scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level Genome assembly and annotation of the *Chirolophis japonicus*

Hongyu Pan[1,4], Hongwei Yan[2,3,4], Xiandong Hu[2], Jingqi Wang[2], Yuying Zhang[1] & Qi Liu[1,3] ✉

With the advancement of molecular biology and sequencing technologies, the genomes of an increasing number of fish species, ranging from model fish to many commercial species, have been sequenced and made publicly available. Thus, this has provided deep insights into accelerating our understanding of various aspects, including species evolution, parallel evolution, molecular regulatory mechanisms, and genetic breeding. *Chirolophis japonicus* is an economically important fish and is distributed in the cold waters of the northwest Pacific Ocean. This study provided a chromosome-level genome of *C. japonicus* using PacBio HiFi long-reads, Hi-C data and MGI short-reads. The *C. japonicus* genome was successfully assembled to a size of 623.67 Mb, with a contig N50 of 23.17 Mb, and a scaffold N50 of 23.77 Mb. It was anchored onto 28 chromosomes. The assembly completeness estimated by the BUSCO method was 99.42%. The repetitive elements accounted for 29.13% (118.67 Mb), primarily consisting of 26.64% transposable elements. We further predicted a total of 23,893 protein-coding genes, of which 23,550 were successfully annotated. In conclusion, the chromosome-level genome of *Chirolophis japonicus* will provide a valuable genomic resource for further aquaculture, genetic breeding and molecular mechanisms underlying its uniquely cold-tolerant traits.

Background & Summary

Chirolophis japonicus belongs to the order Perciformes, family Stichaeidae, and genus Chirolophis. It inhabits the shallow coastal waters of the northwest Pacific Ocean, including the Yellow Sea, Bohai Sea, northern Japanese waters, the Sea of Okhotsk, and the Bering Sea[1]. *C. japonicus* is an omnivorous, cold-temperate fish species, and is most commonly found along rock edges and in tide pools. Research on this species is fairly limited, primarily comprising studies on its distribution areas, reproductive biology[2], life history[3], the origin of head cortical processes[4] and phylogenomic relationships[5]. Due to its delicious flavor and high lipid content, this fish possesses significant market value. However, wild stocks have been limited recently, making aquaculture a promising venture with substantial market potential for the future.

A high-quality reference genome can significantly enhance our understanding of the genetic basis and evolutionary processes underlying the biological characteristics of *C. japonicus*. On the other hand, complete genome data will greatly benefit the following aquaculture process. We integrated multiple sequencing technologies, including PacBio Revio's Circular Consensus Sequencing (CCS), Hi-C sequencing and DNBseq sequencing. The final assembled *C. japonicus* genome size is 623.67 Mb, with 99.17% of contigs anchored onto 28 chromosomes. The contig N50 length is 23.17 Mb, and the scaffold N50 length is 23.77 Mb. The genome contains 29.13% repetitive sequences. We predicted a total of 23,893 protein-coding genes were predicted, of which 23,550 (98.56%) were functionally annotated., reflecting the high quality of the assembled genome. These comprehensive genomic resources will facilitate future research on the molecular and adaptive mechanisms of *C. japonicus*.

[1]College of Marine Science and Environment Engineering, Dalian Ocean University, 116023, Dalian, China. [2]College of Fisheries and Life Science, Dalian Ocean University, 116023, Dalian, China. [3]Key Laboratory of Environment Controlled Aquaculture Dalian Ocean University, Ministry of Education, 116023, Dalian, China. [4]These authors contributed equally: Hongyu Pan, Hongwei Yan. ✉e-mail: liuqisunson@163.com

Fig. 1 Mature female *Chirolophis japonicus*.

Library type	Tissue	Raw data (Gb)	Clean data (Gb)	Average read length (bp)
DNBSEQ	Muscle	83.86	80.86	150
PacBio HiFi	Muscle	94.91	94.91	17,201.75
Hi-C	Muscle	68.94	64.71	150

Table 1. Statistics of sequencing data for *C. japonicus* genome assembly and annotation.

Methods

Ethics statement. This study is reported in accordance with the ARRIVE guidelines. All experimental protocols were approved by the Animal Care and Use Committee of Dalian Ocean University in Dalian, China, and all methods were carried out in accordance with the Guidelines for the Care and Use of Laboratory Animals and relevant regulations.

Fish sample collection and preparation. One healthy, mature female *C. japonicus* specimen was collected from the Longwangtang Fishing Port of Dalian, Liaoning Province, China (38.8233314°N, 121.403927°E) (Fig. 1). The experimental fish measured 27.6 cm in total length and weighed 217.54 g. Multiple tissues (muscle, liver, gill, kidney, lateral line skin and abdominal skin) were collected from the same individual for further transcriptome sequencing.

