Scientific Data

https://doi.org/10.1038/s41597-026-06656-y

Article in Press

Chromosome-level genome assembly of the longfin barb (*Acrossocheilus longipinnis*)

Received: 17 September 2025

Accepted: 19 January 2026

Cite this article as: E, Z., Xiong, F., Zhu, Y. *et al.* Chromosome-level genome assembly of the longfin barb (*Acrossocheilus longipinnis*). *Sci Data* (2026). https://doi.org/10.1038/s41597-026-06656-y

Zechen E, Fangyuan Xiong, Yuansheng Zhu, Li Wang, Jiajun Zhang, Shenghui Dong & Mingxiang Lu

Chromosome-level genome assembly of the longfin barb (*Acrossocheilus longipinnis*)

Zechen E[1,2,3], Fangyuan Xiong[1,2,3], Yuansheng Zhu[1,2,3], Li Wang[1,2,3], Jiajun Zhang[1,2,3], Shenghui Dong[1,2,3], Mingxiang Lu[1,2,3]

Affiliations:

[1]Scientific Institute of Pearl River Water Resources Protection, Guangzhou 510611, China

[2]Hongshui River Rare Fish Conservation Center, Guigang 537200, China

[3]Engineering Research Center of Hongshui River Rare Fish Conservation, Guangxi Zhuang Autonomous Region, Guigang 537200, China

Corresponding:

Fangyuan Xiong

Address: Scientific Institute of Pearl River Water Resources Protection, Guangzhou 510611, China

E-mail: xiongfangyuan@ihb.ac.cn

Abstract

The longfin barb (*Acrossocheilus longipinnis*), a vulnerable cyprinid fish endemic to China's Pearl River basin, is of significant conservation concern and also popular in the ornamental fish trade. To facilitate genetic research and molecular breeding for this species, we generated a high-quality genome by integrating PacBio HiFi long reads and Hi-C sequencing data. The final assembly spans approximately 936.04 Mb, achieving high continuity with a contig N50 of 36.09 Mb. Assessment of genome quality revealed excellent completeness (98.76% BUSCO score) and accuracy (QV = 54.46; GCI =29.76; CRAQ = 96.40). The vast majority of the sequence (927.20 Mb, 99.06%) was successfully anchored to 25 chromosomes. Annotation predicted 24,718 protein-coding genes and identified approximately 553.06 Mb (59.09%) of repetitive elements. This high-quality chromosome-scale reference genome provides a crucial foundation for investigating the genomic underpinnings of *A. longipinnis* evolution and will significantly advance molecular breeding programs aimed at its conservation

and sustainable utilization.

Background & Summary

The cyprinid genus *Acrossocheilus* Oshima, 1919 comprises 26 valid species distributed across East and Southeast Asia, including mainland China, Taiwan, Hainan, Laos, and Vietnam. These small- to medium-sized barbines are principally characterized by a medially interrupted lower lip with two thick lateral lobes, which are anteriorly separated from the lower jaw by a distinct groove running the entire length of the jaw[1]. These species are widely distributed across Laos, Vietnam, and southern China, including Hainan, Taiwan, and other parts of the Chinese mainland[2]. *Acrossocheilus longipinnis*, is an endemic species of mainland China currently known only from the Pearl River basin, exhibits an elongated, laterally compressed body covered in dense scales with a prominent lateral line. Its silver-gray base coloration is adorned with five distinct pale yellow vertical bars. A key morphological trait in males is the elongation of the last branched ray and first unbranched ray of the dorsal fin into filamentous projections. Valued in the ornamental fish trade for its unique morphology and striking coloration, this species has experienced significant wild population declines, as indicated by recent fishery resource assessments. This decline is attributed to multiple anthropogenic threats, including cascading hydropower dam construction, extensive sand mining, overfishing, environmental pollution, and the introduction of invasive fish species. Consequently, *A. longipinnis* has been classified as Vulnerable on the IUCN Red List.

Molecular research on *A. longipinnis* remains limited. To date, only its mitochondrial genome has been sequenced[3]. Crucially, a reference genome assembly for this species is still lacking, which significantly hinders progress in understanding its biology, advancing genetic breeding programs, and developing desirable aquacultural traits. Recent advancements in DNA sequencing technologies, however, offer unprecedented opportunities for genomic research. Notably, Pacific BioSciences' (PacBio) Circular Consensus Sequencing (CCS) mode provides long read lengths (10–20 kb) and high accuracy (>99%), thus greatly facilitating *de novo* assembly studies of both plant and animal genomes[4,5]. According to the comprehensive overview by Li and Durbin (2024)[6], high-fidelity (HiFi) sequencing enables near-telomere-to-telomere assemblies by resolving repetitive regions and segmental duplications that are challenging for short-read approaches. In a parallel manner, Wang *et al*. (2025) emphasize HiFi's applications in complex genomic regions, such as centromeres and ribosomal DNA arrays, and its superiority in variant detection and phasing compared to other long-read platforms like Oxford Nanopore Technologies[7]. When integrated with complementary approaches such as chromosomal conformational capture (Hi-C) sequencing, these technologies enable the generation of highly contiguous, chromosome-level genome assemblies. Such integrated approaches have already been successfully applied in another *Acrossocheilus* species, *Acrossocheilus fasciatus*, demonstrating their utility in resolving

genomic architectures within this genus[8].

Here, we assembled a high-quality genome of *A. longipinnis* by combining short sequencing reads, PacBio HiFi long reads, and Hi-C sequencing data. The final longfin barb genome assembly had a total length of 936.04 Mb, with 99.06% (927.20 Mb) of the sequences successfully anchored to 25 chromosomes. The assembly demonstrated high continuity (contig N50 = 36.09 Mb) and completeness (BUSCO = 98.76%), supported by quality metrics including a QV value of 54.46, a GCI score of 29.76, and a CRAQ value of 96.40. Subsequent annotation identified 24,718 protein-coding genes and 553.06 Mb of repetitive sequences. This high-quality genome assembly not only facilitates population genetic research and evolutionary analyses of *A. longipinnis* but also provides valuable resources for optimizing genetic breeding efforts.

Methods

Sampling, DNA and RNA extraction. This study was carried out according to the recommendations for the care and use of animals for scientific purposes set up by the Animal Care and Use Committee of the Chinese Academy of Fishery Sciences (ACUC-CAFS). Samples of *A. longipinnis* were collected from Hechi City, Guangxi Zhuang Autonomous Region, China (coordinates: 107°33′–108°13′ E, 24°22′–24°55′ N). Tissue samples were promptly collected, snap-frozen in liquid nitrogen, and then stored at -80°C. DNA and RNA extraction, library construction, and sequencing in this study were performed using standard experimental and analytical protocols provided by NextOmics Biosciences (Wuhan, China).

