scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of the *Siniperca obscura*

Haiyang Liu[1]✉, Huijuan Liu[1,2], Kai Cui[3], Wenxuan Lu[3], Jing Li[3], Ting Fang[3], Na Gao[3], Cheng Chen[3], Xiuxia Zhao[3], Kun Yang[3], Yanfeng Lin[4]✉ & Yangyang Liang[3]✉

Siniperca obscura **is an economically valuable and ecologically significant species in China, yet the lack of comprehensive genomic resources has hindered genetic studies and breeding efforts. In this study, we present a high-quality, chromosome-level genome assembly for *S. obscura*, generated by integrating PacBio HiFi long-read sequencing with Hi-C scaffolding. The final assembly spans 734.12 Mb, with 99.85% of the assembled bases anchored and oriented onto 24 chromosomes. It achieves a contig N50 of 24.59 Mb and a scaffold N50 of 30.62 Mb, with high genome completeness further demonstrated by a BUSCO score of 99.37%. We predicted 23,225 protein-coding genes, with a 99.12% BUSCO completeness, and 98.50% of the genes were functionally annotated. Approximately 30.54% of the genome sequences were classified as repeat elements. This high-quality reference genome provides a valuable resource for advancing molecular breeding, comparative genomics, and evolutionary studies of *S. obscura* and closely related species.**

Background & Summary

Siniperca obscura is an economically valuable and ecologically significant freshwater species endemic to China, classified within order Perciformes, family Sinipercidae, and genus *Siniperca*[1]. This species primarily inhabits the benthic to mid-water layers of freshwater systems and has stringent water quality requirements. As a carnivorous predator, it preys on small fish and shrimp, playing a vital role in maintaining the stability of freshwater ecosystems[2]. *S. obscura* has high market demand due to its tender, boneless flesh, demonstrating considerable potential for aquaculture development. Additionally, the distribution of *S. obscura* across distinct river basins, coupled with the historical classification of different geographic populations as separate species, makes it an ideal subject for comparative genetic studies[3]. However, wild populations of *S. obscura* have significantly declined in recent decades due to hydrological engineering, overfishing, and water pollution[4]. As a result, it has been designated as a protected species by several provincial governments (e.g., Hunan Province). Despite its urgent conservation needs and the growing demand for aquaculture, comprehensive genomic studies on *S. obscura* remain limited, significantly hindering advancements in conservation strategies, adaptive evolution studies, and molecular breeding research.

Currently, research on *S. obscura* primarily focuses on mitochondrial DNA sequence analysis[5], SSR molecular marker development[6,7], and the exploration of its evolutionary relationships[2,3,8,9]. Despite these advancements, a deeper investigation into its genomic features is essential to enhance our understanding of its biology and improve its applications in aquaculture. In recent years, significant progress has been made in the genomic research of the Sinipercidae family, including species such as the *Siniperca chuatsi*[10–12], *Siniperca scherzeri*[13], *Siniperca kneri*[14] and *Siniperca roulei*[15]. These genomic studies have provided a solid foundation for research on adaptive evolution, predatory behaviour, and sensory-related genes. However, *S. obscura*, a key species in the genus *Siniperca*, still lacks a high-quality reference genome, which limits further research on its adaptive evolution, population genetics, and comparative genomics with other Sinipercidae species.

In this study, we present the first chromosome-level genome assembly of *S. obscura*, generated by integrating PacBio long-read sequencing, Hi-C chromatin conformation capture, and high-throughput transcriptomic data. The assembly demonstrates high continuity and completeness and includes annotations for

[1]Key Laboratory of Tropical and Subtropical Fishery Resources Application and Cultivation, Ministry of Agriculture and Rural Affairs, Pearl River Fisheries Research Institute, Chinese Academy of Fishery Sciences, Guangzhou, 510380, China. [2]School of Marine Sciences, Ningbo University, Ningbo, 315211, China. [3]Key Laboratory of Freshwater Aquaculture and Enhancement of Anhui Province, Fisheries Research Institute, Anhui Academy of Agricultural Sciences, Hefei, 230001, China. [4]Fishery Bureau of Xiuning County, Huangshan, 245400, China. ✉e-mail: hyliu@prfri.ac.cn; linyf@yeah.net; liangyy10214@126.com

Types	Method	Raw data (Gb)	Clean data (Gb)
DNA	MGI	75.26	73.74
DNA	PacBio	35.58	35.29
DNA	Hi-C	121.48	117.57
RNA	MGI	23.22	21.76

Table 1. Statistics of sequencing data.

