scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of *Chirolophis japonicus* Herzenstein, 1890 (Stichaeidae, Perciformes)

Kaiqiang Liu[1,2,5], Qi Liu[3,5], Yinquan Qu[4] & Tianxiang Gao[4]✉

Chirolophis japonicus **is notable for its slender body and distinctive fringed head projections, features that help it camouflage among rocky reefs and make it an ecologically intriguing species in cold-temperate seas. Here, we report the first chromosome-level genome of *C. japonicus* by combining PacBio HiFi, Hi-C, and Illumina short reads. The genome size was 617.85 Mb, with repetitive content of 38.74% and contig N50 of 9.70 Mb. About 98.51% of genomic sequences were anchored onto 28 chromosomes, including six chromosomes with both telomeres and 17 chromosomes with one-end telomeres. A total of 22,165 protein-coding genes were predicted, of which 98.39% were functionally annotated. The genomic quality was further confirmed by Merqury with an average consensus quality value (QV) of 45.26 and the Benchmarking Universal Single-Copy Orthologs (BUSCO) value of 98.65%. The availability of the chromosome-level *C. japonicus* genome will provide a foundation for detailed analyses of population genetics, adaptation, and conservation biology, thereby filling a critical knowledge gap in the genomic resources for this species.**

Background & Summary

Chirolophis japonicus (family Stichaeidae, order Perciformes) is a cold-temperate demersal fish species widely distributed in the northwestern Pacific Ocean, including the Yellow Sea, Bohai Sea, northern Sea of Japan, Okhotsk Sea, and Bering Sea[1,2]. It typically inhabits shallow rocky reef environments that provide both shelter and feeding grounds. Morphologically, the species is characterized by a slender laterally compressed body, small cycloid scales, and distinctive fringed cortical projections on the head. Ecologically, *C. japonicus* is an omnivorous benthic fish that feeds on small fish, algae, and benthic shellfish. It reaches sexual maturity at approximately two years, with spawning usually occurring between September and November. In recent decades, however, population declines have been documented in certain regions, largely attributed to anthropogenic pressures such as overfishing, marine pollution, and habitat degradation[3,4]. These trends highlight the need for molecular resources to improve the understanding and management of this species.

Despite its ecological significance, genomic resources for *C. japonicus* remain scarce. Most previous genomic studies within the Stichaeidae family have targeted taxa with different ecological niches or phylogenetic positions, leaving *C. japonicus* underrepresented. A recent genome-wide phylogenomic survey has provided valuable insights into the evolutionary relationships of *C. japonicus* within the family[5], but no chromosome-level reference genome is currently available. This absence of genomic resources restricts comprehensive investigations into the genetic basis of its life history traits, population structure, and adaptation to cold-temperate marine environments[6,7].

Here, we present the first chromosome-level reference genome for *C. japonicus*. High-molecular-weight DNA was extracted from muscle tissue, and PacBio HiFi long-read sequencing was performed to generate high-quality

[1]State Key Laboratory of Mariculture Biobreeding and Sustainable Goods, Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao, Shandong, 266071, China. [2]Laboratory for Marine Fisheries Science and Food Production Processes, Qingdao Marine Science and Technology Center, Qingdao, Shandong, 266237, China. [3]College of Fisheries, Southwest University, Chongqing, 400715, China. [4]Fishery College, Zhejiang Ocean University, Zhoushan, 316022, Zhejiang, China. [5]These authors contributed equally: Kaiqiang Liu, Qi Liu. ✉e-mail: gaotianxiang0611@163.com

