scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Telomere-to-telomere gapless genome assembly of *Siniperca scherzeri*

Yannian Wu[1,3], Zhiqiang Cheng[1,3], Yang Li[1,3], Hao Xu[1], Maoyuan Wang[2], Xiaojun Ye[2], Mingyong Lai[2], Zhiyong Wang[1] & Dongling Zhang[1] ✉

The mandarin fish (*Siniperca scherzeri*), renowned as the "freshwater grouper", has emerged as a commercially significant aquaculture species in China due to its superior flesh quality, disease resistance, and domestication adaptability. With the rapid development of bioinformatics, higher standards of genome analysis are now required compared to previous reference genomes. In this study, we integrated PacBio HiFi long-read sequencing, Oxford Nanopore Technologies ultralong-read sequencing, and Hi-C chromatin conformation capture to assemble a near-complete telomere-to-telomere genome. The gapless assembly spans 24 chromosomes, with telomeric repeats detected at both ends of 20 chromosomes and at only one end of the remaining four chromosomes. BUSCO evaluation against the Actinopterygii database (actinopterygii_odb10) revealed 98.7% genome completeness. Alignment analyses using minimap2 demonstrated >97% mapping rates for ONT ultralong reads, PacBio HiFi reads, and Hi-C data against the assembled genome. We annotated 23,296 protein-coding genes, establishing a crucial genomic resource for elucidating the species' evolutionary biology and advancing molecular breeding strategies.

Background & Summary

The mandarin fish (*Siniperca scherzeri*), also known as black mandarin fish, brown mandarin fish, or rock mandarin fish, belongs to the class Osteichthyes, order Perciformes, family Serranidae, and genus *Siniperca*. It is an endemic freshwater fish species to East Asia, widely distributed in the inland waters of China, and has populations in North Korea and Vietnam[1]. *S. scherzeri* primarily inhabits flowing waters rich in gravel substrates and rock crevices. As a typical benthic carnivore, it preys on small fish and aquatic insect larvae, occasionally exhibiting cannibalistic behavior under food scarcity. In natural environments, males generally reach sexual maturity at 1 year old, while females require 2 years, though maturation age varies across geographic populations[2]. Under aquaculture conditions, both sexes achieve gonadal maturity at 1 year old, with peak spawning occurring from May to July. The species exhibits marked sexual dimorphism, with mature females being 10%~30% heavier than males[3]. This pronounced dimorphism provides inherent advantages for developing monosex culture systems to enhance commercial productivity. *S. scherzeri* has earned the epithet "freshwater grouper" due to its superior flesh quality, strong disease resistance, and adaptability to captive breeding, with annual production reaching 374,000 tons, establishing itself as a crucial freshwater aquaculture species in China[4]. In summary, as a premium fish species integrating both economic and medicinal value, *S. scherzeri* demonstrates broad developmental prospects in aquaculture.

S. scherzeri, as a crucial freshwater aquaculture species in China, has gained prominence in artificial cultivation due to its notable disease resistance and economic value. The continuous expansion of aquaculture scale has driven significant progress in fundamental research areas including ecological adaptation mechanisms, germplasm resource evaluation, physiological metabolism regulation, and disease immunity mechanisms. However, it should be noted that early genomic reference sequences developed through Next-generation sequencing technology exhibit substantial limitations: The initial genome assembly published by He *et al.*[5] demonstrated highly fragmented characteristics with a Contig N50 of only 83.8 kb. Although subsequent studies improved genome continuity by 190-fold through integrating third-generation sequencing with Hi-C technology, persistent

[1]Key Laboratory of Healthy Mariculture for the East China Sea, Ministry of Agriculture and Rural Affairs, Jimei University, Xiamen, China. [2]Freshwater Fisheries Research Institute of Fujian Province, Fuzhou, China. [3]These authors contributed equally: Yannian Wu, Zhiqiang Cheng, Yang Li. ✉e-mail: zhangdongling@jmu.edu.cn

Type	Sequencing platform	Data size (Gb)	Average depth (×)	Mean Read length	N50
Pacbio	Pacbio Sequel II	34.6	44	17,995	16,945
Nanopore	Oxford Nanopore	77.3	108	80,440	137,504
Illumina	Illumina NovaSeq 6000	29.6	39	150	150
Hi-C	Illumina NovaSeq S2	93.5	125	150	150
RNA-seq	Illumina HiSeq X Ten	75.7	96	2,039	2,311