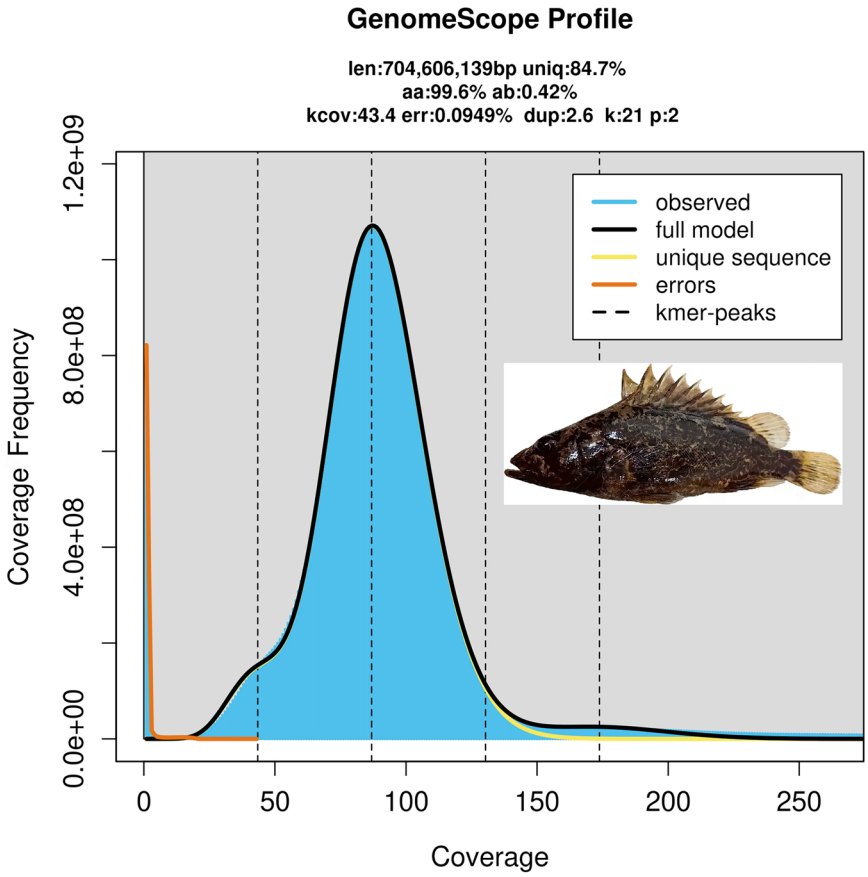

Fig. 1 Genome survey of *S. obscura* was conducted using 21-mer analysis with GenomeScope2.

23,225 protein-coding genes. This dataset not only serves as a valuable resource for evolutionary studies within Sinipercidae but also facilitates applied research in molecular marker development, disease-resistant trait selection, and wild resource conservation.

Methods

Sample collection and DNA extraction. A male *S. obscura* was collected from a tributary of the Yangtze River in Qimen County, Huangshan City, Anhui Province, China. Genomic DNA was extracted from muscle tissue using the Qiagen DNeasy Blood and Tissue Kit (Qiagen, USA), following the manufacturer's instructions. DNA integrity and purity were assessed through 1% agarose gel electrophoresis, and the concentration was measured using a NanoDrop One spectrophotometer (Thermo Scientific, USA). All procedures strictly followed the guidelines approved by the Ethics Committee of the Pearl River Fisheries Research Institute, Chinese Academy of Fishery Sciences (Approval No. PRFRI-2024-023).

Genome sequencing. For short-read sequencing, a 350 bp insert-size library was prepared from sheared DNA using the MGIEasy PCR-Free DNA library prep set (BGI, China) and sequenced on the DNBSEQ-T7 platform (MGI, China), generating 75.26 Gb of PE150 raw reads. For long-read sequencing, high-fidelity (HiFi) libraries were constructed with the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, USA) and sequenced on the PacBio Sequel II system, producing 35.29 Gb of PacBio HiFi long reads with an average length of 18.21 kb. Additionally, Hi-C libraries were prepared from approximately 1 g of *S. obscura* muscle tissue using the GrandOmics Hi-C kit (DpnII digestion) and sequenced on the DNBSEQ-T7 platform (MGI, China), resulting in 121.48 Gb of Hi-C data (Table 1).