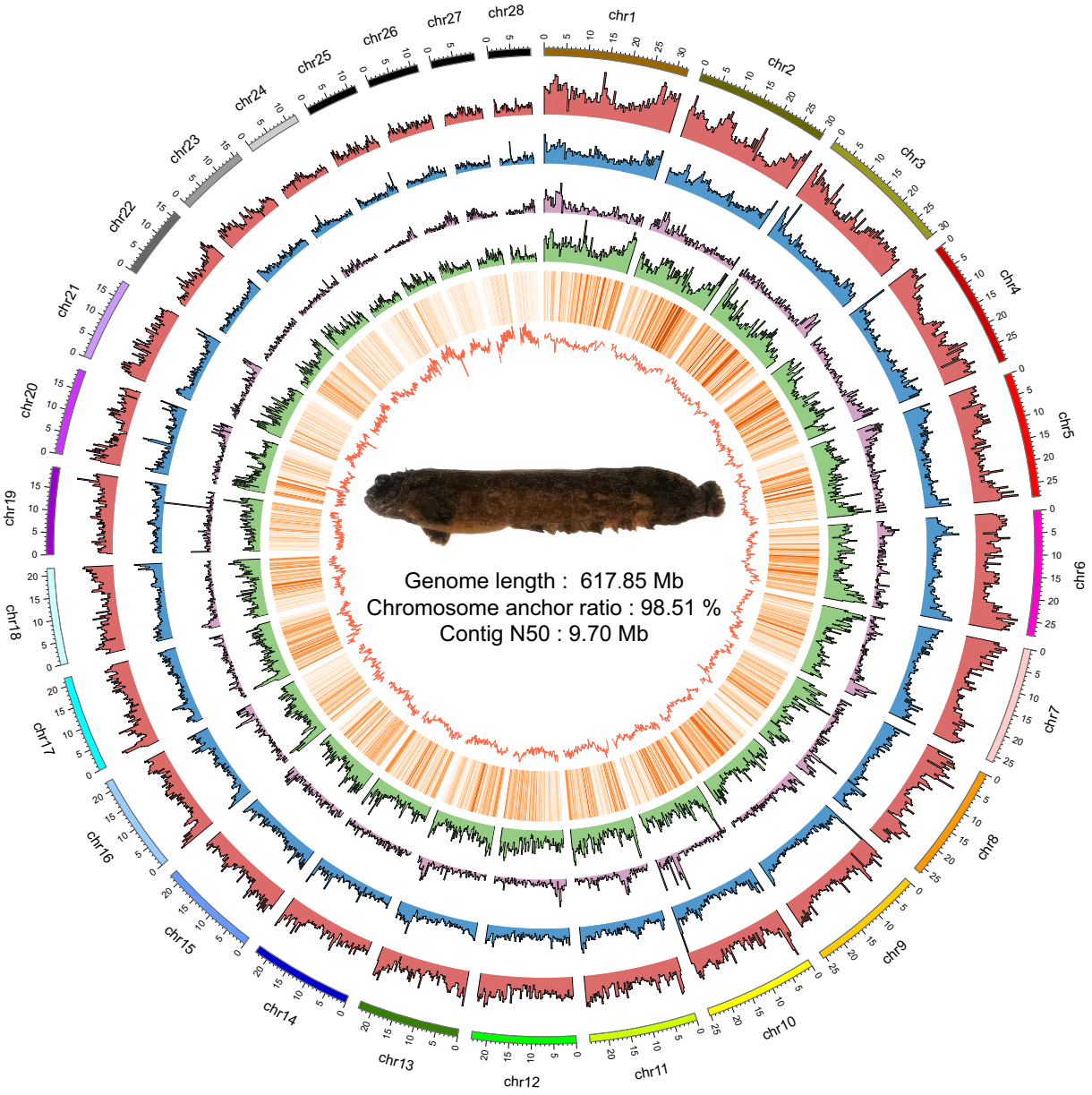

Fig. 1 The circos plot shows fundamental information about the genome. The center displays a photograph of the *C. japonicus* along with the basic genome assembly results. The circular tracks surrounding the central region, from the innermost to the outermost, correspond to GC content, gene density, LTR density, SINE density, LINE density, DNA transposon density, and the length of chromosome and their name (the most outer layer), respectively.

data. Hi-C chromatin interaction sequencing was further applied to achieve chromosome-scale scaffolding. The integration of these data allowed us to assemble a high-quality genome that will serve as a foundational resource for evolutionary and ecological genomics. This assembly enables more detailed analyses of population genetics, adaptation, and conservation biology, thereby filling a critical knowledge gap in the genomic resources for this species.

Methods

Sampling and sequencing. *Sampling.* One male individual with body length 186 mm and body weight 225 g of *C. japonicus* was collected from coastal waters of Qingdao (35°40′N, 119°30′E), China in July 2021 (Fig. 1). To obtain high-quality data, different tissues were carefully dissected and immediately preserved in liquid nitrogen, then preserved in −80 °C ultra-low temperature freezer for subsequent analyses. Specifically, muscle tissue was reserved for DNA extraction and Hi-C library preparation. Besides, five different tissues, including muscle, gill, skin, heart and kidney, were used for RNA extraction. All experimental procedures were conducted in strict accordance with relevant animal welfare regulations, and the sampling protocol was reviewed and approved by the Ethics Committee for Animal Experimentation of Zhejiang Ocean University (ZJOU-ECAE20211876).