Table 1. Sequencing data statistics.

technical challenges remain, including telomere region deletions, 227 unscaffolded contigs, and assembly gaps in repetitive sequence-enriched regions[6]. The recent emergence of long-read sequencing platforms (PacBio HiFi, ONT Ultra-long) combined with innovative assembly algorithm development has significantly enhanced genomic resolution capabilities. These developments have enabled successful telomere-to-telomere (T2T) complete genome assemblies for multiple economically important fish species, including *Mastacembelus armatus*[7] and *Epinephelus lanceolatus*[8]. Notably, despite being one of the three primary cultured mandarin fish species in China's aquaculture industry, the T2T genome assembly for *S. scherzeri* remains unreported to date. This technological gap has substantially impeded the development of molecular marker-assisted breeding systems and the innovation of germplasm resources.

In this study, we employed an integrated approach combining PacBio HiFi long-read sequencing, ONT ultra-long reads, and Hi-C scaffolding technologies to assemble a near-complete T2T genome of *S.scherzeri*. This high-quality reference genome not only provides crucial theoretical foundations for elucidating molecular mechanisms, tracing genetic variation origins, selecting superior traits, and conserving endangered species, but also serves as a critical catalyst for driving innovative advancements in genetic engineering, precision breeding, and biodiversity conservation.

Methods

Ethics statement. Animal procedures reported in this study were approved by the Jimei University Animal Ethics Committee (Approval No. Jmu20240830, approved on 30 August 2024), issued specifically for this project, and all experiments were conducted in accordance with institutional and national regulations.

Sample collection. A healthy live female *S. scherzeri* was collected from Huiyuan Aquatic Products Company in Changting County, Fujian Province. Twelve fresh tissue samples including brain, gonad, heart, liver, spleen, kidney, head kidney, stomach, intestine, muscle, skin, and fin rays were taken. The tissues were immediately frozen in liquid nitrogen and stored at −80°C.Following the standard protocol of QIAGEN DNeasy Blood & Tissue Kit (Qiagen, Shanghai, China), genomic DNA (gDNA) of muscle was extracted. Total RNA was extracted from ten tissues by a TRIzoL kit (Invitrogen, Shanghai, China) and mixed equally for RNA-seq. The quality of nucleic acid was detected by 1.0% agarose gel electrophoresis.

Library construction and sequencing. For PacBio long-read sequencing, genomic DNA was isolated by DNeasy Plant Mini Kit (QIAGEN), circular consensus sequencing (CCS) libraries was constructed using the Pacific Biosciences SMRT bell template prepkit 1.0 and was size-selected by Sage ELF for molecules 14–17 kb,followed by primer annealing and the binding of SMRT bell templates to polymerases with the DNA/Polymerase Binding Kit. Sequencing was performed on the PacBio Sequel II System (https://www.pacb.com/products-and-services/pacbio-systems/sequel/) based on HiFi mode. Five cells for each sample were obtained. Finally, 34.6 Gb of PacBio circular consensus sequencing (CCS) data (coverage of 44×) was obtained (Table 1). For Oxford Nanopore Technologies (ONT) sequencing, qualified DNA samples were subjected to random fragmentation via g-TUBE, followed by library preparation using the Ligation Sequencing Kit SQK-LSK112. The constructed libraries were quantified using Qubit 4.0 and subjected to fragment size distribution quality control prior to loading onto PromethION Flow Cells for single-molecule real-time sequencing. Raw data were base-called through Guppy (v6.4.6)[9], with subsequent filtering to remove low-quality reads and adapter contaminants. This process ultimately yielded 77.3 GB of clean data, achieving a sequencing depth of approximately 108× (Table 1). For Illumina short-read sequencing, genomic DNA was randomly fragmented into small segments. Following adapter ligation to construct short-insert libraries, paired-end sequencing was performed on the Illumina NovaSeq 6000 platform. The acquired sequencing data were utilized for whole-genome survey analysis to estimate genome size and perform base-level error correction for third-generation sequencing data. Ultimately, a total of 29.6 Gb of clean reads (150 bp in length) with 39× genomic coverage was generated (Table 1). For Hi-C sequencing, library preparation according to the standard procedure, nuclear DNA was cross-linked *in situ*, extracted, and then digested by Mbo I restriction endonuclease, the sticky ends of the digested fragments were biotinylated, diluted and ligated randomly. The biotinylated DNA fragments were enriched to construct sequencing library, and sequencing on Illumina NovaSeq S2 platform, which generated 150-bp paired-end reads and generated 93.5 Gb clean reads overall (coverage of 125×). (Table 1) The RNA-seq library was constructed using Illumina standard protocol (San Diego, CA, USA) and sequenced on the Illumina HiSeq X Ten platform. Finally, we obtained 75.7 Gb paired-end clean reads for the following gene prediction (Table 1).