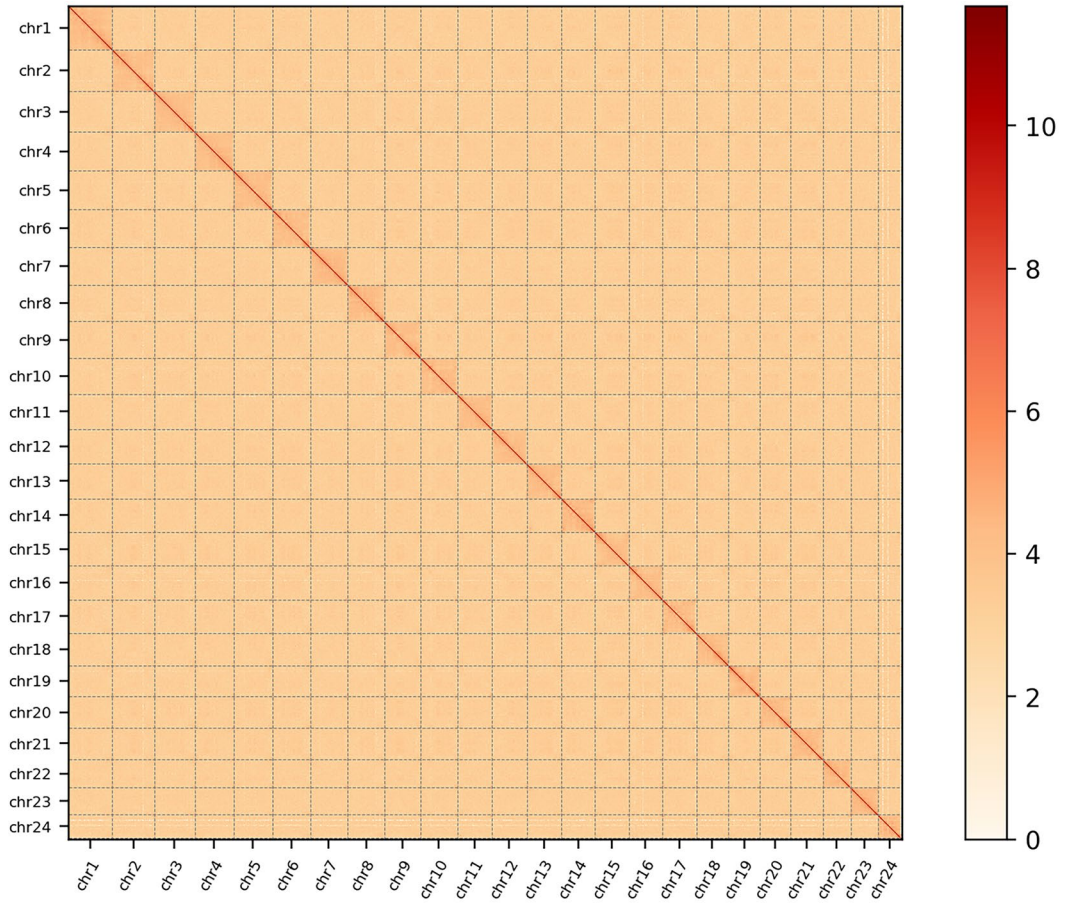

Fig. 2 Hi-C interactive heatmap of 24 chromosomes in *S. obscura* genome.

RNA extraction and transcriptome sequencing. For comprehensive genome annotation, RNA was extracted from ten tissues (spleen, kidney, brain, muscle, ovary, liver, intestine, heart, gill, and swim bladder). RNA quality was evaluated using both a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific) and an Agilent 2100 Bioanalyzer. Equal amounts (1 μg per tissue) of RNA were pooled to construct a cDNA library with the MGIEasy PCR-Free DNA library prep set (BGI, China), followed by PE150 sequencing mode on the DNBSEQ-T7 platform (MGI, China). This resulted in 23.22 Gb of transcriptomic data to support genome annotation.

Genome size and heterozygosity estimation. MGI raw reads were quality-filtered and adapter-trimmed using fastp v0.23.4[16] with the parameter -l 50, yielding a total of 73.74 Gb of clean data. The clean reads were then used for genome size estimation of *S. obscura* through k-mer analysis. A 21-mer frequency table was generated using Jellyfish v2.3.0[17] and subsequently analyzed with GenomeScope v2.0[18]. The analysis revealed an estimated genome size of approximately 704.61 Mb, with a heterozygosity rate of 0.42% and repeat content of 15.3% (Fig. 1).

Genome assembly. Raw PacBio HiFi reads were filtered prior to assembly by removing reads shorter than 100 bp, and a total of 35.29 Gb of high-quality HiFi data were retained. The genome of *S. obscura* was assembled using the filtered HiFi reads with Hifiasm v0.19.9[19] under default parameters, generating an initial contig-level assembly of 735.19 Mb, with a contig N50 of 24.73 Mb and a maximum contig length of 34.67 Mb. Raw Hi-C reads were filtered using fastp v0.23.4[16] with the parameter -l 50, yielding 117.57 Gb of clean data. Hi-C–based scaffolding was performed using Juicer v1.6[20] in combination with the 3D-DNA pipeline v20100821[21]. Briefly, the contig-level assembly was indexed with BWA v0.7.17[22], and clean Hi-C paired-end reads were aligned to the contigs to generate contact maps. This procedure produced 242.51 million uniquely mapped read pairs, corresponding to a unique mapping rate of 59.92%. After stringent filtering, 110,460,838 non-redundant valid interaction pairs were retained, accounting for 45.55% of the uniquely mapped alignments (detailed statistics are provided in Supplementary Table S1). Initial chromosome-level scaffolds were generated using 3D-DNA under default parameters. Chromosome boundaries and misassemblies were subsequently refined through manual curation using Juicebox v1.11.08[23]. As a result, 99.85% of the assembled sequences were successfully anchored and oriented onto 24 chromosomes (Figs. 2, 3). The final assembly comprised 43 scaffolds, with a maximum scaffold length of 38.30 Mb and a scaffold N50 of 30.62 Mb (Tables 2, 3).