Long read DNA preparation and sequencing. A total of 8 μg of high-quality genomic DNA was extracted from muscle tissue using a Qiagen DNeasy Blood and Tissue Kit (Qiagen, USA) according to the manufacturer's instructions. The quality and concentration of the extracted DNA were assessed using a NanoDrop One spectrophotometer (Thermo Scientific, USA) and 1% agarose gel electrophoresis. PacBio long insert libraries were prepared using the SMRTbell Express Template Prep Kit 2.0 according to manufacturers' instructions, with an insert size of approximately 20 kb. The libraries were sequenced on the PacBio Revio system in CCS mode. Subreads were processed with SMRTLink (v11.1.0)[9] using the parameters "–minPasses 3 –minPredictedAccuracy 0.99 –minLength 500", producing approximately 114.37 Gb HiFi reads with an N50 size of 16,728 (Table 1). The parameter "minPredictedAccuracy" set to 0.99 in the context of PacBio SMRTLink software means that, during the data processing of sequencing reads, only those reads that have a predicted accuracy of 99% or higher will be retained for further analysis.

Short read DNA preparation and sequencing. The extracted DNA (~5 μg) was randomly sheared into approximately 350 bp fragments, and a short fragment library was constructed using the MGIEasy Universal

DNA Library Prep Set (MGI, China). Sequencing was conducted on the MGISEQ T7 platform (MGI, China), resulting in a total of 56.50 Gb of short sequencing reads, each 150 bp in length (Table 1).

Hi-C DNA library preparation and sequencing. A Hi-C library was generated using the DpnII restriction enzyme (GrandOmics, China). Muscle tissue samples were treated with 1% formaldehyde at room temperature for 10–30 minutes to crosslink chromatin-interacting proteins. Subsequently, the DNA was digested with the restriction enzyme, and the 5' overhangs were repaired with a biotinylated residue. A paired-end library with insert sizes of approximately 300 bp was prepared and then sequenced on the MGISEQ T7 platform (MGI, China). A total of 127.92 Gb of clean data was obtained from 129.09 Gb of sequencing data using the software fastp (v0.19.5)[10] with parameters "-w 16 --length_required 150" (Table 1).

RNA library preparation and sequencing. For the purpose of RNA sequencing, we extracted total RNA from muscle, heart, liver, spleen, gill, kidney, skin, and fin tissues using the TRIzol reagent (Invitrogen, USA) following the manufacturer's protocol. Mixed total RNA purity was assessed with a NanoPhotometer spectrophotometer (IMPLEN, CA, USA), while RNA concentration was quantified using the Qubit RNA Assay Kit with a Qubit 2.0 Fluorometer (Life Technologies, CA, USA). RNA-seq libraries were prepared using the TruSeq Stranded mRNA Library Prep Kit (Illumina, USA) according to the manufacturer's instructions. Sequencing was performed on a MGISEQ T7 platform (MGI, China), generating 150 bp paired-end reads.

Genome size estimation. The genome size of *A. longipinnis* was estimated through *k*-mer profiling. First, raw short sequencing reads underwent quality control using fastp (v0.19.5)[10]. Using K-mer analysis ($K = 21$) of quality-filtered short reads, the genome size of *A. longipinnis* was first estimated with findGSE (v1.94.R)[11]. The genome size of *A. longipinnis* was estimated to be 961,326,620 bp (Fig. 1).

***De novo* assembly and Hi-C assembly.** Primary contigs were assembled from HiFi reads using Hifiasm (v 0.25.0)[12] with parameters: -t 100 --n-hap 2 --telo-m TTAGGG hifi.fa. Genome base errors (single-nucleotide variants and small indels) were corrected using NextPolish (v1.4.1)[13], integrating both HiFi reads and quality-filtered short reads. This yielded 132 contigs spanning 936.78 Mb with an N50 of 33.36 Mb. For chromosomal anchoring, BWA (v0.7.12)[14] was used to align the Hi-C clean data to the assembled contigs. Low-quality reads were filtered using the HiC-Pro pipeline[15] with default parameters. The remaining valid reads were employed to anchor chromosomes using Juicer[16] and the 3d-dna pipeline[17], followed by manual correction with Juicebox (v2.13.07)[18]. In the 3d-DNA pipeline, a default gap size of 500 bp was inserted between consecutive sequences. Next, we applied the LR_Gapcloser[19] program to close the gaps in the assemblies. To enhance genome quality, the assemblies were polished with NextPolish2 (v0.2.0)[20] using HiFi reads and quality-filtered short reads. Ultimately, 99.06% of contig sequences were anchored to 25 pseudochromosomes, with only two gaps

remaining (one each in pseudochromosomes 5 and 20) (Table 2 and Fig. 2). The sizes of these two gaps were 3bp and 151bp, respectively. The longest and shortest pseudochromosomes measured 56.97 Mb and 28.75 Mb, respectively (Table 3). The final assembly totaled 936.04 Mb with a contig N50 of 36.09 Mb (Table 2 and Fig. 3).

Repetitive sequence annotation. Repeat elements in the *A. longipinnis* genome were annotated employing a combined methods of homology alignment and *de novo* searches. The homology-based blast was performed against the RepBase database (http://www.girinst.org/repbase/)[21] using RepeatMasker (v4.0.7)[22] and Proteinmask software for known repeat elements. For *de novo* annotation, we firstly employed LTR_FINDER (v1.06)[23] and RepeatModeler (v1.0.4)[24] to bulid a *de novo* repeat library, and then was used to predict repeat elements using RepeatMasker (v4.0.7)[22] with default parameters. Additionally, Tandem Repeat Finder (v4.10.0)[25] was used to discern tandem repeats with default parameters. In detail, a total of 553.06 Mb (~59.09%) of repetitive sequences were obtained. Among the interspersed repeats, long terminal repeats were the most prevalent type, accounting for 32.67% of the genome (Table 4).