DNA extraction and genome sequencing. To prepare the genomic library, high-quality genomic DNA was isolated from muscle tissue using the sodium dodecyl sulfate (SDS) method. The DNA was then purified and quantified. For the short-read DNBseq sequencing, paired-end sequencing by the DNBSEQ-T7 sequencing platform (MGI) was used to generate a total of 83.86 Gb of raw reads (Table 1). Reads were filtered using Fastp (v 0.24.0) with default parameters to remove problematic reads (e.g., low quality, too short, or with too many uncalled bases), and to trim adapters and polyG sequences, retaining 80.86 Gb of clean data[6]. Subsequently, genome size and heterozygosity were estimated by K-mer analysis via GCE(v 1.0.0) software[7]. A k-mer ($k = 17$) distribution analysis was conducted using the clean reads, resulting in an estimated genome size of C. japonicus of approximately 586 Mb, corrected to 583 Mb, with a heterozygosity rate of 0.42% and a repetitive sequence proportion of 30.12% (Fig. 2).

For PacBio's Circular Consensus Sequencing (CCS), libraries were constructed using the Pacific Biosciences SMRTbell Prep Kit 3.0. Fragments of 10 k - 50 k were selected using the PippinHT system. The library was combined with sequencing primers using the Revio Polymerase Kit, and subsequently bound to the sequencing enzyme. Finally, the diluted library was loaded onto the sequencing plate and sequenced on the PacBio Revio platform. After eliminating low-quality reads, a total of 94.91 Gb of reads with an average length of 17.201 kb were retained. Using these PacBio long-reads, an initial contig assembly was generated using Hifiasm (v 0.19.6) with the default algorithm (https://github.com/chhylp123/hifiasm). Subsequently, purge haplotigs (v1.0.4) was used to eliminate redundant sequences[8]. This process yielded a contig-level assembly of approximately 623.67 Mb (623,668,508 bp ≈ 623.67 Mb), consisting of 170 contigs, with an N50 and maximum contig sizes of 23.17 Mb and 30.52 Mb, respectively (Table 2).

Hi-C library preparation, sequencing, and chromosome-level assembly. To create a chromosome-level genome for *C. japonicus*, a Hi-C library was generated from muscle tissue. The library was prepared according to standard protocol: cells were treated with formaldehyde to fix DNA conformation, then biotin-labeled. After crosslink reversal, DNA was purified and randomly sheared into 300–500 bp fragments. Subsequently, junction fragments were captured using streptavidin magnetic beads and sequenced on the DNBSEQ-T7 platform (MGI, Shenzhen, China). A total of 68.94 Gb of Hi-C reads were obtained (Table 1). After filtering reads with an average quality score below 20 and removing adapters with fastp (v 0.20.0) under default parameters, 64.71 Gb of clean data were retained (Table 1). We also used HICUP (v0.7.2) to remove erroneous alignments and duplicate contigs, generating an interaction matrix[9]. This matrix served as the basis for anchoring contigs to chromosomes using 3D-DNA (v180,922) with the optimized parameter "-r0" (https://github.com/theaidenlab/3d-dna). Scaffolds were manually evaluated and refined using Juicebox Assembly Tools (v1.11.08) to correct any chromosomal translocations and inversions[10]. By integrating this Hi-C data during the assembly

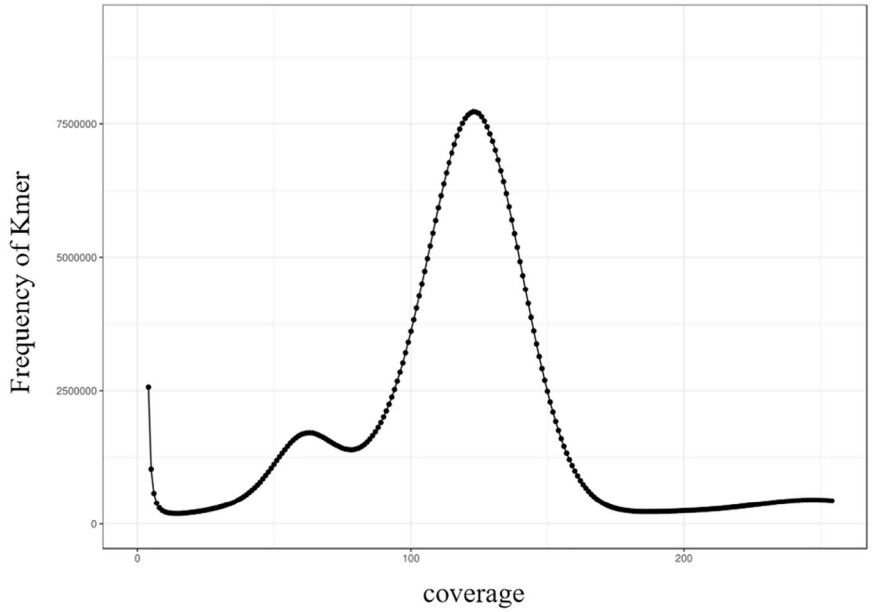

Fig. 2 K-mer distribution map (k = 17) of the *C. japonicus* genome based on the MGI short-reads.