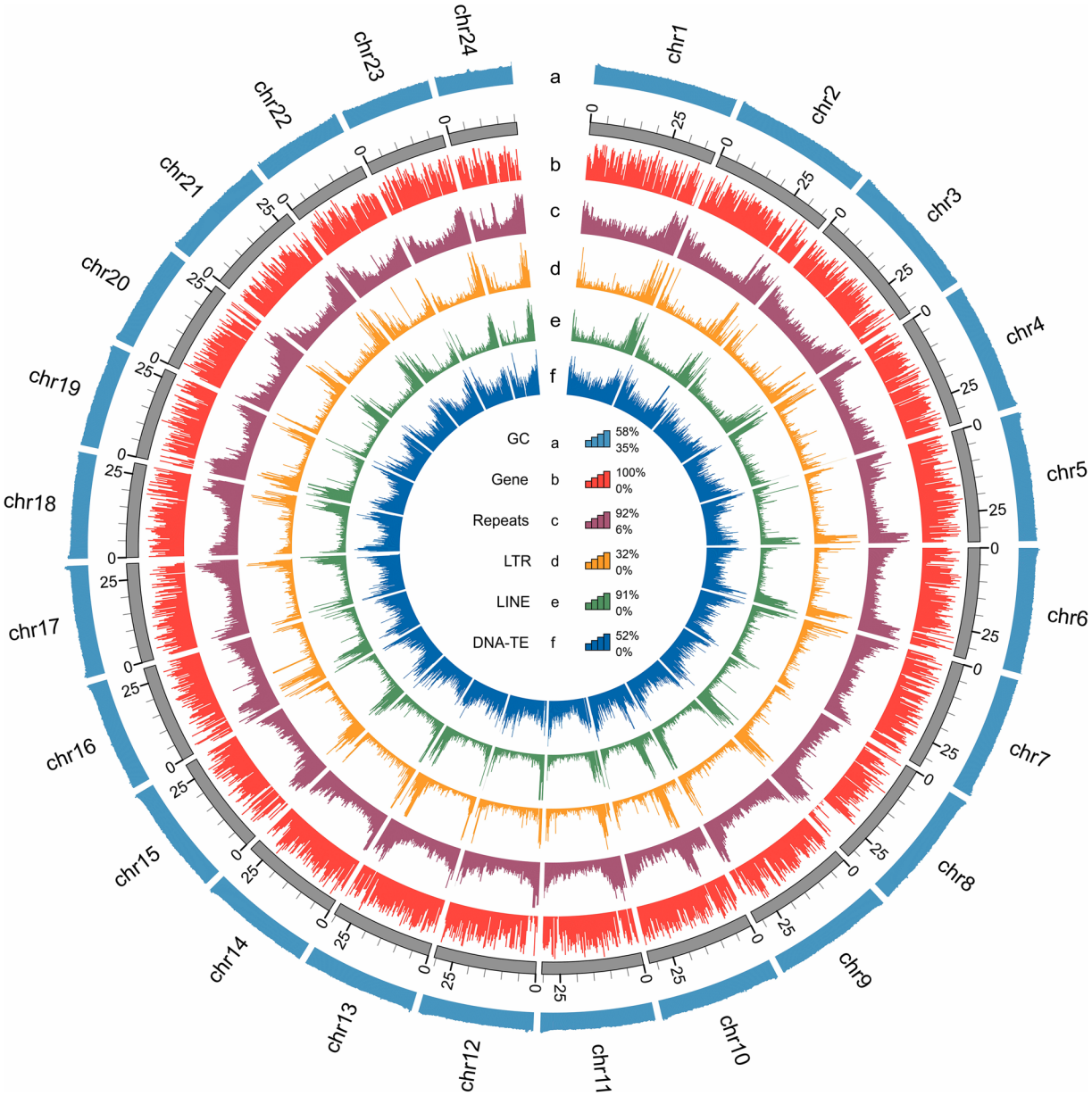

Fig. 3 Features of the *S. obscura* genome. From outside to inside: (a) GC content, (b) Gene density, (c) Repeats content, (d) LTR content, (e) LINE content, (f) DNA-TE content.

Metric	Contig-level	Chromosome-level
Assembly genome length (Mb)	735.19	734.12
GC (%)	40.42	40.42
contigs number	132	127
contig N50 (Mb)	24.73	24.59
contig N90 (Mb)	3.89	4.20
scaffolds number	132	43
scaffold N50 (Mb)	24.73	30.62
scaffold N90 (Mb)	3.89	27.17
Longest scaffold (Mb)	34.67	38.30
Number of chromosomes		24

Table 2. Summary statistics of *S. obscura* contig-level and chromosome-level genome assembly.