Library	Raw data			Clean data			
	Total bases (Gb)	Q20	Sequencing depth (×)	Total bases (Gb)	Q20	Sequencing depth (×)	Mean read length (bp)
Genomic PacBio Hi-Fi long reads	/	/	/	22.37	/	36.20	18957
Genomic short reads	66.98	97.53%	108.38	65.43	97.88%	105.87	150
Hi-C short reads	72.25	97.69%	116.91	71.64	97.68%	115.92	150
Transcriptome short reads	7.17	97.84%	/	6.92	97.84%	/	150

Table 1. The summary of genomic PacBio Hi-Fi long reads, genomic short reads, Hi-C short reads and transcriptome short reads used in this study (The genome size, 618 Mb, was used to determine the sequencing depth).

PacBio HiFi long reads sequencing. High-molecular-weight genomic DNA was isolated from muscle tissue using the Blood & Cell Culture DNA Mini Kit (Genomic-tip), ensuring minimal fragmentation. The integrity and purity of the extracted DNA were assessed through agarose gel electrophoresis and spectrophotometric analysis. Qualified DNA samples were subsequently used to construct a PacBio HiFi library following the manufacturer's standard protocol. The final library was sequenced on the PacBio Sequel II platform, yielding a total of 22.37 Gb of high-quality HiFi reads, with an average read length of 18.96 kb (Table 1). These data provided a solid foundation for subsequent genome assembly.

Genomic short reads sequencing. Genomic DNA was extracted from freshly dissected muscle using Magbead Tissue DNA kit with RNase A treatment. DNA quality was evaluated by agarose gel electrophoresis, NanoDrop, and Qubit assays. Short-read libraries (~350 bp insertion size) were prepared following the Illumina protocol and sequenced on the NovaSeq 6000 platform, producing 150 bp paired-end reads. In total, 66.98 Gb of raw reads were generated (Table 1).

Hi-C short reads sequencing. To obtain chromatin interaction information for genome scaffolding, Hi-C libraries were constructed following the standard protocol. In brief, fresh tissue was cross-linked with formaldehyde to preserve chromatin architecture, digested with restriction enzymes, and proximity-ligated to form chimeric DNA fragments representing physical interactions within the genome. The resulting Hi-C libraries were purified, size-selected, and subjected to paired-end sequencing on the Illumina NovaSeq 6000 platform, producing a total of 72.25 Gb of high-quality data with a Q20 score of 97.69% (Table 1).

RNA-seq sequencing. The transcriptome data were used to support genome annotation. Firstly, total RNA was extracted from multiple tissues of the same individual, including muscle, gill, skin, heart and kidney, using a commercial RNA isolation kit. RNA integrity was assessed using the Agilent 2100 Bioanalyzer to ensure RIN values above 7. Then the pooled RNA-seq library was constructed and sequenced on the Illumina NovaSeq 6000 platform, generating 7.17 Gb of raw data (Table 1). The raw reads were subsequently processed with fastp (v0.20.1) to remove adapter sequences and low-quality bases, yielding 6.92 Gb of clean reads, which is used for downstream gene annotation (Table 1).

Genome assembly. Raw Illumina reads were initially processed using fastp[8] (v0.20.1) with default parameters to remove adapter sequences and filter out low-quality bases, generating 65.43 Gb of clean reads suitable for downstream analyses (Table 1). Genome size estimation was performed by k-mer frequency analysis with Jellyfish[9] (v2.3.0), using a 21-mer. The resulting k-mer distribution was subsequently modeled with GenomeScope[10] (v2.0) to estimate genome size of 547.12 Mb, based on the Illumina short-read dataset of 65.43 Gb (Fig. 2).