	Contig Length (bp)	Contig Number	Scaffold Length (bp)	Scaffold Number
N90	9,708,868	27	9,708,868	27
N80	16,246,819	22	16,246,819	22
N70	19,366,678	18	19,366,678	18
N60	21,946,249	15	21,946,249	15
N50	23,168,778	13	23,168,778	13
Total length	623,668,508	—	623,668,508	—
Number(>=100 bp)	—	170	—	170
Number(>=2 kb)	—	170	—	170
Max length	30,517,861	—	30,517,861	—

Table 2. Sequencing data used for the *C. japonicus* genome assembly.

correction process, the original 170 contigs were fragmented at misassembly points based on the interaction map, ordered, and ultimately assembled into 28 chromosomes and 104 scaffolds, with a total length of 0.62 Gb. The contig N50 was 23.17 Mb, the scaffold N50 was 23.77 Mb, and the chromosome anchoring rate reached 99.17% (Fig. 3).

RNA extraction and transcriptome sequencing(RNA-seq). Total RNA was extracted from six tissues (muscle, liver, gill, kidney, lateral line skin and abdominal skin) separately according to a standard Trizol protocol (Invitrogen, Frederick, MD, USA) method, RNA purity was checked using the Nanodrop (OD 260/280), RNA concentration was measured using Qubit RNA Assay Kit in Qubit2.0 Flurometer (Life Technologies, CA, USA), RNA integrity was assessed using the Agilent 2100 (Agilent Technologies, Santa Clara, CA, USA). Only those RNA samples with OD 260/280 $\geq$ 1.8 and RNA integrity number $\geq$ 7.0 were selected for transcriptome sequencing. Six RNA samples were initially sequenced on a DNBSEQ-T7 platform, respectively (MGI, BGI Shenzhen, China). Then RNA from those six different tissues was mixed into a total RNA, which was also sequenced on a PromethION sequencer (Oxford Nanopore Technologies, Oxford, UK) for full-length transcriptome sequencing. Finally, a total of 165.01 Gb RNA-seq data were generated for assistance in gene annotations

Repeat sequence annotation. The annotation of repetitive sequences involved a comprehensive approach combining homology-based searches against known repeat databases and *de novo* prediction methods, using TRF, RepeatMasker, and RepeatProteinMask. Tandem repeats within the genomic sequences were identified using TRF (v4.09)[11] with parameters "2 7 7 80 10 50 2000 -d -h". Homology-based prediction used RepeatMasker (v4.0.9) with the optimized parameter"-nolow -no_is -norna", and its internal RepeatProteinMask to identify repeat features at the DNA and protein levels based on the RepBase database. *De novo* prediction used a non-redundant repeat library generated by RepeatModeler (v1.0.11)[12] and LTR-FINDER (v1.0.7)[13] to annotate repeats in the genome using RepeatMasker (v4.0.9). Finally, after combining the results from the three strategies above and removing overlapping regions, we identified 181,668,016 bp of repetitive sequences, accounting for 29.13% of the assembled genome (Table 3). The specific classification results were shown in Table 4.