Chromosome	Length (bp)	Number of Contigs
Chr01	38,295,284	3
Chr02	36,747,327	5
Chr03	35,686,122	4
Chr04	34,240,585	4
Chr05	34,171,263	6
Chr06	33,394,294	9
Chr07	32,981,857	6
Chr08	32,369,141	10
Chr09	31,985,866	4
Chr10	32,241,401	11
Chr11	30,471,572	1
Chr12	30,621,129	3
Chr13	30,530,749	2
Chr14	29,727,018	6
Chr15	29,749,904	3
Chr16	29,574,930	10
Chr17	29,860,058	3
Chr18	28,313,692	1
Chr19	27,518,391	6
Chr20	27,167,291	3
Chr21	28,418,504	2
Chr22	24,393,381	3
Chr23	24,053,660	2
Chr24	20,485,290	1
Total	732,998,709	108

Table 3. Summary statistics of *S. obscura* genome assembly.

Repeat class	Total length (bp)	Genome proportion (%)
DNA	81,935,614	11.16
LINE	58,374,657	7.95
SINE	3,571,906	0.49
LTR	17,819,282	2.43
Satellite	2,552,444	0.35
Unknown	71,065,081	9.68
Total	224,193,574	30.54

Table 4. Composition of repetitive elements in the *S. obscura* genome.

Repeat annotation. We conducted a comprehensive repeat analysis using multiple computational methods. For tandem repeats, GMATA v2.2.1[24] and Tandem Repeats Finder v4.10.0[25] were used to identify Satellite repeats, which covered 0.35% of the genome. Dispersed repeats were characterized using an integrated pipeline. First, MITEs[26] were detected with MITE-hunter to create a MITE library. The genome was then hardmasked and analyzed with RepeatModeler v2.0.6[27] to generate a de novo repeat library (RepMod.lib). TEclass was used to classify unknown repeats in RepMod.lib, which was subsequently combined with MITE.lib and Repbase v19.06[28] to construct a combined repeat library. RepeatMasker v4.1.6[29] applied this library for genome-wide repeat annotation. The analysis revealed that 30.54% of the genome consists of dispersed repeats (Table 4). DNA transposons were the most abundant (11.16%), followed by LINEs (7.95%), LTR retrotransposons (2.43%), and SINEs (0.49%). In total, repetitive sequences accounted for 224,193,574 bp (30.54%) of the genome (Table 4, Fig. 4).

Gene prediction and function assignment. Gene prediction was performed through an integrative approach combining transcriptomic evidence, homology-based methods, and ab initio predictions. Raw RNA-seq reads were first quality-filtered and adapter-trimmed using fastp v0.23.4 with the parameter -l 50, yielding a total of 21.76 Gb of clean data. For transcriptome-guided annotation, clean RNA-seq reads were aligned to the genome using HISAT2 v2.2.1[30], followed by transcript assembly with StringTie v2.2.3[31]. Transcript models were further refined using PASA v2.3.3[32], and coding sequences were predicted with TransDecoder v5.7.1[33]. Homology-based gene predictions were performed using protein sequences from five evolutionarily related teleost species (*Oryzias latipes*, *Danio rerio*, *Takifugu rubripes*, *Siniperca chuatsi*, and *Micropterus salmoides*). All

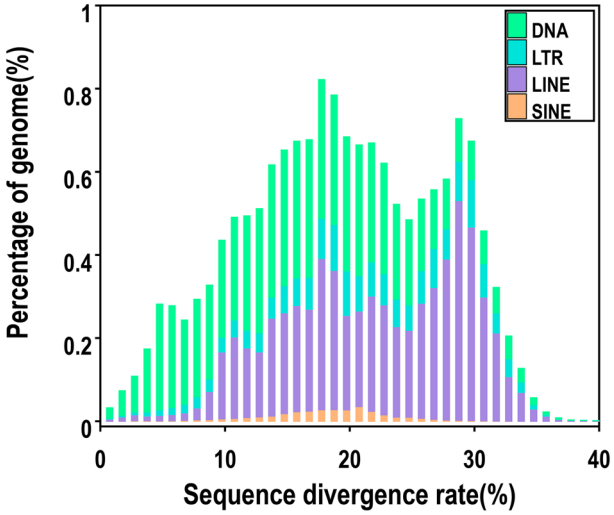

Fig. 4 Four types of TE sequence divergence distribution diagram annotated by RepeatMasker.