Gene prediction and functional annotation. Gene prediction was performed using a multifaceted approach incorporating transcriptome-based, homology-based, and *ab initio* methods. For the transcriptome-based prediction, a total of 8.73 Gb of RNA-seq clean reads were aligned to the *A. longipinnis* assembly using Hisat2 (v2.2.1)[26] (Table 5). Stringtie (v1.2.2)[27] was then utilized to assemble transcripts based on the alignment results. In addition, the RNA-seq data were *de novo* assembled by Trinity (v2.15.2)[28] with parametrs: --seqType fq --max_memory 200G --min_kmer_cov 2 --min_glue 2 --CPU 60 --min_contig_length 200. Afterwards, the assembled transcripts were aligned against the *A. longipinnis* assembly using Program to Assemble Spliced Alignment (PASA; v2.4.1)[29]. For homology-based prediction, we utilized Miniport (v0.11) to conduct a comparative analysis of the protein sequences from seven vertebrate species, including *A. fasciatus*[8], *Ctenopharyngodon idella*[30], *Cyprinus carpio*[31], *Poropuntius huangchuchieni*[32], *Onychostoma macrolepis* (GCF_012432095.1), *Danio rerio* (GCF_049306965.1), and *Homo sapiens* (GCF_009914755.1). For *ab initio* prediction, 2,000 high-quality genes from PASA were randomly selected as the training set for model training with AUGUSTUS (v3.2.3)[33]. AUGUSTUS (v3.2.3)[33] was then employed to predict coding regions in the repeat-masked genome. In addition, Fgenesh (v2.4.5)[34] was also used for *ab initio* prediction. Finally, all gene models were integrated using EvidenceModeler (v2.1.0)[35]. The final comprehensive gene set comprised 24,718 genes (Table 6), with an average of 10.44 exons per gene, an exon length of 170.64 bp, and a coding sequence (CDS)

length of 1781.09 bp.

After gene prediction, the finalized gene sets derived from the preceding methods underwent functional annotation through matching with a variety of databases. Briefly, amino-acid sequences were aligned to SwissProt[36], Kyoto Encyclopedia of Genes and Genomes (KEGG)[37], and the NCBI nonredundant database (NR) using the Diamond (v 2.1.10)[38] with an E-value cutoff of 1e-05. Protein domains were identified using the InterProScan (v5.30)[39] program, and Gene Ontology (GO) terms for each gene were also extracted through InterProScan. Overall, 24,228 (98.02%) of the predicted protein-coding genes were functionally annotated (Table 6).

Data Records

The raw sequencing data have been deposited into the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database with accession number SRP604471[40] under BioProject number PRJNA1297891. Additionally, the genome assembly and annotation are available at the Figshare dataset[41].

Technical Validation

Genome assembly and gene prediction quality assessment. We employed a multi-faceted approach to rigorously evaluate the precision and integrity of the *A. longipinnis* genome assembly. First, we utilized Merqury (v1.3)[42] with a combination of HiFi long reads and short reads, setting the *K*-mer value at 19, to calculate the consensus QV. The analysis yielded a QV of 54.46, indicating a high level of accuracy in the assembled genome sequence (Table 2). Subsequently, we aligned the HiFi reads and quality-filtered short reads to the assembly using minimap2 (v2.24-r1122)[43] and BWA (v0.7.12)[14], respectively. This process demonstrated an exceptional alignment rate, with 99.99% of the HiFi reads and 99.85% of the short sequencing reads successfully mapped to the genome (Table 2). Centromeric regions were predicted following the method described in the recent telomere-to-telomere genome study of *Cyprinus carpio*[31]. We found the centromeric regions displayed the canonical features of centromeres: high repetitive sequence content, low gene density, and low HiFi read coverage depth, aligning with the previous research reports[31,44] (Fig. 4). Additionally, both assembly gaps were located within highly repetitive regions, one of which lay within a centromere. The HiFi read coverage in the regions flanking these gaps was notably lower compared to the genome-wide average. Clipping information for revealing assembly quality (CRAQ, v1.10)[45] was used to assess the accuracy of our genome assembly based on PacBio HiFi and quality-filtered short reads, resulting in a S-AQI of 96.40, confirming high assembly quality. In addition, genome continuity inspector (GCI, v1.0)[46] yielded a value of 29.76, which was comparable to that of the chicken complete genome[47]. To assess genome completeness, we performed an analysis with Benchmarking Universal Single-Copy Orthologs (BUSCO) (v5.5.0)[48] using the actinopterygii_odb10 database.

The results showed that 98.76% of the BUSCO genes were complete, including 97.53% single-copy and 1.24% duplicated orthologs, while only 0.93% of the genes were fragmented (Fig. 5). Furthermore, BUSCO analysis of the genome annotation revealed 97.14% of the recognized BUSCOs were complete, consisting of 95.11% single-copy and 2.03% duplicated genes (Fig. 5). Collectively, these comprehensive evaluation metrics strongly suggest that the *A. longipinnis* genome assembly has achieved a high standard of quality, providing a reliable resource for subsequent genetic and biological studies.

Data availability

Raw sequencing data have been deposited in the NCBI SRA database under BioProject accession number PRJNA1297891, with accession numbers as follows: PacBio HiFi: SRR34770991[49]; Hi-C: SRR34770992[50]; RNA sequencing: SRR34770990 [51]; DNA short-read sequencing: SRR34770993[52]. The genome assembly has been uploaded to the GenBank database under the accession GCA_054083375.1[53]. Moreover, the genome assembly, annotation files (GFF3, FASTA), and gene functional annotation datasets, are available via Figshare[41]. All datasets are publicly accessible without restrictions.

Code availability

No specific code or script was used in this work. Commands used for data processing were all executed according to the manuals and protocols of the corresponding software.

Acknowledgements

This work is supported by Operating funds of Hongshui River Rare Fish conservation Center.

Author contributions

Zechen E conceived this study, designed the experiment, and performed data analysis. Fangyuan Xiong contributed to the experimental design, collected samples, and performed data analysis. Yuansheng Zhu and Li Wang provided funding and contributed to conceptualization. Jiajun Zhang and Shenghui Dong assisted in methodology and data curation. All authors have read and approved the final manuscript.

Ethics declarations

Competing interests

The authors declare no competing interests.

Ethical approval

The study did not involve any wild animals. All experimental procedures involving fish were conducted in strict compliance with the Guide for the Hongshui River Rare Fish Conservation Center to minimize animal suffering and ensure animal welfare.

Figure legends

Fig. 1 *K*-mer frequency distribution estimated. The observed *K*-mer (raw *K*-mer) frequencies (in grey), fitted *K*-mer frequencies (in blue) with skew normal distribution model, and overall fitting (in red) that concatenated observed and fitted *K*-mer frequencies.

