 10. The most frequently matched genus was *Cyprinus*, accounting for 43.60% of the split genome fragments and 43.67% of all matched fragments, with a median sequence identity of 98.15%. Other major matches were assigned to *Carassius* (17.76%), *Danio* (12.18%), *Onychostoma* (6.46%), and *Sinocyclocheilus* (3.81%). The overall proportion of genome fragments with NT matches was 99.80%. These results showed that most matched fragments were assigned to fish-related taxa, especially cyprinid fishes, and no high-proportion non-vertebrate contaminant signal was detected in the final assembly. The completeness of the genome assembly and gene annotation was assessed using BUSCO (v5.7.1; default parameters)[51], aligned against the actinopterygii_odb10 lineage. The genome assembly recovered 99.37% of BUSCOs, with 30.80% in single-copy, 68.57% in duplicated, 0.33% fragmented, and 0.30% missing (Table 11). The gene annotations recovered 99.21% of BUSCOs, with 26.79% in single-copy, 72.42% in duplicated, 0.19% fragmented, and 0.60% missing (Table 11).

Type	Gene		mRNA	
	Number	Percent (%)	Number	Percent (%)
Total	49,781	100.00	62,499	100.00
Annotated	48,819	98.07	61,387	98.22
NR	48,712	97.85	61,263	98.02
Swiss-Prot	40,854	82.07	52,364	83.78
TrEMBL	48,735	97.90	61,288	98.06
KOG	37,786	75.90	48,471	77.55
TF	13,289	26.69	16,415	26.26
InterPro	45,596	91.59	57,532	92.05
GO	40,667	81.69	51,535	82.46
KEGG_ALL	47,995	96.41	60,502	96.80
KEGG_KO	31,416	63.11	40,576	64.92
Pfam	43,243	86.87	54,392	87.03
Unannotated	962	1.93	1,112	1.78

Table 8. Functional annotation of the WuLi carp genome's protein-coding genes. Note: Functional annotation leveraged nine databases (Nr, Swiss-Prot, TrEMBL, KOG, TF, InterPro, GO, KEGG, Pfam).

Type	Copy	Average length (bp)	Total length (bp)	Percentage of genome (%)
miRNA	4,196	89	374,925	0.024236
tRNA	20,088	76	1,522,359	0.098407
rRNA	8,226	121	992,476	0.064155
snRNA	3,853	144	553,397	0.035772

Table 9. Statistics of non-coding RNAs in the WuLi carp genome.

Genus	No. of hits to this genus	Total genome fragments	No. of fragments with NT hits	Percentage of genome matching this genus (%)	Percentage of hits to this genus (%)	Median identity (%)
Cyprinus	67,533	154,892	154,627	43.60	43.67	98.15
Carassius	27,508	154,892	154,627	17.76	17.79	89.82
Danio	18,867	154,892	154,627	12.18	12.20	85.24
Onychostoma	9,999	154,892	154,627	6.46	6.47	87.88
Sinocyclocheilus	5,895	154,892	154,627	3.81	3.81	93.36

Table 10. Taxonomic assignment of genome fragments for contamination assessment based on NT database alignment.

BUSCO genome completeness score	Genome	Annotation
Complete BUSCOs	3,617 (99.37%)	3,611 (99.21%)
Complete Single-Copy BUSCOs	1,121 (30.80%)	975 (26.79%)
Complete Duplicated BUSCOs	2,496 (68.57%)	2,636 (72.42%)
Fragmented BUSCOs	12 (0.33%)	7 (0.19%)
Missing BUSCOs	11 (0.30%)	22 (0.60%)
Total BUSCO groups searched	3,640 (100%)	3,640 (100%)

Table 11. BUSCO completeness and accuracy evaluation of the WuLi carp genome and annotations.

ONT and PacBio HiFi reads were aligned to the assembly using minimap2 (v2.24; parameters: ont: -ax map-ont, ccs: -ax map-hifi)[19], resulting in 99.99% mapping and 99.81% coverage for ONT reads, and 99.96% mapping with 99.97% coverage for PacBio HiFi reads. Short reads were aligned using bwa (v0.7.12-r1039; parameters: sort -m 2G)[52,53], achieving 99.65% mapping and 99.85% coverage.

A telomere-to-telomere assembly was achieved for 34 chromosomes, while telomeres were identified at only one end, or were not detected at either end, for the remaining chromosomes. The inability to achieve full T2T status for these chromosomes is likely due to the presence of complex, highly repetitive genomic regions (Fig. 4). Despite the use of ultra-long ONT reads, the structural complexity of these regions remains challenging to resolve completely. It is expected that future improvements in sequencing technology, particularly longer read lengths, and advancements in T2T assembly algorithms will address these challenges and allow for the full assembly of the WuLi carp genome.

Data availability

All raw sequencing reads generated in this study, including PacBio HiFi, ONT ultra-long, Hi-C, genome survey, and RNA-seq data, have been deposited in the NCBI Sequence Read Archive under accession number SRP38622118. The complete near telomere-to-telomere WuLi carp genome assembly is deposited in GenBank under the accession number GCA_060414195.1. Annotated coding and protein sequences are accessible on Figshare via https://doi.org/10.6084/m9.figshare.30814787.

Code availability

No custom code was created for this research. All genome assembly, annotation, and validation were conducted using publicly accessible bioinformatics tools, following standard protocols and default settings, as detailed in the Methods section.