Gene structure statistics		
Feature	**Count / average**	**Average length (bp)**
Gene	23,225	15,842.44
Exon	10.26 (average)	175.8
Intron	9.26 (average)	1,516.76
Functional annotation summary		
Database	**Annotated genes**	**Proportion (%)**
Swissprot	20,242	87.16
KEGG	22,558	97.13
KOG	18,632	80.22
GO	16,546	71.24
NR	22,852	98.39
ALL	22,877	98.50

Table 5. Statistics of gene structure and functional annotation in the *S. obscura* genome.

protein sequences were retrieved from the NCBI database and aligned to the *S. obscura* genome assembly using miniprot v0.13[34] with stringent alignment parameters. For *ab initio* gene prediction, we used BRAKER v2.1.6[35], integrating Augustus v3.5.0[36] and GeneMark-ETP v1.31[37], trained on reference proteins from the OrthoDB v12[38] database. The integrated annotation pipeline yielded 22,333 genes supported by RNA-seq evidence, 23,256 genes via homology-based search, and 25,109 genes through de novo prediction. These three evidence-based gene sets were merged using EVidenceModeler v2.1.0[39], resulting in a final consensus set of 23,225 high-confidence protein-coding genes. The final gene models featured an average gene length of 15,842.44 bp, an average coding sequence (CDS) length of 1,803.12 bp, and a mean of 10.26 exons per gene (Table 5).

Functional annotation of the predicted protein sequences was conducted using Diamond v2.1.10[40] against the SwissProt[41], KEGG[42], KOG[43], GO[44], and NR[45] databases, with an e-value cutoff of 1e-5. A total of 22,877 genes were annotated, representing 98.50% of all inferred genes (Fig. 5 and Table 5).

Data Records

The raw sequencing reads of all libraries have been deposited into NCBI SRA database via the accession number PRJNA1245649[46]. The assembled genome has been deposited at Genbank under the accession number GCA_049996615.1[47]. Moreover, data of the genome annotations, predicted coding sequences and protein sequences are available at Figshare[48].

Technical Validation

Assessment of genome assembly. The completeness of the *S. obscura* genome was evaluated using BUSCO v5.7.1[49] against the actinopterygii_odb10 lineage. The assembly exhibited exceptional integrity, with 99.37% of the benchmark genes detected (99.23% single-copy, 0.14% duplicated, 0.52% fragmented, and 0.11% missing; Table 6). Further quality assessment via Merqury v1.3[50] with HiFi reads yielded a QV of 63.22 (error rate $4.76 \times 10{-}^{7}$), while CRAQ v1.10[51] analysis indicated a scaffold anchoring quality (S-AQI) of 99.05%. Detailed per-chromosome metrics are provided in Supplementary Table S1, confirming high chromosomal accuracy and continuity. Additionally, BUSCO evaluation of the predicted proteome against the same database yielded a

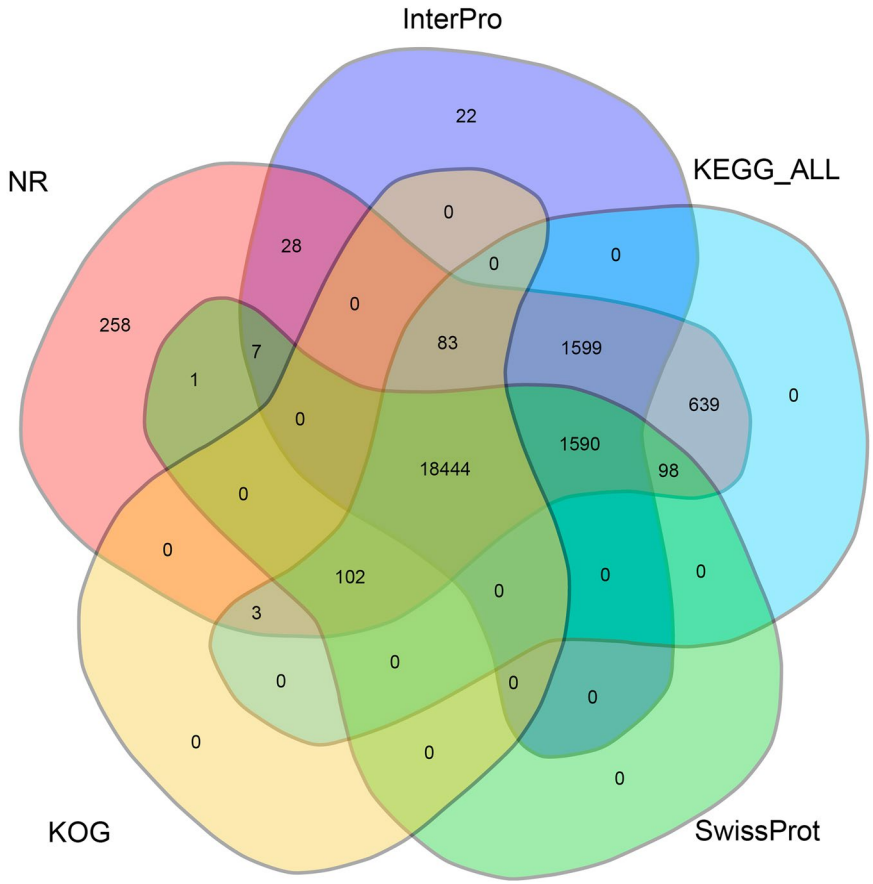

Fig. 5 Venn diagram of function annotations from various databases.