Fig. 2 Hi-C assembly of chromosome interactive heat map. The abscissa and ordinate represent the order of each bin on the corresponding chromosome group. The colour block illuminates the intensity of interaction from white (low) to red (high).

Fig. 3 Snail plot showing the features of the assembled *A. longipinnis* genome. The main plot is divided into 1,000 size-ordered bins around the circumference with each bin representing 0.1% of the 936,040,231bp assembly. The distribution of chromosome lengths is shown in dark grey with the plot radius scaled to the longest chromosome present in the assembly. Orange and pale-orange arcs show the N50 and N90 chromosome lengths (36,094,363 and 29,100,020 bp), respectively. The pale grey spiral shows the cumulative chromosome count on a log scale with white scale lines showing successive orders of magnitude. The blue and pale-blue area around the outside of the plot shows the distribution of GC, AT and N percentages in the same bins as the inner plot.

Fig. 4 Characterization of centromeric regions and gap locations visualized by a circos plot. From inside to outside: Gene density in 1 Mb sliding windows; Percentage of repetitive sequence in 1 Mb sliding windows; Centromere density in 1 Mb sliding windows; Gap locations; HiFi reads coverage depth; The length of pseudochromosome in the size of Mb.

Fig. 5 BUSCO assessments of *A. longipinnis* genome and gene sets.

Tables

Table 1. Summary of DNA sequencing data of *A. longipinnis* genome.

Types	Reads Number	Total length (Gb)	Genome depth*	N50 length of reads (bp)
Short sequencing raw reads	376,689,082	56.50	58.78	150
Hi-C raw data	860,583,098	129.09	134.28	150
PacBio HiFi reads	6,815,862	114.37	118.97	16,728

Note: * Estimated based on the assembly size of 961.33 Mb.

Table 2. Summary statistics of *A. longipinnis* assembly.

Item	*Acrossocheilus longipinnis*
Size of assembly (Mb)	936.04
Contig N50 (Mb)	36.09
Scaffold N50 (Mb)	36.09
Gap number	2

Pseudochromosome number	25
Hi-C anchored ratio	99.06%
GC content	37.34%
Genome complete BUSCOs	98.76%
Quality value	54.46
GCI score	29.76
CRAQ S-AQI	96.40
Short reads mapping rate	99.85%
HiFi reads mapping rate	99.99%
Repetitive sequences	59.09%
Number of protein-coding genes	24,718

Note: GCI: genome continuity inspector; CRAQ: Clipping information for Revealing Assembly Quality; The lineage dataset used in BUSCO is actinopterygii_odb10.

Table 3. Pseudo-chromosome length statistics after Hi-C assisted assembly.

Pseudomolecule	Length (bp)	GC content	Gap number
Chr01	56,969,854	37.67%	0
Chr02	51,036,322	37.33%	0
Chr03	48,695,360	37.19%	0
Chr04	43,963,889	36.91%	0
Chr05	43,193,524	37.08%	1
Chr06	41,074,459	38.25%	0
Chr07	39,989,793	37.14%	0
Chr08	39,090,749	37.17%	0
Chr09	38,848,481	36.91%	0
Chr10	36,771,069	38.00%	0
Chr11	36,094,363	37.25%	0
Chr12	35,977,621	36.84%	0
Chr13	34,662,433	37.36%	0
Chr14	33,594,154	37.27%	0
Chr15	33,577,499	37.25%	0
Chr16	33,551,769	37.40%	0
Chr17	33,057,203	37.09%	0
Chr18	32,905,817	37.27%	0
Chr19	32,771,601	37.00%	0
Chr20	31,839,502	37.86%	1
Chr21	31,683,128	37.08%	0
Chr22	31,088,712	37.10%	0
Chr23	29,100,020	36.66%	0
Chr24	28,912,336	36.93%	0
Chr25	28,746,342	37.19%	0

Table 4. Statistics of interspersed repetitive sequences in *A. longipinnis* assembly.

Type	Repbase TEs		TE protiens		*De novo*		Combined TEs	
	Length (Bp)	% in genome	Length (Bp)	% in genome	Length (Bp)	% in genome	Length (Bp)	% in genome
DNA	177,942,007	19.01	446,123	0.05	159,800,106	17.07	254,780,347	27.22
LINE	36,012,676	3.85	21,443,429	2.29	93,369,750	9.97	109,976,318	11.75
SINE	4,683,147	0.50	0	0.00	1,364,525	0.15	6,045,978	0.65
LTR	31,646,572	3.38	12,418,421	1.33	301,000,263	32.16	305,799,237	32.67
Other	16,949	0.00	0	0.00	0	0.00	16,949	0.00
Unknown	0	0.00	0	0.00	88,695,758	9.48	88,695,758	9.48
Total	242,533,048	25.91	34,301,627	3.66	536,071,526	57.27	545,573,809	58.29

Note: This statistical table does not contain Tandem Repeats, some elements may partly include another element domain. Combined: the non-redundant consensus of all repeat prediction/classification methods employed. LINE, long interspersed nuclear elements; SINE, short interspersed nuclear elements; LTR, long terminal repeat.

Table 5. Summary of RNAseq sequencing data of *A. longipinnis* genome.

Raw data (Gb)	Clean data	Clean Q20	GC rate	Total MappingRatio	Uniquely MappingRatio
10.13	8.73	98.79%	46.20%	98.04%	88.95%

Table 6. Statistics of functional annotation result.

Type		Gene number	Percentage (%)
Total		24,718	100.00
NR		24,002	97.10
Swissprot		21,942	88.77
KEGG		20,909	84.59
TrEMBL		23,957	96.92
Interpro	All	23,778	96.20
	GO	17,179	69.50
Annotated		**24,228**	**98.02**
Unannotated		490	1.98