Genome size estimation. Genome size estimation was performed using a k-mer-based analytical strategy[10]. The detailed workflow comprised the following steps: Illumina PE150 sequencing data were processed through Jellyfish (v2.3.0)[11] for k-mer frequency enumeration. The Lander-Waterman model was applied to

Sample	k-mer	k-mer number	Genome size (bp)	Heterozygosity rate (%)
XX	21	19,524,911,653	706,061,784	0.33

Table 2. Genome size assessment of *S. scherzeri* (k-mer = 21).

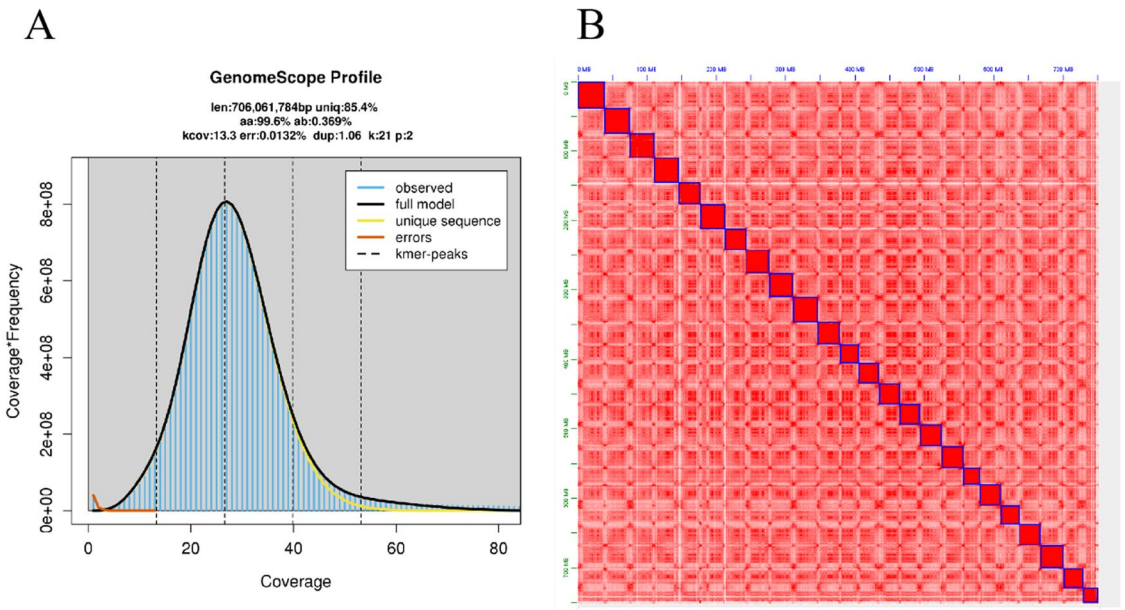

Fig. 1 (**A**) K-mer frequency distribution of *S. scherzeri* genome. (**B**) Hi-C interaction heatmap of *S. scherzeri* chromosomes. Blue rectangles represent individual chromosomes, with interaction intensity increasing through a color gradient from light to dark.

perform Gaussian fitting on k-mer depth distribution, thereby validating sequencing data quality (kurtosis coefficient < 0.5). Subsequent parameter estimation was conducted via GenomeScope 2.0[12], employing the computational model: genome size equals the ratio of total k-mer count to the expected sequencing depth, with simultaneous evaluation of heterozygosity rate. The final estimation revealed a mandarin fish genome size of 706.06 Mb and a heterozygosity level of 0.33% (Table 2, Fig. 1A).