De novo assembly was carried out with hifiasm[11] (v0.19.6), which uses PacBio HiFi long reads to construct contig-level genome. The resulting draft assembly was subsequently purged of redundant sequences using purge_haplotigs[12] (v1.0.4), leveraging read depth distribution and pairwise alignment to identify and remove heterozygous duplications. Raw Hi-C reads were initially processed using fastp (v0.20.1) with default parameters to remove adapter sequences and filter out low-quality bases, generating 71.64 Gb of clean reads suitable for scaffolding (Table 1). Those clean Hi-C reads were mapped to the draft assembly using Bowtie2[13] (v2.3.4.3) with default parameter set. A Hi-C contact matrix was generated with HiCUP[14] (v0.7.2) and used as input for the 3D-DNA pipeline[15] (v180922) to cluster, order, and orient contigs into chromosomes. To further improve assembly accuracy, manual curation was performed with Juicebox[16] (v1.11.08), enabling the correction of potential mis-assemblies through inspection of Hi-C contact maps (Fig. 3a). The final *C. japonicus* genome assembly comprised 617.85 Mb, with a scaffold N50 of 23.43 Mb and a scaffold N90 of 12.32 Mb (Fig. 4). A total of 98.51% of assembled sequences were anchored to 28 chromosomes (Fig. 3a, Table 2). These assembly statistics indicate a high level of completeness and accuracy, providing a robust genomic framework for downstream analyses.

Telomeric sequences were identified in the *C. japonicus* genome using the quarTeT package[17] (v.1.1.1). In total, 29 telomeres were detected across the assembled chromosomes (Fig. 3b). Among these, 6 chromosomes contained telomeric sequences at both ends, while 17 chromosomes harbored telomeric repeats at only one end. Notably, 5 chromosomes lacked detectable telomeric sequences. These results suggest that the majority of chromosome termini in the assembly were successfully captured, further supporting the high completeness and structural integrity of the genome assembly.

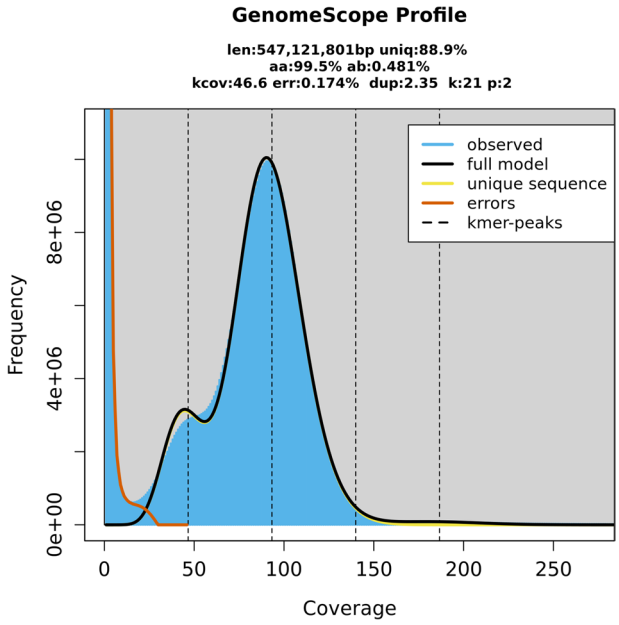

Fig. 2 Genome size estimates for *C. japonicus* using kmer-based (k = 21) method.