Category	Assembly		Annotation	
	Proteins	Percentage (%)	Proteins	Percentage (%)
Complete BUSCOs	3617	99.37	3608	99.12
Complete Single-Copy BUSCOs	3612	99.23	3582	98.41
Complete Duplicated BUSCOs	5	0.14	26	0.71
Fragmented BUSCOs	19	0.52	9	0.25
Missing BUSCOs	4	0.11	23	0.63
Total BUSCO groups searched	3640	100	3640	100
Database	actinopterygii_odb10		actinopterygii_odb10	

Table 6. BUSCO analysis of the genome assembly and gene annotation.

completeness score of 99.12% (98.41% single-copy and 0.71% duplicated), further validating the high quality of the gene models (Table 6)

To evaluate assembly consistency, mapping rates were determined by aligning PacBio HiFi, Illumina, and RNA-seq reads to the final assembly using Minimap2 v2.24[52], BWA v0.7.17[22] and HISAT2 v2.2.1[30], respectively. The mapping rates were 99.78% for PacBio, 99.18% for Illumina, and 97.36% for RNA-seq reads, further validating the high quality and completeness of the genome. Correspondingly, to assess the structural accuracy of gene models, the genomic distribution of RNA-seq alignments was analyzed via Qualimap v2.3[53]. This revealed that 70.41% of the reads mapped to exonic regions, with only 4.68% in introns and 24.91% in intergenic regions, thereby confirming the robust transcriptional support and structural reliability of the predicted gene structures.

Genome synteny analysis. To investigate whole-genome synteny, the chromosome-level genome assembly of *S. chuatsi*[12] was aligned to that of *S. obscura* using MCscan v0.8[54], and the syntenic relationships were visualized with JCVI v1.1.12[55]. The collinearity analysis revealed a high degree of conserved synteny between the two genomes, supporting the high quality and completeness of our genome assembly (Fig. 6).

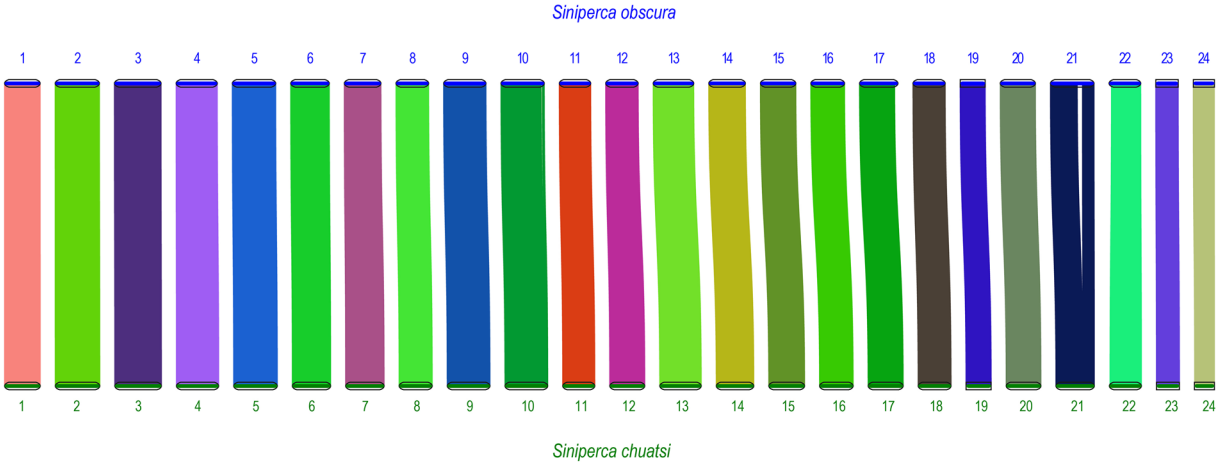

Fig. 6 A synteny analysis of the chromosomes among genomes of *S. obscura* and *S. chuatsi*.

Data availability

All data generated in this study have been deposited in public repositories and are freely available. Raw sequencing data, including PacBio HiFi, Hi-C, MGI short-read, and transcriptome datasets, are available in the NCBI Sequence Read Archive under accession number PRJNA1245649[46]. The chromosome-level genome assembly of *Siniperca obscura* has been deposited in NCBI GenBank under accession GCA_049996615.1[47]. The corresponding genome annotation files have been deposited in the Figshare repository[48].

Code availability

No special codes or scripts were used in this work, and data processing was carried out based on the protocols and manuals of the corresponding bioinformatics software.