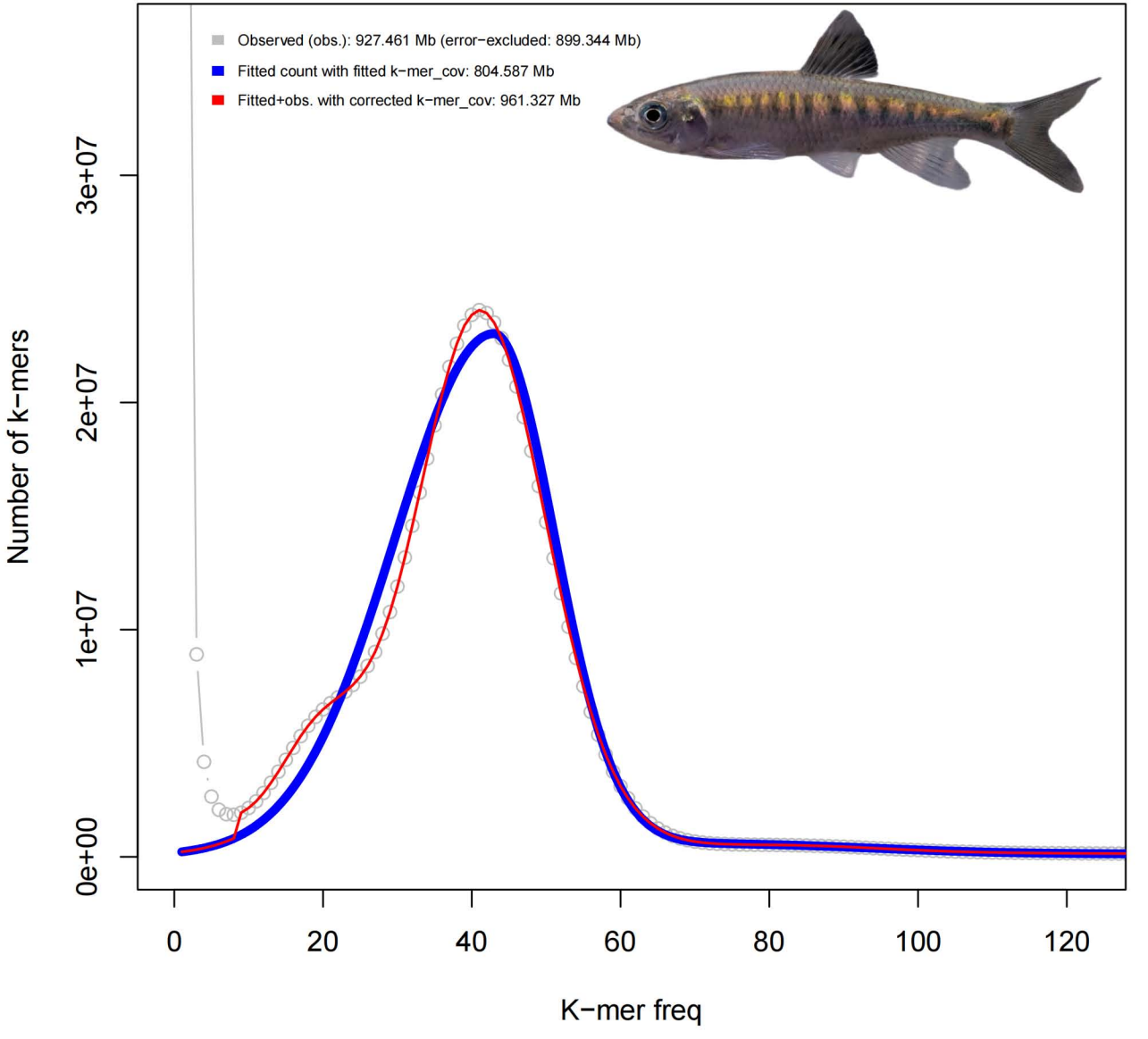
Note: Total, all annotated genes; Annotation, gene products could be annotated by at least one of the databases; KEGG, Kyoto Encyclopedia of Genes and Genomes; NR, Non-Redundant Protein Sequence Database; Swissport, Swiss-Prot Protein Knowledgebase; GO, Gene Ontology; TrEMBL, Translation of European Molecular Biology Laboratory; Interpro, Integrative Protein Signature Database.