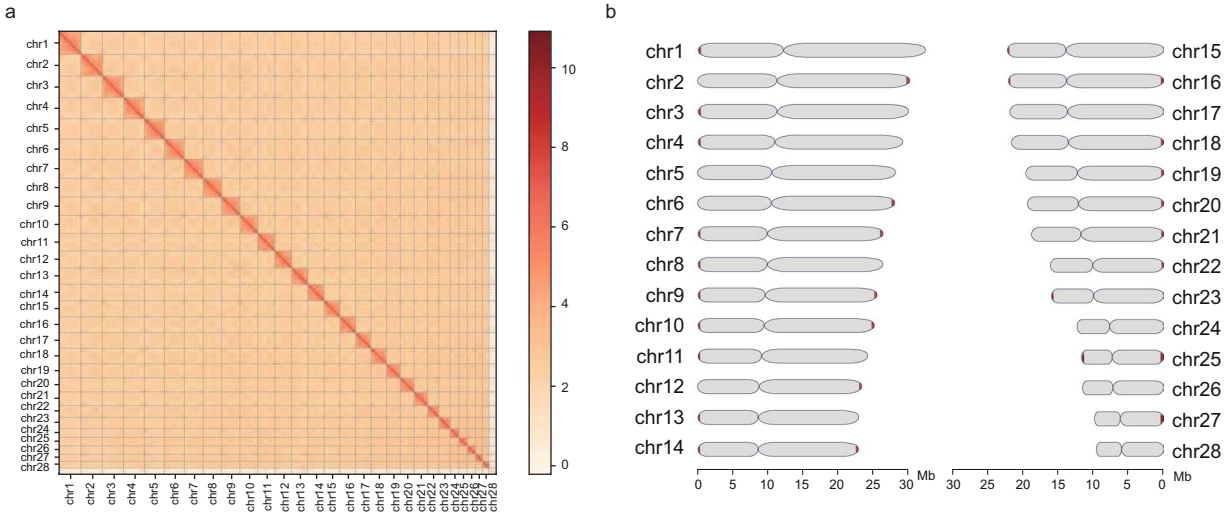

Fig. 3 Overview of chromosome structure of *C. japonicus* genome. (**a**) The Hi-C heatmap illustrates the interaction frequencies among various chromosomes in *C. japonicus* genome. (**b**) Overview of telomeres identified in the genome, with chromosome ends marked in red indicating the presence of telomeres, while the positions of the centromeres are depicted schematically and are not based on actual data.

Genome annotation. Repeat, including tandem repeat and transposon elements, annotation was conducted through a combination of *de novo* prediction and homology-based searches. Tandem repeats were detected with Tandem Repeat Finder[18] (TRF, v4.09). For *de novo* identification, LTR_FINDER_parallel[19] (v.1.0.7) and RepeatModeler[20] (v.1.0.11) were employed to generate a repeat library specific to *C. japonicus*, further predicted novel repeat with RepeatMasker[21] (v.4.0.9). Homology-based searches were then performed against Repbase (20181026) repeat databases, and the identified elements were subsequently masked across the genome using RepeatMasker (v.4.0.9) and RepeatProteinMask (v.4.1.0). In total, repetitive sequences accounted for 38.74% of the genome (Table 3), with DNA transposons (DNA) being the most abundant class (26.56%), followed by long interspersed nuclear elements (LINEs) retrotransposons (8.63%), long terminal repeat (LTR) retrotransposons (5.99%), and short interspersed nuclear elements (SINEs) retrotransposons (0.36%).

Gene prediction was performed with repeat-masked genome, which integrates ab initio prediction, transcriptome evidence and protein homology information. Ab initio gene prediction was performed using AUGUSTUS[22] (v.3.3.2) and GenSCAN[23] with closely related species models according to their taxonomic status. These protein sequences of (*Notothenia coriiceps*, *Danio rerio*, *Cottoperca gobio*, *Perca flavescens*)