***De novo* genome assembly.** The chromosome-level genome assembly was constructed through an integrative approach combining multiple sequencing technologies. PacBio HiFi and Oxford Nanopore ultra-long reads were co-assembled using Verkko (v1.4)[13] with specified parameters:–hifi,–nano,–hic1,–hic2. Chromosomal scaffolding was enhanced through Hi-C data integration, with raw sequencing reads preprocessed via HiCUP (v0.8.1)[14] for restriction enzyme site identification and read end-trimming. Juicer (v1.6)[15] converted alignment results into 3D interaction heatmaps, enabling 3D-DNA (v180419)[16] to perform spectral clustering of contigs onto chromosomal coordinates. Manual curation in Juicebox (v1.11.08)[17] resolved topological discordances (including inversions and translocations) by aligning contig orientations with interaction frequency gradients. Residual assembly gaps were resolved using third-generation sequencing data with TGS-GapCloser (v1.2.1)[18], resulting in fully continuous chromosomal sequences. Telomeric completeness was validated through quarTeT's motif scanning algorithm, detecting ≥ 8 consecutive TTAGGG repeats[19]. The final T2T assembly spans 757.9 Mb with contig N50 of 31.97 Mb, achieving gapless chromosomal integration, with 99.83% sequences anchored to 24 chromosomes (Fig. 1B and Table 3). Telomeric repeats were detected at both ends of 20 chromosomes (40 telomeres in total), while four additional telomere signals were identified on chromosomes exhibiting only a single terminal repeat (Table 4). Thus, the 20 T2T chromosomes account for 40 canonical telomeric ends, and the remaining 4 signals correspond to chromosomes with only one detectable telomeric repeat. For chromosomes 14, 18, 19, and 23, although only a single canonical telomere motif was detected, the terminal regions exhibited continuous HiFi, ONT, and Hi-C read coverage without structural interruptions, supporting the completeness and accuracy of these chromosome ends.

Repeat annotation. Repetitive sequences were predicted by two strategies: homology-based and *de novo* methods. For homology-based repeat identification, we used RepeatMasker[20] and RepeatProteinMask (version4.0.5)[20] to scan the whole genome for known transposable elements in RepBase database[21]. *De novo* repeats were identified using RepeatModeler[22]. The results showed that the genome repetitive sequence size was 236.86 Mb, accounting for 31.39% of the assembled genome. Among the repeat elements, short interspersed nuclear elements (SINEs) accounted for 0.37% of genome size and long interspersed nuclear elements (LINEs) accounted for 4.85%. Long terminal repeats (LTRs) and DNA elements accounted for 0.67% and 1.57%, respectively (Table 5).

Type	Data
Contig N50(bp)	31,976,378
Contig N90(bp)	27,557,260
Longest Contig (bp)	39,334,400
Contig Number	26
Length ≥ 1 kb	26
Length ≥ 2 kb	26
Length ≥ 5 kb	26
T2T Number	20
Gap Number	0
Telomere Number	44
Scaffolding Rate (%)	99.83
Chromosome Length	756,661,368
Total Length(bp)	757,906,621

Table 3. The statistics of *S. scherzeri* genome assembly.

Chromosome	Length(bp)	Number of telomere
Chr1	29,279,161	2
Chr2	31,511,270	2
Chr3	36,875,188	2
Chr4	36,849,970	2
Chr5	34,915,548	2
Chr6	26,885,724	2
Chr7	34,242,438	2
Chr8	31,473,800	2
Chr9	34,341,398	2
Chr10	34,335,361	2
Chr11	30,224,197	2
Chr12	30,266,081	2
Chr13	28,983,447	2
Chr14	22,025,399	1
Chr15	24,669,220	2
Chr16	39,334,400	2
Chr17	33,038,002	2
Chr18	31,396,749	1
Chr19	31,976,378	1
Chr20	34,089,650	2
Chr21	32,915,887	2
Chr22	30,457,355	2
Chr23	27,557,260	1
Chr24	29,017,485	2

Table 4. Assembly statistics of chromosomes.

Gene prediction and annotation. Structural annotation employed three distinct strategies for gene structure prediction: *de novo* prediction, homology-based prediction, and transcriptome-based gene prediction.

De novo prediction: Augustus (v3.5.0)[23] was used for *de novo* gene prediction, constructing models based on the zebrafish training set and transcriptome data from this study. RNA-seq data were aligned to the repeat-masked genome using TopHat (v2.1.1)[24]. After two rounds of iterative training to optimize parameters, a high-confidence prediction model for this species was generated, followed by *de novo* prediction using Augustus.