Received: 16 December 2025; Accepted: 29 September 2026;
Published: xx xx xxxx

References

1. Liang, Z. *et al*. Genetic origin and differentiation of ten paddy field-farmed *Cyprinus carpio* strains in China. *Aquaculture* **561**, 738573 (2022).
2. Xu, Y. *et al*. Estimation of genetic parameters for the main growth traits of Quanzhou Hehua Carp. Guangxi. *Sci*. **29**, 801–808 (2022).
3. Zhou, K. *et al*. Analysis of the relationship between morphological traits and body mass of Guangxi Hehua Carp. *Henan Agric. Sci* **49**, 159–165 (2020).
4. Wen, J., Teng, Q. & Shen, J. Reproduction and cultivation techniques of Hehua carp. Guide Fish. *Fortune* **11**, 37–41 (2023).
5. Zhou, A. *et al*. Analysis on the aquaculture technology and benefits of Lian Shan paddy field Hehua carp. Sci. Fish. *Farming* **2**, 82 (2021).
6. Wang, T. *et al*. Analysis of nutritional components and safety evaluation of Hehua. *carp muscle. J. South. Agric.* **50**, 1579–1586 (2019).
7. Fan, J. *et al*. Effects of morphological traits on body mass of the F5 generation of Hehua carp. *Guangdong Agric. Sci* **48**, 124–130 (2021).
8. Chen, S. *et al*. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
9. Liu, B. *et al*. Estimation of genomic characteristics by analyzing k-mer frequency in de novo genome projects. *Quant. Biol.* **35**, 62–67 (2013).
10. Vurture, G. W. *et al*. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204 (2017).
11. 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).
12. Xu, P. *et al*. The allotetraploid origin and asymmetrical genome evolution of the common carp *Cyprinus carpio*. *Nat. Commun.* **10**(1), 4625 (2019).
13. Cheng, H. *et al*. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
14. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 460 (2018).
15. Burton, J. N. *et al*. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat. Biotechnol.* **31**, 1119–1125 (2013).
16. Kajitani, R. *et al*. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res* **24**, 1384–1395 (2014).
17. Durand, N. C. *et al*. Juicer Provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst* **3**, 95–98 (2016).
18. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst* **3**, 99–101 (2016).
19. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
20. 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).
21. Walker, B. J. *et al*. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
22. Li, Z. *et al*. A high-quality chromosome-level genome assembly of the common carp (*Cyprinus carpio* var. *Xiangxi*). *Sci Data* **12**, 1639 (2025).
23. 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).
24. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573–580 (1999).
25. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinformatics* 25, 4.10.1-4.10.14 (2009).
26. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics* **21**, i351–i358 (2005).
27. Ou, S. & Jiang, N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mob. DNA* **10**, 48 (2019).
28. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
29. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2021).
30. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34**, W435–W439 (2006).
31. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
32. Kim, D. *et al*. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**(8), 907–915 (2019).
33. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
34. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
35. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
36. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27–30 (2000).
37. Coudert, E. *et al*. Annotation of biologically relevant ligands in UniProtKB using ChEBI. *Bioinformatics* **39**, btac793 (2023).
38. Sayers, E. W. *et al*. Database resources of the national center for biotechnology information. *Nucleic Acids Res* **50**, D20–D26 (2022).
39. Shen, W. *et al*. AnimalTFDB 4.0: a comprehensive animal transcription factor database updated with variation and expression annotations. *Nucleic Acids Res* **51**, D39–D45 (2023).
40. Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).

41. Bateman, A. *et al*. The Pfam protein families database. *Nucleic Acids Res* **32**, D138–D141 (2004).
42. Paysan-Lafosse, T. *et al*. InterPro in 2022. *Nucleic Acids Res* **51**, D418–D427 (2023).
43. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
44. Bu, D. *et al*. KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acids Res* **49**, W317–W325 (2021).
45. Altschul, S. F. *et al*. Basic local alignment search tool. *J. Mol. Biol* **215**, 403–410 (1990).
46. 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).
47. Griffiths-Jones, S. *et al*. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* **33**, D121–D124 (2005).
48. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP713457 (2026).
49. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_060414195.1 (2026).
50. Wang, C. Q. Near telomere-to-telomere and gap-free reference genome assembly of the WuLi carp (*Cyprinus carpio* var. *Quanzhounensis*). *Figshare* https://doi.org/10.6084/m9.figshare.30814787 (2026).
51. Manni, M. *et al*. BUSCO update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Mol. Biol. Evol.* **38**, 4647–4654 (2021).
52. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

Author contributions

Chongqing Wang conducted the writing and data analysis. Xiaowei Xu, Xu Huang, Xidan Xu, Kun Zhang, Yue Zhou, Jinhai Bai, Zhengkun Liu and Yuchen Jiang participated in data analysis. Siyang Li, Enkui Hu, Wanjing Peng, Ling Xiong, Zeyang Li, Qian Xiao, Yuhan Yang and Qinbo Qin provided guidance on experimental design and manuscript revision. Qinbo Qin also provided project support.

Funding

This work was financially supported by Yuelushan Laboratory Breeding Program (YLS-2025-ZY02043), Hunan Province Major Basic Research Projects (2026JC0008), the Earmarked Fund for HARS (HARS-07), the Aid Program for Science and Technology Innovative Research Team in Higher Educational Institutions of Hunan Province, and the Key Research and Development Program of Hunan Province (2024JK2034, 2025QY2004).

Competing interests

The authors declare no competing interests.

Additional information

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

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 4.0 International License, which permits use, sharing, adaptation, 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 changes were made. 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/4.0/.

© The Author(s) 2026

This Article in Press is shared early to give you faster access to new research. It is citable and carries a permanent DOI. The final edited version will replace it automatically.