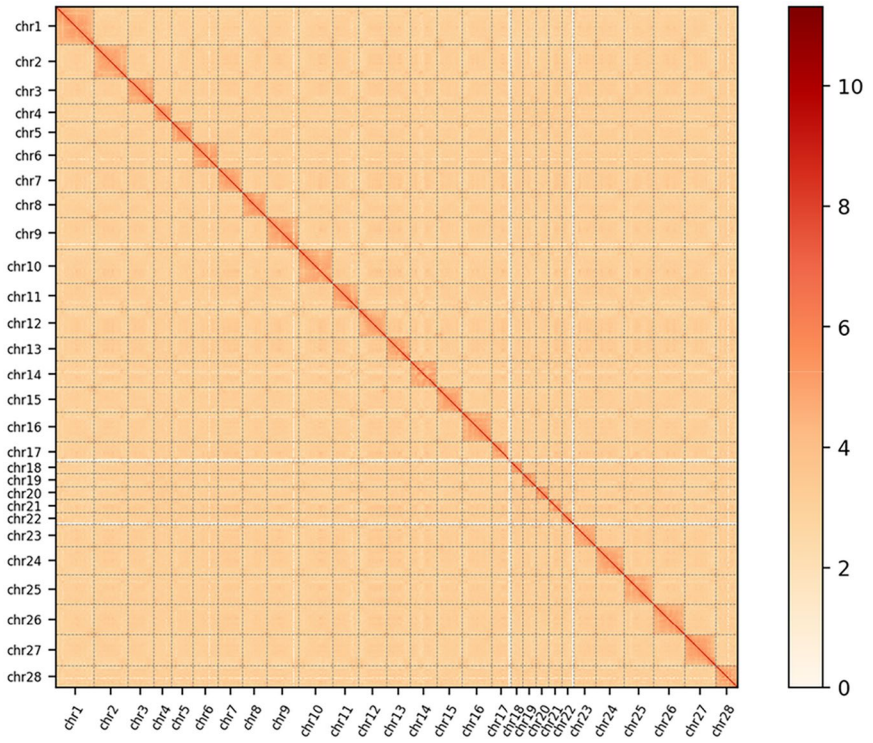

Fig. 3 *C. japonicus* genome contig contact matrix using Hi-C data. Chr1–28 correspond to the 28 pseudo chromosomes. The depth of the red color indicates the contact density.

Type	Repeat Size (bp)	% of genome
Trf	43,898,397	7.04
Repeatmasker	82,279,024	13.19
Proteinmask	19,562,248	3.14
De novo	108,821,414	17.45
Total	181,668,016	29.13

Table 3. Repeat sequence statistics.

	RepBase TEs		TE Proteins		*De novo*		Combined TEs	
	Length (bp)	%in Genome	Length (bp)	%in Genome	Length (bp)	%in Genome	Length (bp)	%in Genome
DNA	51,942,116	8.33	4,604,340	0.74	47,768,988	7.66	83,710,932	13.42
LINE	20,498,910	3.29	11,540,238	1.85	11,736,033	1.88	27,584,598	4.42
SINE	1,525,101	0.24	0	0.00	2,273,641	0.36	3,136,779	0.50
LTR	14,336,267	2.30	3,420,199	0.55	6,746,517	1.08	19,441,604	3.12
Satellite	5,645,713	0.91	0	0.00	650,262	0.10	6,250,387	1.00
Simple_repeat	0	0.00	0	0.00	66,998	0.01	66,998	0.01
Other	3,716	0.00	234	0.00	0	0.00	3,950	0.00
Unknown	563,648	0.09	10,389	0.00	41,853,397	6.71	42,272,068	6.78
Total	82,279,024	13.19	19,562,248	3.14	108,821,414	17.45	166,119,699	26.64

Table 4. Summary of transposable elements (TEs) annotation in *C. japonicus* genome.

Protein-coding gene identification and annotationgene. First, AUGUSTUS (v3.3.2)[14], Genscan (v1.0)[15], and GlimmerHMM (v3.0.4)[16] were used to perform ab initio gene prediction on a repeat-masked genome sequence. Secondly, miniprot (v0.11-r234)[17] and Liftoff (v1.6.3)[18] were used for homology-based prediction against closely related species. We aligned homology proteins from six representative fish species, including *Zoarces viviparus* (viviparous blenny, GCA_040110945.1), *Gasterosteus aculeatus* (three-spined stickleback, GCA_016920845.1), *Pungitius pungitius* (ninespine stickleback, GCA_949316345.1), *Cyclopterus*

	Gene set	Protein-coding gene number	Average gene length (bp)	Average CDS length (bp)	Average exon per gene	Average exon length (bp)	Average intron length (bp)
De novo	Genscan	29,371	14,641	1,605	9.02	177.98	1,626
	AUGUSTUS	29,861	10,334	1,412	7.94	177.83	1,286
Homolog	*Zoarces viviparus*	51,542	6,310	1,061	5.56	190.63	1,150
	Anoplopoma fimbria	33,504	14,548	1,868	10.88	171.61	1,283
	Gasterosteus_aculeatus	47,962	18,489	2,147	12.35	173.77	1,440
	Pungitius pungitius	42,927	18,984	2,218	12.64	175.57	1,441
	Cyclopterus lumpus	40,476	15,446	2,042	11.88	171.93	1,232
	Pseudoliparis swirei	41,124	17,637	2,094	12.19	171.77	1,389
Liftoff	*Zoarces viviparus*	24,195	9,554	1,536	8.86	173.32	1,020
	Gasterosteus_aculeatus	20,710	11,483	1,595	8.88	179.53	1,254
Trans ORF	RNAseq	15,517	18,082	1,945	12.11	362.13	1,233
	ISOseq	8,011	9,976	1,392	9.65	243.38	881.90
BUSCO		3,660	9,184	1,655	10.40	159.02	800.65
MAKER		24,157	15,231	1,618	10.27	338.60	1,268
Integrated		23,893	14,059	1,753	10.23	270.85	1,223

Table 5. Basic statistics of gene prediction.

lumpus (lumpfish, GCA_009769545.1), *Anoplopoma fimbria* (sablefish, GCA_027596085.2), *Pseudoliparis swirei* (GCA_029220125.1), which were downloaded from the NCBI.