Homology-based prediction: Protein sequences from nine closely related teleost species were obtained from the NCBI database (https://www.ncbi.nlm.nih.gov/), including *Danio rerio*, *Larimichthys crocea*, *Oreochromis niloticus*, *Cynoglossus semilaevis*, *Oryzias latipes*, *Lates calcarifer*, *Gasterosteus aculeatus*, *Takifugu rubripes*, and *Siniperca chuatsi*. Tblastn combined with Genomethreader (v1.7.3)[25] was selected for protein-nucleotide alignment to locate splice sites and predict exon regions for homology-based gene prediction.

Transcriptome-based prediction: RNA-seq data were aligned to the genome using Hisat2 (v2.2.1)[26], with reference-based transcript assembly performed by StringTie (v3.0.0)[27]. Trinity (v2.15.2)[28] was simultaneously employed for *de novo* transcript assembly. Both types of transcripts were mapped to the genome through PASA

Repeat element	Fragments	Total length (bp)	% of genome
SINE	28,055	2,810,758	0.37
LINE	91,708	36,614,692	4.85
LTR element	6,952	5,026,155	0.67
DNA element	38,783	11,840,401	1.57
Unclassified	808,166	148,659,355	19.70
Total interspersed repeats	—	204,951,361	27.16
Small RNA	3,917	828,964	0.11
Satellite	2,208	452,162	0.06
Simple repeat	340,385	20,196,587	2.68
Low complexity	38,764	2,327,993	0.31
Total	—	228,757,067	30.32

Table 5. Annotation of repetitive elements in *S. scherzeri* genome.

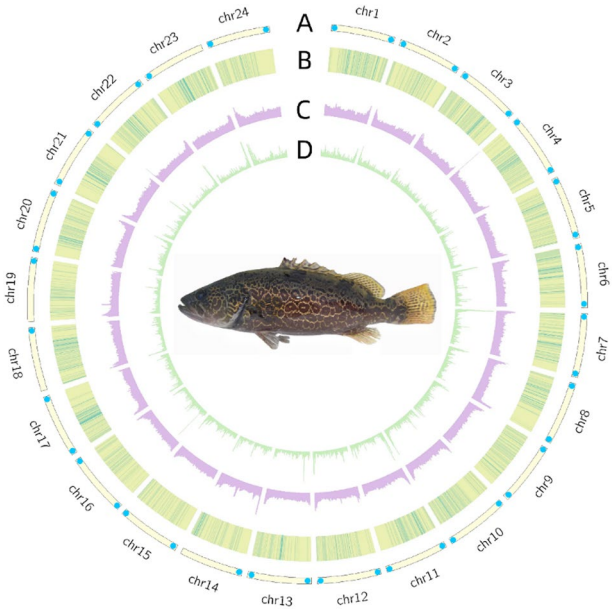

Fig. 2 Circos of *S. scherzeri* genome. (A) represents chromosomes (blue dots denote telomeres); (B) represents gene density (darker colors indicate higher densities); (C) represents repeat sequence density; (D) represents GC content.

(v2.0.2)[29], combined with open reading frame prediction using TransDecoder (v5.7.1) (https://github.com/TransDecoder/TransDecoder). Final integration of all results into a non-redundant gene set was accomplished using EvidenceModeler (v2.1.0)[30].

Functional annotation: Blastp[31] alignment against multiple databases (Nr[32], Swissprot[33], GO[34], KEGG[35], Pfam (http://pfam.xfam.org), TrEMBL[36]) was performed for predicting motifs, domains, protein functions, and metabolic pathways. tRNAscan-SE[37] was used for tRNA prediction, rRNA databases for rRNA prediction, and ncRNAs were identified through Pfam database searches using INFERNAL (http://eddylab.org/infernal/). In summary, 23,296 protein-coding genes were identified in the mandarin fish genome. Functional annotation revealed that 22,455 genes (96.39%) obtained functional annotations in databases including NR and Swiss-Prot. A circos plot was generated to visualize the global feature distribution of the genome (Fig. 2).