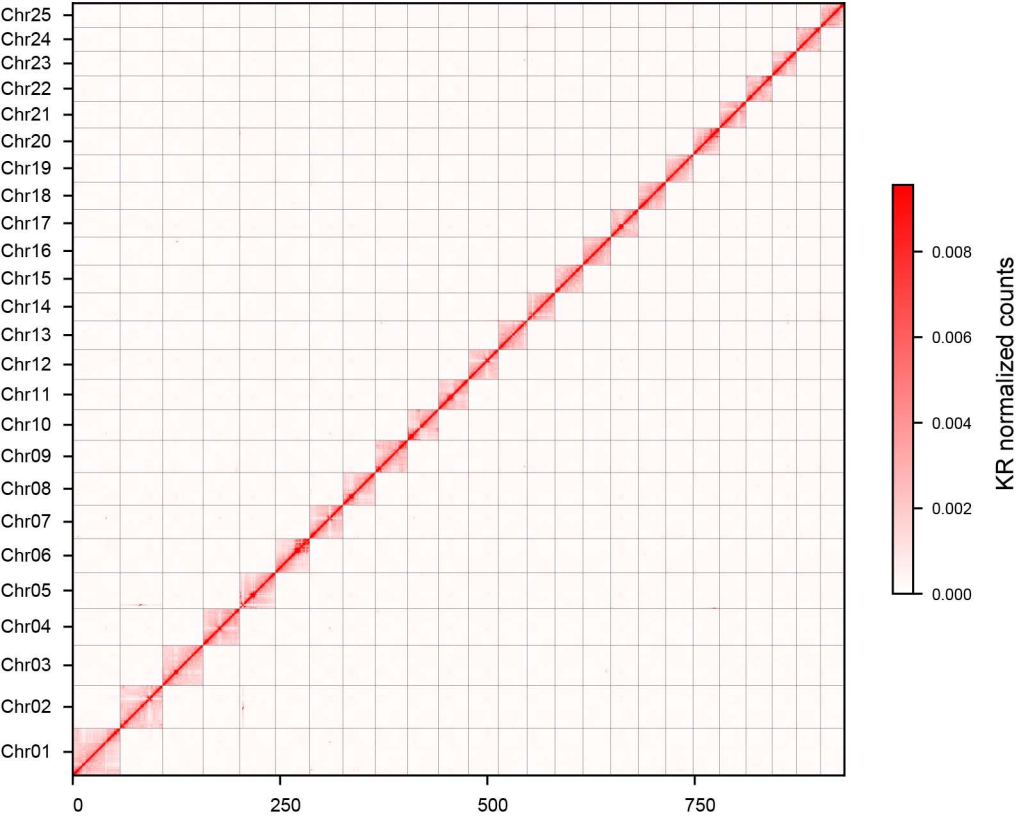
References

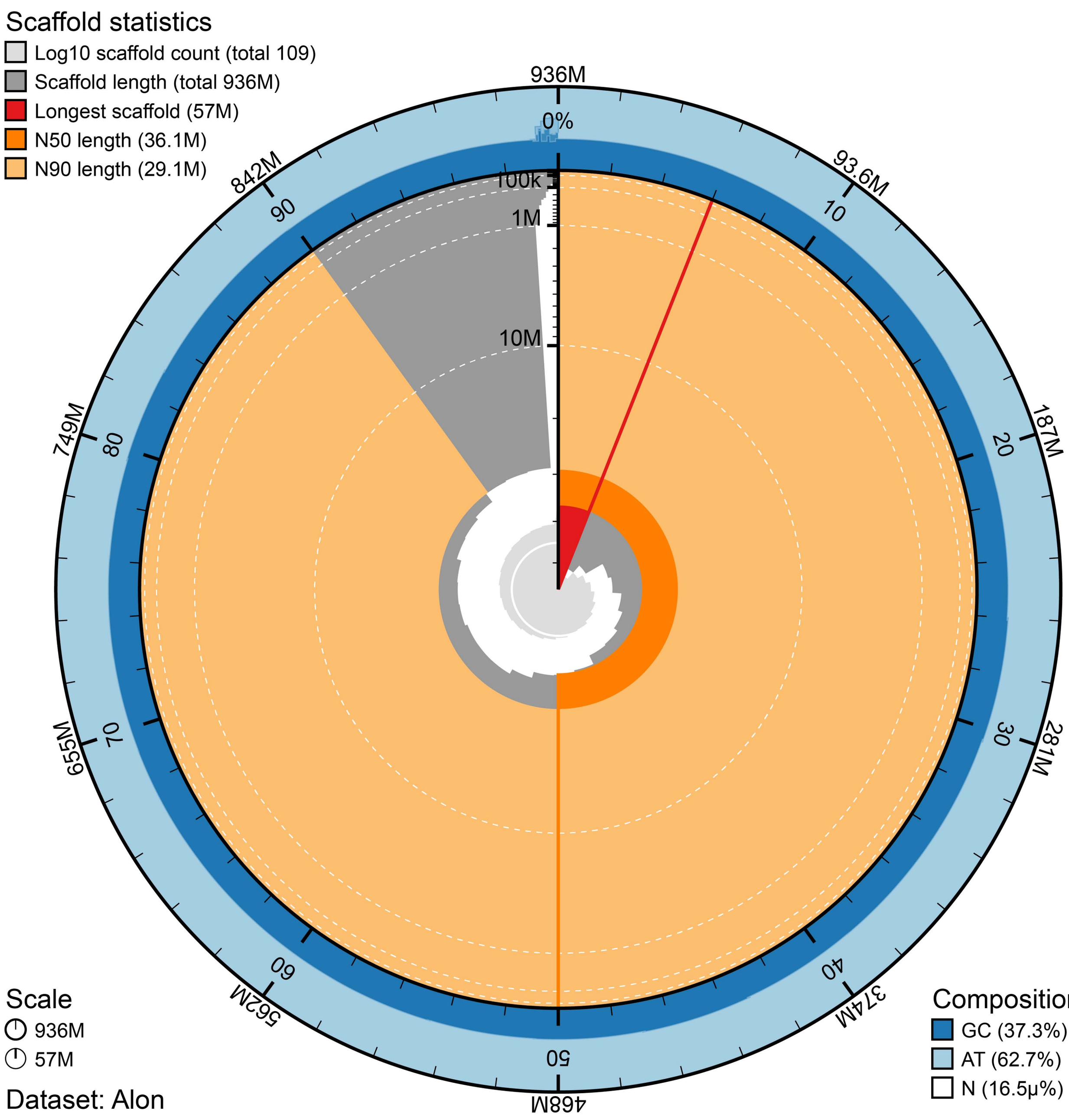
1. Yuan, L.Y., Liu, X.X. & Zhang, E. Mitochondrial phylogeny of Chinese barred species of the cyprinid genus Acrossocheilus Oshima, 1919 (Teleostei: Cypriniformes) and its taxonomic implications. *Zootaxa* **4059**, 151-168 (2015).
2. Chen, T.E. *et al.* A New Species of the Genus Acrossocheilus Oshima, 1919 (Cypriniformes: Cyprinidae) from the Dabie Mountains. *Animals* **15**, 734 (2025).
3. Hou, X.-J. *et al.* Complete mitochondrial genome of the freshwater fish Acrossocheilus longipinnis (Teleostei: Cyprinidae): genome characterization and phylogenetic analysis. *Biologia* **75**, 1871-1880 (2020).
4. Wenger, A.M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature biotechnology* **37**, 1155-1162 (2019).
5. Lovell, J.T. *et al.* Four chromosome scale genomes and a pan-genome annotation to accelerate pecan tree breeding. *Nature Communications* **12**, 4125 (2021).
6. Li, H. & Durbin, R. Genome assembly in the telomere-to-telomere era. *Nature Reviews Genetics* **25**, 658-670 (2024).
7. Wang, B. *et al.* Long and Accurate: How HiFi Sequencing is Transforming Genomics. *Genomics Proteomics Bioinformatics* **23**(2025).
8. Zheng, J. *et al.* Chromosome-level genome assembly of Acrossocheilus fasciatus using PacBio sequencing and Hi-C technology. *Scientific Data* **11**, 166 (2024).
9. Chin, C. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods* **10**, 563-569 (2013).
10. Chen, S., Zhou, Y., Chen, Y. & Jia, G. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884-i890 (2018).
11. Sun, H., Ding, J., Piednoël, M. & Schneeberger, K. findGSE: estimating genome size variation within human and Arabidopsis using k-mer frequencies. *Bioinformatics (Oxford, England)* **34**, 550-557 (2018).
12. 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).
13. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics (Oxford, England)* **36**, 2253-2255 (2020).
14. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754 - 1760 (2009).
15. Servant, N. *et al.* HiC-Pro: An optimized and flexible pipeline for Hi-C data processing. *Genome Biology* **16**(2015).
16. Durand, N. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems* **3**, 95-98 (2016).
17. Dudchenko, O. *et al.* De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, eaal3327 (2017).