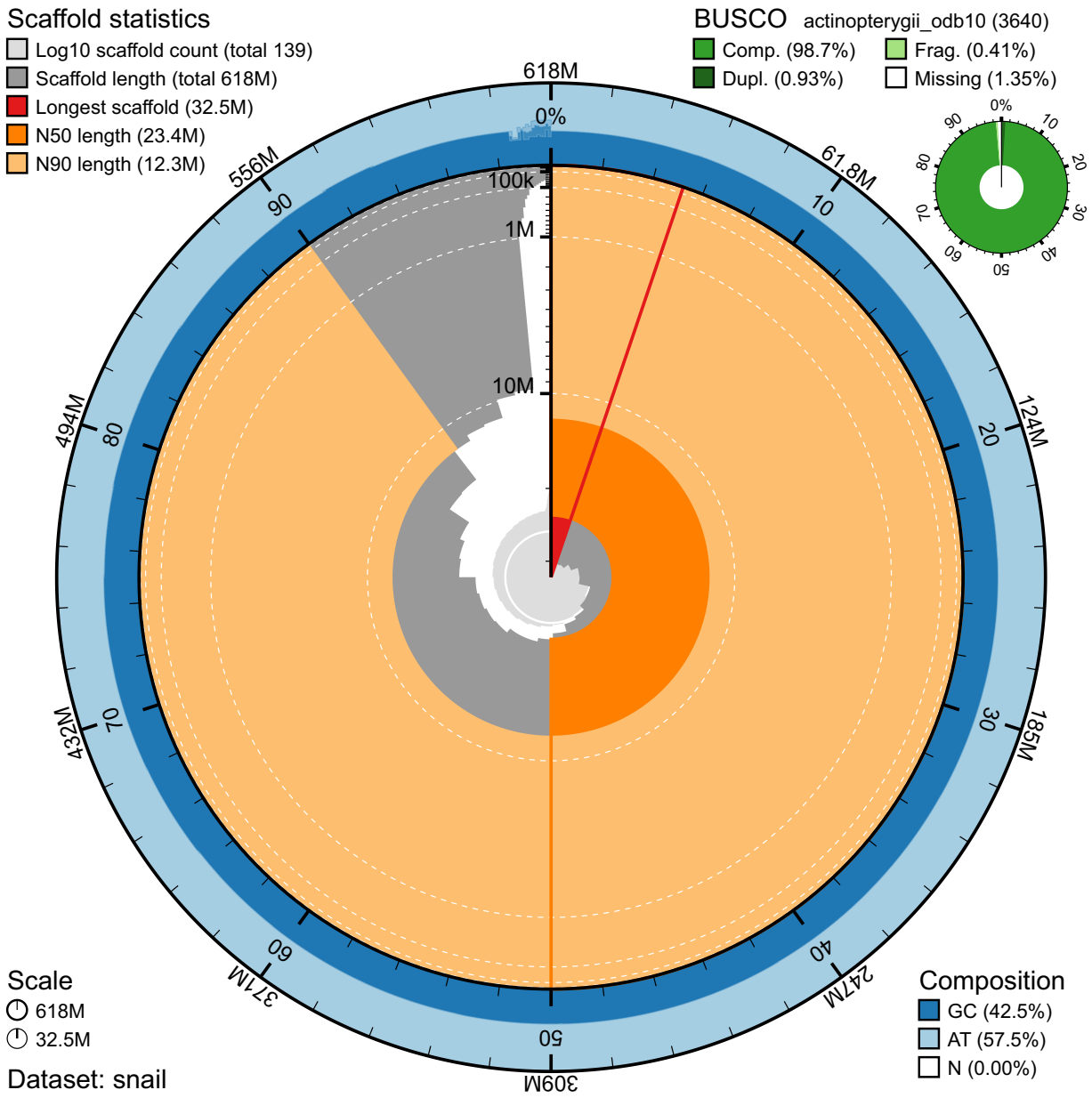

Fig. 4 Snail plot of *C. japonicus* genome. The BlobToolKit snail plot displays the scaffold N50 metric and BUSCO gene completeness. The main plot is divided into 1,000 size-ordered bins, with each bin representing 0.10% of the genome assembly. The gray area shows the distribution of sequence lengths, and the radius of the plot is scaled to the longest sequence in the assembly (32.5 Mb, highlighted in red). The orange and light orange arcs represent the scaffold N50 and N90 sequence lengths (23.4 Mb and 12.3 Mb, respectively). The light gray spiral displays the cumulative sequence count on a logarithmic scale, with white tick marks indicating successive orders of magnitude. The outer blue and light blue regions illustrate the distribution of GC, AT, and N percentages within the same bins as the inner plot. In the upper right corner, a summary of BUSCO genes from the actinopterygii_odb10 dataset is shown, including complete, fragmented, duplicated, and missing BUSCOs.

Total length (bp)	Length of chromosome	Anchor ratio	Contig N50 length (bp)	Contig N90 length (bp)	Maximum length (bp)	GC content (%)
617,853,943	608,623,247	98.51%	9,701,332	1,504,129	32,466,121	42.53

Table 2. The statistics of genome assembly of *C. japonicus*.

were downloaded from the NCBI database and subjected to alignment with the genome via tblastn[24] (E-value ≤ 1e-5), and following the exonerate[25] (v2.2.0) was used to deal with the alignment results and generate

Type	Repbase TEs	TE protiens	De novo	Combined TEs	
	Length (bp)	Length (bp)	Length (bp)	Length (bp)	% in genome
DNA	51,245,569	4,608,249	132,854,489	164,084,947	26.56
LINE	20,621,384	11,802,811	40,115,157	53,312,801	8.63
SINE	1,454,901	0	1,208,324	2,216,743	0.36
LTR	14,061,436	3,283,828	25,273,910	36,995,987	5.99
Satellite	5,282,466	0	4,350,203	9,135,203	1.48
Simple repeat	0	0	39	39	0
Other	4,090	282	0	4,372	0
Unknown	562,987	7,878	20,309,075	20,866,477	3.38
Total	81,054,075	19,691,641	196,887,100	239,386,482	38.74

Table 3. Repeat annotation statistics of *C. japonicus*.