Thirdly, prediction was conducted using transcriptomic RNA-seq reads. HISAT2 (v2.1.0)[19] was used for data alignment, and StringTie (v1.3.5)[20] was applied for transcript prediction. ISOseq3 was employed to obtain full-length transcripts from PacBio sequencing, and TransDecoder (v5.5.0) was utilized for coding region prediction. Fourthly, BUSCO (v5.7.1)[21] was used to perform prediction based on AUGUSTUS (v3.3.2)[22] and a conserved homologous gene database selected according to species classification. Subsequently, MAKER2 (v2.31.10)[23] was used to integrate the gene sets predicted by the above methods into a non-redundant, more comprehensive gene set. Finally, by combining all the above results, a high-quality final gene set of 23,893 protein-coding genes was annotated. The average length of these genes is 14,059 bp, the average coding sequence (CDS) length is 1,753 bp, and each gene contains an average of 10.23 exons (Table 5). A comparison of the statistical characteristics of gene set elements revealed similar distribution patterns between *Chirolophis japonicus* and the six other related species (Fig. 4).

To annotate the functional roles of these protein sequences, we employed primarily two methods: first, based on sequence similarity, the protein sequences of the annotated gene set were searched against protein databases such as TrEMBL, SwissProt, NR, KOG/COG, and KEGG using the diamond software (v2.0.14)[24]. Simultaneously, KOBAS software (v3.0)[25] was used to correlate the annotated KEGG information with KEGG ORTHOLOGY and PATHWAY pathways. Second, based on domain or motif similarity, InterProScan (v5.61-93.0)[26] was used to search a series of sub-databases within InterPro to obtain information on conserved protein sequences, motifs, and domains. HMMER3 (v3.3.1)[27] was used based on multiple sequence alignment and hidden Markov models to annotate transcription factors, Pfam domains, motifs, and other conserved sequence information. Ultimately, we obtained valid annotations for 23,550 genes in these databases (approximately 98.56% of the total predicted genes) (Table 6).

Non-coding RNA annotation. During non-coding RNA annotation, different methods were employed to predict and annotate based on the characteristics of various RNA types. Using the covariance models of Rfam families, the INFERNAL bundled with Rfam (v14.8)[28] was used to annotate miRNA and snRNA sequences in the genome. Due to the high conservation of rRNA, rRNA sequences from closely related species were selected as reference sequences, and BLASTN (v2.11.0+)[29] was used to search for rRNA in the genome. Based on tRNA's structural features, tRNAscan-SE (v1.3.1)[30] was used to search for tRNA sequences in the genome. In total, annotations predicted 951 miRNAs, 2,437 tRNAs, 2,088 rRNAs and 2,451 snRNAs (Table 7).

Data Records

The complete genomic data generated in this study have been deposited in public repositories as follows: Whole-genome sequencing raw reads and transcriptome data are available under NCBI BioProject PRJNA1355350[31]. The raw sequencing reads are available under the accessions SRR35920331, SRR35920332[32,33] and SRR35942771[34]. While the RNA sequencing data can be found under accessions SRR35920325 to SRR35920330[35–40]. The final chromosome-level genome assembly has been deposited in GenBank under the accession number JBTOFL000000000[41] and and the comprehensive annotations are available at Figshare[42].

Technical Validation

To assess the quality of the *C. japonicus* genome assembly, we comprehensively evaluated its assembly performance by comparing the genome with the NT database, its short-read data, long-read data, and transcriptome data, and through methods such as BUSCO. To confirm that the assembly results belong to the target species, we fragmented the genome sequences into 10 kb intervals and compared them to the NCBI nucleotide database (NT