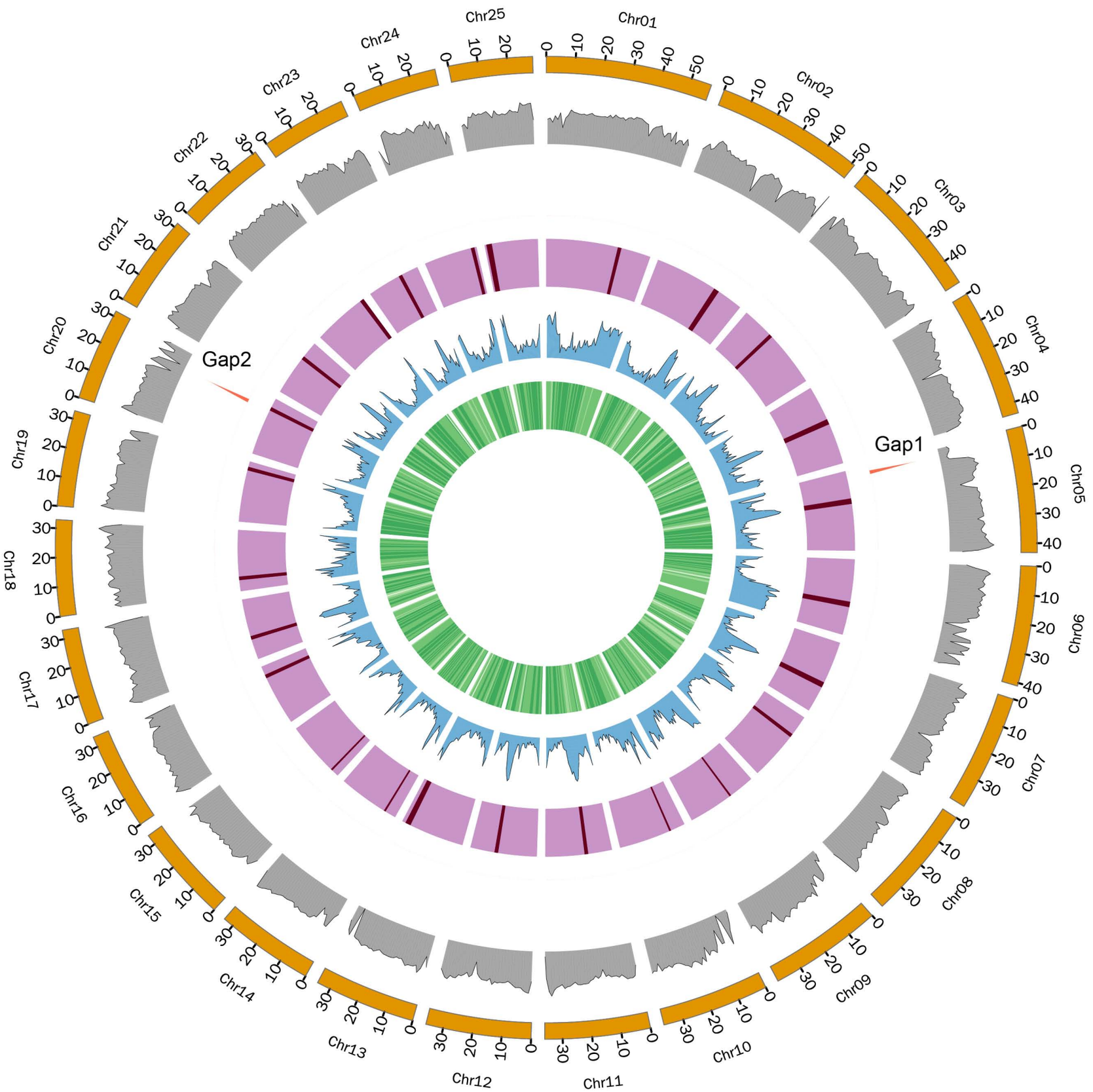
18. Durand, N.C. *et al.* Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell systems* **3**, 99-101 (2016).
19. Xu, G.-C. *et al.* LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. *GigaScience* **8**(2018).
20. Hu, J. *et al.* NextPolish2:a repeat-aware polishing tool for genomes assembled using HiFi long reads. (bioRxiv, 2023).
21. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research* **110**, 462-467 (2005).
22. Price, A.L., Jones, N.C. & Pevzner, P.A. *De novo* identification of repeat families in large genomes. *Bioinformatics (Oxford, England)* **21 Suppl 1**, i351-8 (2005).
23. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic acids research* **35**, W265-8 (2007).
24. Chen, N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Current Protocols in Bioinformatics* **5**(2004).
25. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573-580 (1999).
26. Kim, D., Langmead, B. & Salzberg, S.L. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* **12**, 357-360 (2015).
27. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology* **20**, 278 (2019).
28. Grabherr, M.G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**, 644-652 (2011).
29. Haas, B. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31**, 5654-5666 (2003).
30. Liu, F. *et al.* The telomere-to-telomere gapless genome of grass carp provides insights for genetic improvement. *GigaScience* **14**(2025).
31. Yuan, J. *et al.* A telomere-to-telomere genome assembly of koi carp (*Cyprinus carpio*) using long reads and Hi-C technology. *GigaScience* **14**(2025).
32. Chen, L. *et al.* Chromosome-level genome of Poropuntius huangchuchieni provides a diploid progenitor-like reference genome for the allotetraploid Cyprinus carpio. *Molecular ecology resources* **21**, 1658-1669 (2021).
33. Stanke, M. & Morgenstern, B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic acids research* **33**, W465-7 (2005).
34. Solovyev, V., Kosarev, P., Seledsov, I. & Vorobyev, D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome biology* **7 Suppl 1**, S10.1-12 (2006).
35. Haas, B.J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* **9**, R7 (2008).