	Number	Percent (%)
Total	22165	100
Annotated	21809	98.39
InterPro	19864	89.62
GO	15175	68.46
KEGG_ALL	21428	96.67
KEGG_KO	13945	62.91
Swissprot	19290	87.03
TrEMBL	21677	97.8
Pfam	19156	86.42
NR	21773	98.23
KOG	17799	80.3
Unannotated	356	1.61

Table 4. The summary of gene functional annotation of *C. japonicus*.

the homolog-based gene models. RNA-seq data, generated from a pooled five tissues, was aligned to the assembly using HISAT2[26] (v.2.1.0), and putative transcripts were detected and predicted by StringTie[27] (v.1.3.5). All gene models underwent merging to eliminate redundancy using MAKER2[28] (v.2.31.10) and HiCESAP (Wuhan OneMore Tech Co., Ltd., https://www.onemore-tech.com/) with default settings. Finally, a total of 22,165 protein-coding genes were predicted.

Functional annotation was achieved through searches against multiple databases, including InterPro[29] (v.93.0, 2023-03-02), GO (2023-04-01), KEGG[30] (v.105.0, 2023-01-01), Swiss-Prot[31] (2023-03-01), TrEMBL[31] (2023-03-01), Pfam (v.14.1), NR[32] (2023-04-01), and KOG (2003-03-01, https://ftp.ncbi.nih.gov/pub/COG/KOG/). A total of 22,165 sequences were annotated, with 21,809 (98.39%) successfully annotated and 356 (1.61%) unannotated. Databases with the highest coverage included NR (21,773 sequences, 98.23%), TrEMBL (21,677 sequences, 97.80%), and KEGG_ALL (21,428 sequences, 96.67%); InterPro (89.62%), Swissprot (87.03%), Pfam (86.42%), and KOG (80.30%) also showed strong performance. Specialized annotations covered GO (15,175 sequences, 68.46%), and KEGG_KO (13,945 sequences, 62.91%) (Table 4). These functional annotations fully confirm the reliability of the above gene annotation results.

In addition to protein-coding genes, a comprehensive annotation of non-coding RNAs (ncRNAs) was performed. Specifically, the ribosomal RNA (rRNA), microRNA (miRNA) and small nuclear RNA (snRNA) genes were predicted using the INFERNAL based on the rfam[33] (v1.4.8) and miRBase[34] databases, respectively. Transfer RNA (tRNA) genes were identified using the tRNAscan-SE[35] (v.1.3.1) software. This analysis identified 886 microRNAs (miRNAs), 1,344 ribosomal RNAs (rRNAs), and 2,314 transfer RNAs (tRNAs), which contribute to gene regulation and essential cellular processes. These results provide a complete view of both coding and non-coding elements in the *C. japonicus* genome, establishing a valuable reference for further functional and evolutionary studies.

Data Records

The complete dataset for *C. japonicus*, including raw sequencing data and the assembled genome, is available through the National Center for Biotechnology Information (NCBI) and figshare. The raw data comprising Illumina short reads, PacBio HiFi, Hi-C, and RNA sequencing data have been deposited in the NCBI Sequence Read Archive (SRA) under accession numbers SRR21530970[36], SRR35386379[37], SRR35182048[38], and SRR35182047[39]. The genome assembly has been deposited in NCBI GenBank under the WGS accession number JBSGKO000000000[40]. Additionally, the chromosome-level genome assembly of *C. japonicus*, along with its predicted coding sequences and protein sequences, is available at figshare[41].

Technical Validation

The high completeness of BUSCO genes and the high read mapping rate demonstrate the high quality of the assembly. BUSCO (v.5.2.2)[42] analysis, using the actinopterygii_odb10 dataset, showed that 98.65% of the BUSCO genes were complete, among which 97.72% were single-copy and 0.93% were duplicated. The fragmented BUSCO genes accounted for 0.41%, and the missing ones were 1.35%. Moreover, the assembly quality value (QV) was quantified using Merqury[43] (v1.4), resulting in a QV score of 45.26, reflecting a high-quality assembly. Hi-C contact maps clearly showed strong diagonal signals (Fig. 3), indicating the high accuracy of the chromosome-level assembly. The mapping rate of PacBio HiFi and Illumina short reads to the assembled genome was 100% and 99.29%, respectively, further validating the assembly quality.