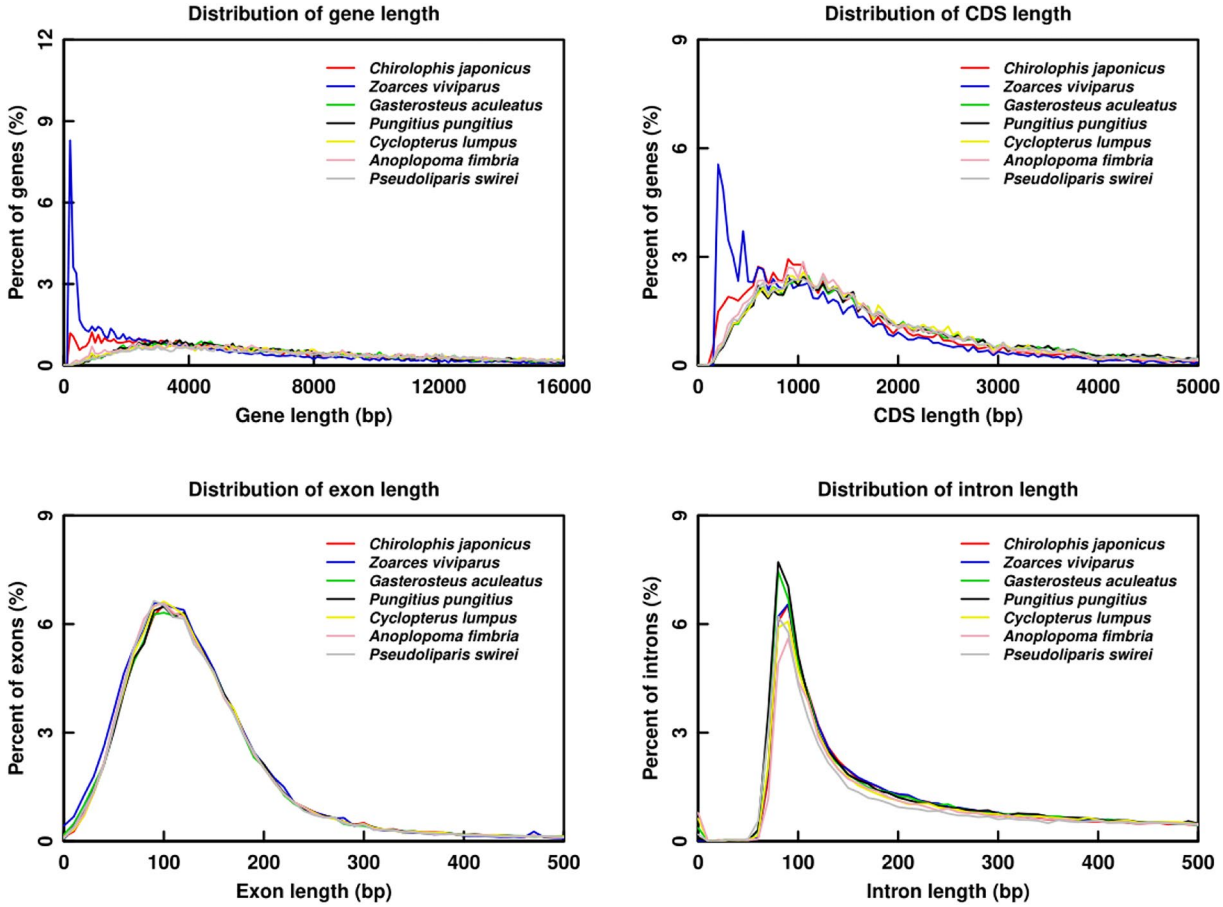

Fig. 4 Gene set component length distribution statistics (comparison with related species' gene components.

		Gene		mRNA	
		Number	Percent (%)	Number	Percent (%)
	NR	23,524	98.46	39,276	98.24
	SwissProt	20,616	86.28	34,692	86.78
	TrEMBL	23,427	98.05	39,146	97.92
	KOG	19,068	79.81	32,105	80.30
	TF	5,962	24.95	10,155	25.40
	InterPro	22,128	92.61	36,647	91.67
	GO	16,894	70.71	27,630	69.11
	KEGG_ALL	23,192	97.07	38,859	97.20
	KEGG_KO	16,557	69.30	28,221	70.59
	Pfam	20,968	87.76	34,170	85.47
Annotated		23,550	98.56	39,327	98.37
Unannotated		343	1.44	652	1.63
Total		23,893	100.00	39,979	100.00

Table 6. Functional annotation statistics.

database) using the Blast software. Based on the median percentage of identity in comparisons with sequences from the same genus reached 94.23%, the assembly error rate was determined to be low. Subsequently, we used the software BWA (v0.7.12-r1039)[43] and Minimap2 (v2.24-r1122)[44] to align the short-read and long-read data to the genome, respectively. The alignment rate and coverage for short-read data were 99.90% and 99.69%, respectively. In contrast, for long-read data, the consistency was 99.99% and 99.97%, indicating a high degree of consistency between the assembly results and the reads. Finally, a genomic circle graph was generated using Circos (v0.69-6)[45] with default parameters to summarize the evaluation results visually (Fig. 5). Additionally, the completeness of the genome was assessed using BUSCO (v5.7.1) and Compleasm (v0.2.6)[46]. The results showed

Type		Copy	Average length (bp)	Total length (bp)	Percentage of genome (%)
miRNA		951	87	82,493	0.013227
tRNA		2,437	75	183,514	0.029425
rRNA	rRNA	2,088	328	684,939	0.109824
	18S	260	1,767	459,501	0.073677
	28S	0	0	0	0.000000
	5.8S	244	154	37,490	0.006011
	5S	1,584	119	187,948	0.030136
snRNA	snRNA	2,451	148	361,948	0.058035
	CD-box	170	101	17,213	0.002760
	HACA-box	87	147	12,794	0.002051
	splicing	2,186	151	330,104	0.052929
	scaRNA	8	230	1,837	0.000295

Table 7. Non-coding RNA annotation statistics.