Technical Validation

Genome assembly quality was systematically assessed through a multidimensional evaluation framework. Completeness evaluation was performed using BUSCO (v5.4.3)[38] with the Actinopterygii database (actinopterygii_odb10), which benchmarks the presence of evolutionarily conserved single-copy orthologs. The BUSCO analysis of assembly showed that 3,593 (98.7%) of the complete orthologs, including 3,564 (97.9%) single-copy orthologs and 292 (0.8%) duplicated orthologs, as well as 73 (0.2%) fragmented orthologs were identified (Table 6). To further validate assembly quality, raw sequencing data were aligned to the genome using two complementary tools: BWA-MEM (v0.7.18)[39] for short reads and Minimap2 (v2.28)[40] for long-read optimization. Multi-platform sequencing data (HiFi, ONT, Hi-C) demonstrated robust alignment efficiency (>97% mapping rates; Table 7), validating the high fidelity of the assembled sequences. To complement these evaluations,

Type	Data
Complete BUSCOs (C)	3,593(98.7%)
Complete and single-copy BUSCOs(S)	3,564(97.9%)
Complete and duplicated BUSCOs(D)	292(0.8%)
Fragmented BUSCOs (F)	73(0.2%)
Missing BUSCOs (M)	40(1.1%)

Table 6. Genome assembly completeness assessment by BUSCO.

Type	Mapping rate
HiFi	99.98%
ONT	99.08%
Hi-C	97.93%

Table 7. Mapping efficiency of clean reads.

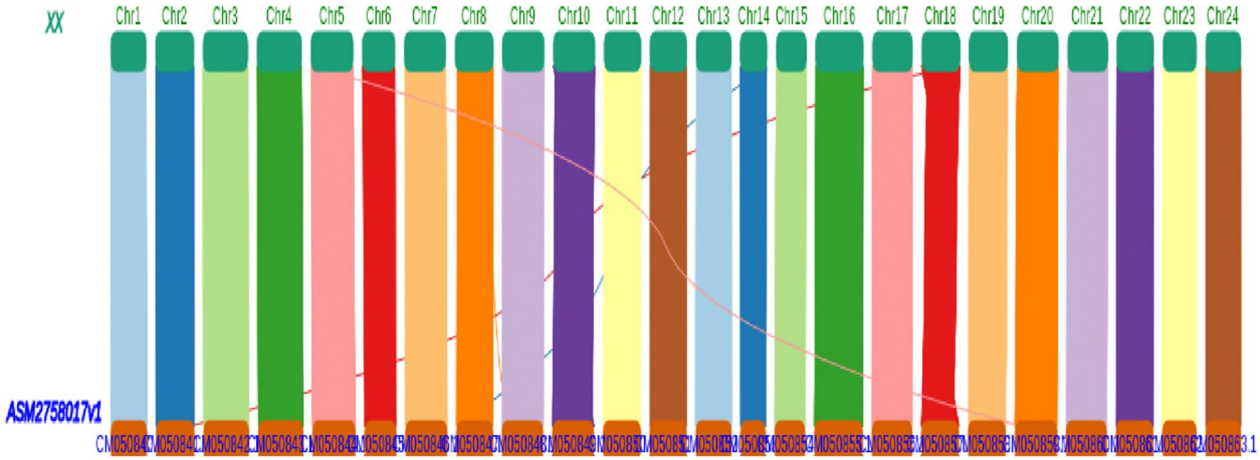

Fig. 3 Syntenic analysis between the XX genomes of *S. scherzeri* assembled in this study and the reference genome ASM2758017v1.

we also assessed the base-level consensus accuracy using Merqury (v1.3), which estimates quality values (QV) based on k-mer concordance between Illumina short reads and the assembled genome. The assembly achieved a QV score of 46.07, corresponding to an estimated per-base error rate of 2.47×10^{-5}. These base-level metrics are consistent with the high BUSCO completeness and mapping rates, further confirming the accuracy and reliability of the final genome assembly. Finally, whole-genome collinearity analysis between the T2T genome assembled in this study and ASM2758017v1 was performed through Minimap2 combined with NgenomeSyn[41]. The collinearity analysis confirmed that the genome assembled in this study showed high collinearity with the latest ASM2758017v1 genome (Fig. 3), with high consistency in chromosomal structure.

Data Records

The genome sequencing data generated in this study and the assembled genome as well as annotation data have been deposited into the China National GeneBank Sequence Archive (CNSA) under the project accession CNP0007951[42]. The sequencing data have been also archived in the NCBI SRA under accession SRP652857[43], and the assembled genome and annotation have been archived in the NCBI GenBank under accession JBVQOQ000000000[44]. The genome assembly data and annotations have also been deposited at Figshare[45].