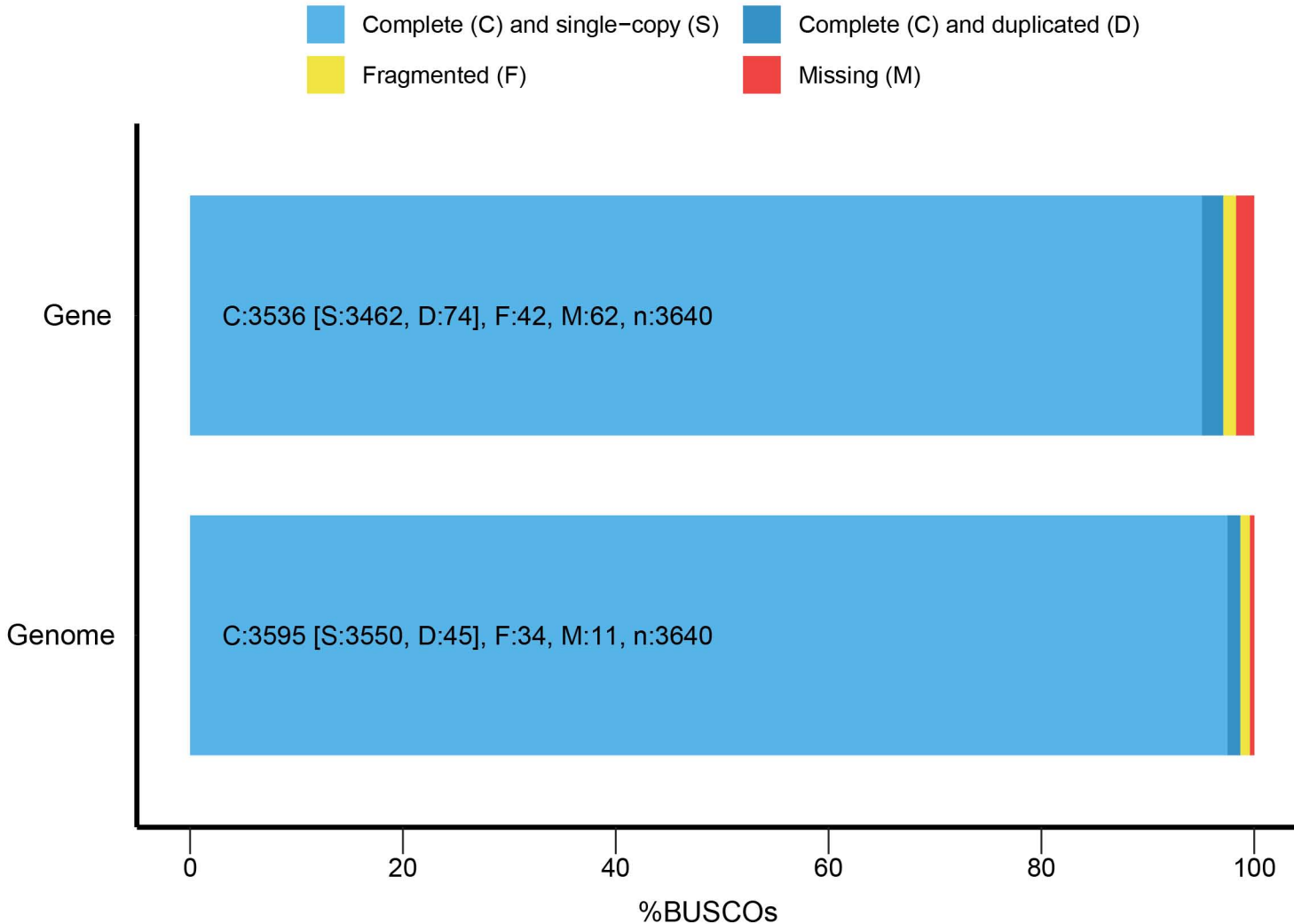
36. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Research* **27**, 49-54 (1999).
37. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**, 27-30 (2000).
38. Buchfink, B., Xie, C. & Huson, D.H. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**, 59-60 (2015).
39. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240 (2014).
40. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP604471 (2025).
41. Li, Jiang. Chromosome-level genome assembly of *Acrossocheilus longipinnis* using PacBio sequencing and Hi-C technology. *Figshare. Dataset*. https://doi.org/10.6084/m9.figshare.29665907.v1 (2025).
42. Rhie, A., Walenz, B.P., Koren, S. & Phillippy, A.M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21**(2020).
43. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100 (2018).
44. Yin, D. *et al.* Telomere-to-telomere gap-free genome assembly of the endangered Yangtze finless porpoise and East Asian finless porpoise. *GigaScience* **13**(2024).
45. Li, K., Xu, P., Wang, J., Yi, X. & Jiao, Y. Identification of errors in draft genome assemblies at single-nucleotide resolution for quality assessment and improvement. *Nature Communications* **14**, 6556 (2023).
46. Chen, Q., Yang, C., Zhang, G. & Wu, D. GCI: a continuity inspector for complete genome assembly. *Bioinformatics* **40** (2024).
47. Huang, Z.A.-O. *et al.* Evolutionary analysis of a complete chicken genome. *Proc Natl Acad Sci U S A*. 120(8):e2216641120 (2023).
48. Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. & Zdobnov, E.M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212 (2015).
49. *NCBI sequence read archive* https://identifiers.org/ncbi/insdc.sra:SRR34770991 (2025).
50. *NCBI sequence read archive* https://identifiers.org/ncbi/insdc.sra:SRR34770992 (2025).
51. *NCBI sequence read archive* https://identifiers.org/ncbi/insdc.sra:SRR34770990 (2025).
52. *NCBI sequence read archive* https://identifiers.org/ncbi/insdc.sra:SRR34770993 (2025).
53. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_054083375.1 (2025).


Observed (obs.): 927.461 Mb (error−excluded: 899.344 Mb)
Fitted count with fitted k−mer_cov: 804.587 Mb
Fitted+obs. with corrected k−mer_cov: 961.327 Mb
Number of k−mers
0e+00
1e+07
2e+07
3e+07
0
20
40
60
80
100
120
K−mer freq


Chr25
Chr24
Chr23
Chr22
Chr21
Chr20
Chr19
Chr18
Chr17
Chr16
Chr15
Chr14
Chr13
Chr12
Chr11
Chr10
Chr09
Chr08
Chr07
Chr06
Chr05
Chr04
Chr03
Chr02
Chr01
0
250
500
750
0.008
0.006
0.004
0.002
0.000
KR normalized counts


Scaffold statistics
Log10 scaffold count (total 109)
Scaffold length (total 936M)
Longest scaffold (57M)
N50 length (36.1M)
N90 length (29.1M)
936M
0%
93.6M
187M
281M
374M
468M
562M
655M
749M
842M
100k
1M
10M
10
20
30
40
50
60
70
80
90
Scale
936M
57M
Dataset: Alon
Composition
GC (37.3%)
AT (62.7%)
N (16.5µ%)


Chr01
Chr02
Chr03
Chr04
Chr05
Chr06
Chr07
Chr08
Chr09
Chr10
Chr11
Chr12
Chr13
Chr14
Chr15
Chr16
Chr17
Chr18
Chr19
Chr20
Chr21
Chr22
Chr23
Chr24
Chr25
Gap1
Gap2


Complete (C) and single−copy (S)
Complete (C) and duplicated (D)
Fragmented (F)
Missing (M)
Gene
C:3536 [S:3462, D:74], F:42, M:62, n:3640
Genome
C:3595 [S:3550, D:45], F:34, M:11, n:3640
0
20
40
60
80
100
%BUSCOs