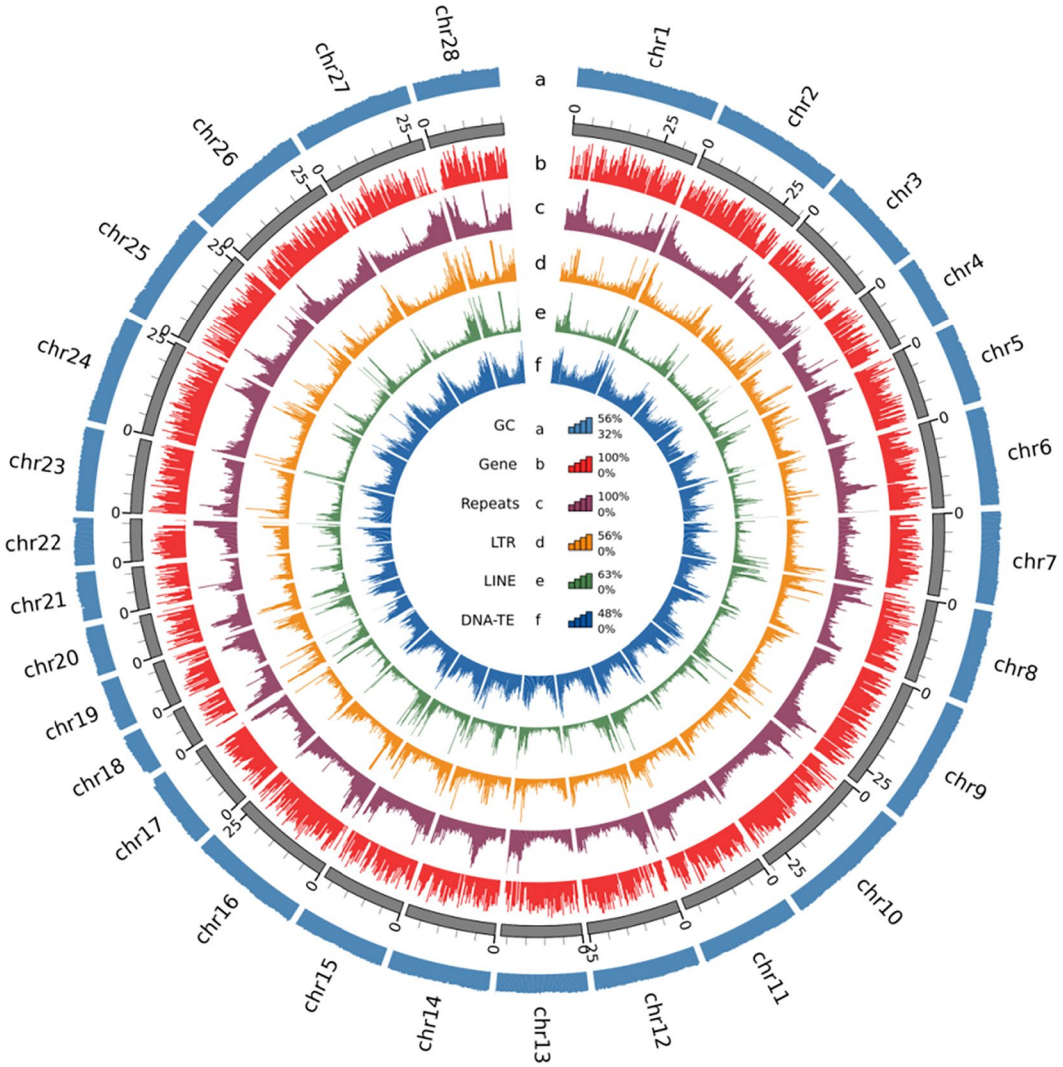

Fig. 5 Genome circle plot visualization.

that the proportion of BUSCO genes completely mapped to the genome used for annotation was 99.42% and 100.0%, respectively, while the evaluation result for the annotated gene set was 99.26% in both cases.

Data availability

All data supporting the findings of this study are publicly available. The sequencing data (MGI, PacBio, Hi-C, and Transcriptome) for *C. japonicus* can be found under NCBI BioProject PRJNA1355350, with individual SRA accessions SRR35920325 to SRR35920332 and SRR35942771. The chromosome-level genome assembly generated in this study is publicly accessible via GenBank (accession: JBTOFL000000000).The genome annotation files are available in the Figshare repository (https://doi.org/10.6084/m9.figshare.30461972).