Data availability

The genome sequencing data generated in this study and the genome assembly as well as annotation data have been deposited into the China National GeneBank Sequence Archive (CNSA) under the project accession CNP0007951[42]. The sequencing data have been also archived in the NCBI SRA under accession SRP652857[43], and the assembled genome and annotation have been archived in the NCBI GenBank under accession JBVQOQ000000000[44]. The genome assembly data and annotations have also been deposited at Figshare[45].

Code availability

No specific code was used in this study. The data analysis used standard bioinformatic tools specified in the methods.

Received: 15 September 2025; Accepted: 19 March 2026;
Published online: 02 April 2026

References

1. Zhou, C., Yang, Q. & Cai, D. On the Classification and Distribution of the Sinipercinae Fishes (Family Serranidae). *Zoolgical Research* **9**, 113–125 (1988).
2. Li, Y. *et al*. Identification of the Sex-Linked Region of *Siniperca Scherzeri* and Development of Sex-Specific Markers. *Aquaculture* **600**, 742231 (2025).
3. Sun, C. *et al*. Construction of a High-Density Linkage Map and Mapping of Sex Determination and Growth-Related Loci in the Mandarin Fish (*Siniperca Chuatsi*). *Bmc Genomics* **18**, 446 (2017).
4. Wang, M. *et al*. Comparison of Growth Performance and Muscle Nutrition Levels of Juvenile *Siniperca Scherzeri* Fed On an Iced Trash Fish Diet and a Formulated Diet. *Fishes.* **8**, 393 (2023).
5. He, S. *et al*. Mandarin Fish (Sinipercidae) Genomes Provide Insights Into Innate Predatory Feeding. *Commun. Biol.* **3**, 361 (2020).
6. Tu, G. *et al*. Long-Read Genome Assemblies Reveal a Cis-Regulatory Landscape Associated with Phenotypic Divergence in Two Sister Siniperca Fish Species. *Zoolgical Research.* **44**, 287–302 (2023).
7. Xue, L. *et al*. Telomere-to-Telomere Assembly of a Fish Y Chromosome Reveals the Origin of a Young Sex Chromosome Pair. *Genome Biol.* **22**, 203 (2021).
8. Zhou, Q. *et al*. Telomere-to-Telomere Gapless Genome Assembly of the Giant Grouper (*Epinephelus Lanceolatus*). *Sci. Data.* **11**, 1342 (2024).
9. Sherathiya, V. N., Schaid, M. D., Seiler, J. L., Lopez, G. C. & Lerner, T. N. Guppy, a Python Toolbox for the Analysis of Fiber Photometry Data. *Sci. Rep.* **11**, 24212 (2021).
10. Liu, Y., Schroder, J. & Schmidt, B. Musket: A Multistage K-Mer Spectrum-Based Error Corrector for Illumina Sequence Data. *Bioinformatics* **29**, 308–315 (2013).
11. Marcais, G. & Kingsford, C. A Fast, Lock-Free Approach for Efficient Parallel Counting of Occurrences of K-Mers. *Bioinformatics* **27**, 764–770 (2011).
12. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. Genomescope 2.0 and Smudgeplot for Reference-Free Profiling of Polyploid Genomes. *Nat. Commun.* **11**, 1432 (2020).
13. Rautiainen, M. *et al*. Telomere-to-Telomere Assembly of Diploid Chromosomes with Verkko. *Nat. Biotechnol.* **41**, 1474–1482 (2023).
14. Wingett, S. *et al*. Hicup: Pipeline for Mapping and Processing Hi-C Data. *F1000Research* **4**, 1310 (2015).
15. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systerms* **3**, 95–98 (2016).
16. Dudchenko, O. *et al*. De Novo Assembly of the Aedes Aegypti Genome Using Hi-C Yields Chromosome-Length Scaffolds. *Science.* **356**, 92–95 (2017).
17. Durand, N. C. *et al*. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Systerms* **3**, 99–101 (2016).
18. Xu, M. *et al*. Tgs-Gapcloser: A Fast and Accurate Gap Closer for Large Genomes with Low Coverage of Error-Prone Long Reads. *Gigascience* **9**, giaa94 (2020).
19. Lin, Y. *et al*. Quartet: A Telomere-to-Telomere Toolkit for Gap-Free Genome Assembly and Centromeric Repeat Identification. *Hortic. Res.* **10**, uhad127 (2023).
20. Tarailo-Graovac, M. & Chen, N. Using Repeatmasker to Identify Repetitive Elements in Genomic Sequences. *Curr Protoc Bioinformatics* Chapter 4, 4–10 (2009).
21. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a Database of Repetitive Elements in Eukaryotic Genomes. *Mob. Dna.* **6**, 11 (2015).
22. Flynn, J. M. *et al*. Repeatmodeler2 for Automated Genomic Discovery of Transposable Element Families. *Proc. Natl. Acad. Sci. USA.* **117**, 9451–9457 (2020).
23. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using Native and Syntenically Mapped Cdna Alignments to Improve De Novo Gene Finding. *Bioinformatics* **24**, 637–644 (2008).
24. Trapnell, C., Pachter, L. & Salzberg, S. L. Tophat: Discovering Splice Junctions with Rna-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
25. Gremme, G., Brendel, V., Sparks, M. E. & Kurtz, S. Engineering a Software Tool for Gene Structure Prediction in Higher Organisms. *Inf. Softw. Technol.* **47**, 965–978 (2005).
26. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. 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. Grabherr, M. G. *et al*. Full-Length Transcriptome Assembly From Rna-Seq Data without a Reference Genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
29. Haas, B. J. *et al*. Improving the Arabidopsis Genome Annotation Using Maximal Transcript Alignment Assemblies. *Nucleic. Acids. Res.* **31**, 5654–5666 (2003).
30. Haas, B. J. *et al*. Automated Eukaryotic Gene Structure Annotation Using Evidencemodeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
31. 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).
32. Pruitt, K. D., Tatusova, T. & Maglott, D. R. Ncbi Reference Sequences (Refseq): A Curated Non-Redundant Sequence Database of Genomes, Transcripts and Proteins. *Nucleic. Acids. Res.* **35**, D61–D65 (2007).
33. Bairoch, A. & Apweiler, R. The Swiss-Prot Protein Sequence Database and its Supplement Trembl in 2000. *Nucleic. Acids. Res.* **28**, 45–48 (2000).
34. Ashburner, M. *et al*. Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29 (2000).
35. Kanehisa, M. & Goto, S. Kegg: Kyoto Encyclopedia of Genes and Genomes. *Nucleic. Acids. Res.* **28**, 27–30 (2000).
36. Bairoch, A. & Apweiler, R. The Swiss-Prot Protein Sequence Data Bank and its Supplement Trembl in 1999. *Nucleic. Acids. Res.* **27**, 49–54 (1999).
37. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. Trnascan-Se 2.0: Improved Detection and Functional Classification of Transfer Rna Genes. *Nucleic. Acids. Res.* **49**, 9077–9096 (2021).
38. Di Tommaso, P. *et al*. Nextflow Enables Reproducible Computational Workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
39. Li, H. & Durbin, R. Fast and Accurate Long-Read Alignment with Burrows-Wheeler Transform. *Bioinformatics* **26**, 589–595 (2010).
40. Li, H. Minimap2: Pairwise Alignment for Nucleotide Sequences. *Bioinformatics* **34**, 3094–3100 (2018).
41. He, W. *et al*. Ngenomesyn: An Easy-to-Use and Flexible Tool for Publication-Ready Visualization of Syntenic Relationships Across Multiple Genomes. *Bioinformatics* **39**, btad121 (2023).
42. China National GeneBank Sequence Archive (CNSA). https://db.cngb.org/data_resources/project/CNP0007951 (2025).
43. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP652857 (2025).

44. *NCBI GenBank* https://www.ncbi.nlm.nih.gov/nuccore/JBVQOQ000000000 (2026).
45. Li, Y. *Siniperca scherzeri* genome. *figshare* https://doi.org/10.6084/m9.figshare.30084370.v1 (2025).

Acknowledgements
This work was financially supported by Fujian Province Seed Industry Innovation and Industrialization Engineering Fishery Project(2021MNZ05), and Agriculture Research System of China (CARS-46).

Author contributions
D.Z. and Z.W. designed the study. Y.L. H.X. and Z.C. were involved in sample collection. Y.L., Z.C. and Y.W. performed the experiment and analyzed the data. Y.L. and Y.W. wrote the paper. M.W., X.Y. and M.L. revised the paper.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to D.Z.

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.