Data availability

The whole-genome shotgun assembly has been deposited at DDBJ/ENA/GenBank under accession JBSGKO000000000 (version JBSGKO000000000). Raw sequencing reads are available in the NCBI SRA under BioProject PRJNA1308732 and PRJNA879413 (PacBio HiFi: SRR35386379; Illumina short reads: SRR21530970; Hi-C: SRR35182048; RNA-seq: SRR35182047). The genome annotation files (GFF3, FA) are accessible through figshare (https://doi.org/10.6084/m9.figshare.30031102.v2). All datasets are publicly available without restriction.

Code availability

All analyses were carried out using standard bioinformatics software. No custom code was used for this study.

Received: 10 September 2025; Accepted: 11 February 2026;
Published online: 03 March 2026

References

1. Nelson, J. S. *Fishes of the World*. 5th edn. Wiley (2016).
2. Froese, R. & Pauly, D. (eds). *FishBase*. www.fishbase.org (2024).
3. Pauly, D. & Zeller, D. Catch reconstructions reveal that global marine fisheries catches are higher than reported and declining. *Nat. Commun.* **7**, 10244 (2016).
4. Cheung, W., Watson, R. & Pauly, D. Signature of ocean warming in global fisheries catch. *Nature* **497**, 365–368 (2013).
5. Liu, L., Liu, Q. & Gao, T. Genome-wide survey reveals the phylogenomic relationships of *Chirolophis japonicus* Herzenstein, 1890 (Stichaeidae, Perciformes). *ZooKeys* **1129**, 55–72 (2022).
6. Conte, M. A. *et al*. A high quality assembly of the Nile Tilapia (*Oreochromis niloticus*) genome reveals the structure of two sex determination regions. *BMC Genom.* **18**, 341 (2017).
7. Peichel, C. L. & Marques, D. A. The genetic and molecular architecture of phenotypic diversity in sticklebacks. *Philos. Trans. R. Soc. B Biol. Sci.* **372**, 20150486 (2017).
8. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
9. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
10. Ranallo-Benavidez, T. R. *et al*. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).
11. Cheng, H. *et al*. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
12. 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).
13. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
14. Wingett, S. *et al*. HiCUP: pipeline for mapping and processing Hi-C data. *F1000Research* **4**, 1310 (2015).
15. Dudchenko, O. *et al*. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
16. Durand, N. C. *et al*. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst.* **3**, 99–101 (2016).
17. Lin, Y. *et al*. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Hortic. Res.* **10**, uhad127 (2023).
18. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
19. Ou, S. & Jiang, N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mob. DNA* **10**, 48 (2019).
20. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA.* **117**, 9451–9457 (2020).
21. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinforma.* **Chapter 4**, 4.10.1–4.10.14 (2009).
22. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–439 (2006).
23. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78–94 (1997).
24. Camacho, C. *et al*. BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
25. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
26. Kim, D. *et al*. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
27. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
28. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
29. Blum, M. *et al*. InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Res.* **53**, D444–D456 (2024).
30. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
31. Coudert, E. *et al*. Annotation of biologically relevant ligands in UniProtKB using ChEBI. *Bioinformatics* **39**, btac793 (2022).
32. Sayers, E. W. *et al*. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **49**, D10–D17 (2020).
33. Griffiths-Jones, S. *et al*. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, D121–124 (2005).
34. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. miRBase: from microRNA sequences to function. *Nucleic Acids Res.* **47**, D155–D162 (2019).

35. 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).
36. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR21530970 (2025).
37. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35386379 (2025).
38. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35182048 (2025).
39. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35182047 (2025).
40. *NCBI GenBank* https://identifiers.org/ncbi/insdc:JBSGKO000000000 (2025).
41. Liu, K. Genome Atlas of *Chirolophis japonicus*: Chromosomal Assembly, Annotation and Evolutionary Insights. *Figshare* https://doi.org/10.6084/m9.figshare.30031102.v2 (2025).
42. 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).
43. Rhie, A. *et al*. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol.* **21**, 245 (2020).

Acknowledgements

This work was supported by the Foundation of State Key Laboratory of Mariculture Biobreeding and Sustainable Goods (No. BRESG202403). We thank Wuhan Onemore Tech Co., Ltd. for their assistance with genome sequencing and analysis.

Author contributions

T.X.G. designed this project. Y.Q.Q. collected the samples. Q.K.L. and Q.L. analyzed the data. Q.K.L. and Q.L. wrote the manuscript and T.X.G. revised the manuscript. All authors read and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to T.G.

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

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