Code availability

All data analyses were performed in accordance with the specified protocols and default parameters of the corresponding bioinformatics tools. The study adhered to these established workflows and did not involve the creation of any new scripts or custom code.

Received: 7 November 2025; Accepted: 15 April 2026;
Published online: 20 April 2026

References

1. Balanov, A. A., Epur, I. V. & Shelekhov, V. A. A Description of *Chirolophis japonicus* and Ch. saitone (Stichaeidae) Pelagic Larvae from Peter the Great Bay, *Sea of Japan*. *J. Ichthyol.* **60**, 364–374 (2020).
2. Chen, B. *et al*. Reproductive Biology of the Fringed Blenny (Azuma emmnion). *Fisheries Science* **37**, 539–543 (2018).
3. Shiogaki, M. On the life history of the stichaeid fish *Chirolophis japonicus*. *Japanese Journal of Ichthyology* **29**, 446–455 (1983).
4. Sato, M. Histological Observations on the Cutaneous Processes on the Head of *Azuma emmnion* and Hemitripterus villosus. *Japanese Journal of Ichthyology* **24**, 12–6 (2010).
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. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, 884–890 (2018).
7. Liu, B. H. *et al*. Estimation of genomic characteristics by analyzing K-mer frequency in de novo genome projects. *Quant. Biol.* **35**, 62–67 (2013).
8. 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).
9. Wingett, S. *et al*. HiCUP:pipeline for mapping and processing Hi-C data. *F1000Research* **4**, 1310 (2015).
10. Durand, N. C. *et al*. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst.* **1**, 99–101 (2016).
11. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research.* **27**, 573–580 (1999).
12. Price, A. L., Jones, N. C. & Pevzner, P. A. De novo identification of repeat families in large genomes. *Bioinformatics.* **21**, 351–3580 (2005).
13. Ou, S. & Jiang, N. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. *Mob DNA.* **10**, 48 (2019).
14. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, 435–439 (2006).
15. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol.* **268**, 78–94 (1997).
16. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics.* **20**, 2878–2879 (2004).
17. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics.* **6**, 31 (2005).
18. Shumate, A. & Salzberg, S. L. Liftoff: accurate mapping of gene annotations. *Bioinformatics.* **37**, 1639–1643 (2021).
19. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology.* **37**, 907–915 (2019).
20. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol.* **33**, 290–295 (2015).
21. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. 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).
22. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–439 (2006).
23. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics.* **12**, 491 (2011).
24. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature methods.* **18**, 366–368 (2021).
25. Bu, D. *et al*. KOBAS-i: intelligent prioritization and exploratory visualization of biological functions for gene enrichment analysis. *Nucleic Acids Res* **49**, 317–325 (2021).
26. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
27. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res* **41**, 121 (2013).
28. Griffiths-Jones, S. *et al*. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* **33**, 121–124 (2005).
29. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403–410 (1990).
30. 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).
31. *NCBI BioProject* https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1355350 (2025).
32. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35920331 (2025).
33. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35920332 (2025).
34. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35942771 (2025).
35. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35920325 (2025).
36. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35920326 (2025).
37. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35920327 (2025).
38. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35920328 (2025).
39. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35920329 (2025).
40. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR35920330 (2025).
41. Liu, Q. *Chirolophis japonicus* isolate QL2025-001, whole genome shotgun sequencing project https://identifiers.org/ncbi/insdc:JBTOFL000000000 (2025).
42. Liu, Q. Chirolophis japonicus assembly gene_annotation. *Figshare* https://doi.org/10.6084/m9.figshare.30461972 (2025).
43. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **1525**, 1754–60 (2009).

44. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**(18), 3094–3100 (2018).
45. Krzywinski, M. *et al*. Circos: an information aesthetic for comparative genomics. *Genome Res* **19**, 1639–1645 (2009).
46. Huang, N. & Li, H. compleasm: a faster and more accurate reimplementation of BUSCO. *Bioinformatics* **39**, 595 (2023).

Acknowledgements

This work was funded by the National Key Research and Development Program of China (2024YFD2400200); Dalian Outstanding Young Talents in Science and Technology (2023RJ010); Liaoning Provincial Major Science and Technology Special Projects (2025JH1/11700001); the General Project of Education Department of LiaoningProvince (LJ212410158023, 2024JBZDZ002).

Author contributions

All authors participated in the design and implementation of the study. H.P. and H.Y. performed the experiments, analyzed the data, and drafted the initial manuscript. X.H., J.W. and Y.Z. were responsible for the investigation and collection of experimental materials. Q.L. contributed to the conceptualization of the research and supervised the overall project. All authors reviewed and approved the final manuscript prior to submission.

Competing interests

The authors declare no competing interests.

Additional information

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

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.