Received: 29 July 2025; Accepted: 22 January 2026;
Published online: 02 February 2026

References

1. Froese, R. & Pauly, D. (Fisheries Centre, University of British Columbia Vancouver, BC, 2010).
2. Lu, L., Jiang, J., Zhao, J. & Li, C. Comparative genomics revealed drastic gene difference in two small Chinese perches, Siniperca undulata and S. obscura. *G3: Genes, Genomes, Genetics* **13**, jkad101 (2023).
3. Song, S., Zhao, J. & Li, C. Species delimitation and phylogenetic reconstruction of the sinipercids (Perciformes: Sinipercidae) based on target enrichment of thousands of nuclear coding sequences. *Mol Phylogenet Evol* **111**, 44–55 (2017).
4. Song, Y. *et al.* Effects of the Three Gorges Dam on the mandarin fish larvae (*Siniperca chuatsi*) in the middle reach of the Yangtze River: Spatial gradients in abundance, feeding, growth, and survival. *Ecology of Freshwater Fish* **33**, e12795 (2024).
5. Chen, D.-X. *et al.* The phylogenetic placement of Siniperca obscura base on complete mitochondrial DNA sequence. *Mitochondrial DNA* **25**, 218–219 (2014).
6. Huang, W., Liang, X.-F., Qu, C.-M., Zhao, C. & Cao, L. Isolation and characterization of 31 polymorphic microsatellite markers in Siniperca obscura Nichols. *Conserv Genet Resour* **5**, 153–156 (2013).
7. Qu, C., Liang, X., Huang, W. & Cao, L. Isolation and characterization of 46 novel polymorphic EST-simple sequence repeats (SSR) markers in two Sinipercine fishes (Siniperca) and cross-species amplification. *Int J Mol Sci* **13**, 9534–9544 (2012).
8. Chen, D., Guo, X. & Nie, P. Phylogenetic studies of sinipercid fish (Perciformes: Sinipercidae) based on multiple genes, with first application of an immune-related gene, the virus-induced protein (viperin) gene. *Mol Phylogenet Evol* **55**, 1167–1176 (2010).
9. Li, C., Ortí, G. & Zhao, J. The phylogenetic placement of sinipercid fishes ("Perciformes") revealed by 11 nuclear loci. *Mol Phylogenet Evol* **56**, 1096–1104 (2010).
10. Ding, W. *et al.* A chromosome-level genome assembly of the mandarin fish (*Siniperca chuatsi*). *Frontiers in genetics* **12**, 671650 (2021).
11. He, S. *et al.* Mandarin fish (Sinipercidae) genomes provide insights into innate predatory feeding. *Communications Biology* **3**, 361 (2020).
12. Yang, C. *et al.* Screening of genes related to sex determination and differentiation in mandarin fish (*Siniperca chuatsi*). *Int J Mol Sci* **23**, 7692 (2022).
13. Tu, G.-X. *et al.* Long-read genome assemblies reveal a cis-regulatory landscape associated with phenotypic divergence in two sister Siniperca fish species. *Zoological Research* **44**, 287 (2023).
14. Lu, L., Zhao, J. & Li, C. High-quality genome assembly and annotation of the big-eye mandarin fish (*Siniperca knerii*). *G3: Genes, Genomes. Genetics* **10**, 877–880 (2020).
15. Jiang, M. *et al.* The telomere-to-telomere gap-free reference genome and taxonomic reassessment of Siniperca roulei. *GigaScience* **14**, giaf068 (2025).
16. Chen, S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *Imeta* **2**, e107 (2023).
17. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
18. 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).
19. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods* **18**, 170–175 (2021).
20. Durand, N. C. *et al.* Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell systems* **3**, 95–98 (2016).

21. Dudchenko, O. *et al*. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
22. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics* **25**, 1754–1760 (2009).
23. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell systems* **3**, 99–101 (2016).
24. Wang, X. & Wang, L. GMATA: an integrated software package for genome-scale SSR mining, marker development and viewing. *Frontiers in plant science* **7**, 215951 (2016).
25. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573–580 (1999).
26. Han, Y. & Wessler, S. R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res* **38**, e199–e199 (2010).
27. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**, 9451–9457 (2020).
28. Jurka, J. *et al*. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research* **110**, 462–467 (2005).
29. Hubley, R. *et al*. The Dfam database of repetitive DNA families. *Nucleic Acids Res* **44**, D81–D89 (2016).
30. 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).
31. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology* **33**, 290–295 (2015).
32. Haas, B. J. *et al*. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res* **31**, 5654–5666 (2003).
33. Haas, B. J. *et al*. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols* **8**, 1494–1512 (2013).
34. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics* **39**, btad014 (2023).
35. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR genomics and bioinformatics* **3**, lqaa108 (2021).
36. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34**, W435–W439 (2006).
37. Bruna, T., Lomsadze, A. & Borodovsky, M. GeneMark-ETP: automatic gene finding in eukaryotic genomes in consistency with extrinsic data. *BioRxiv*, 2023–2001 (2023).
38. Kuznetsov, D. *et al*. OrthoDB v11: annotation of orthologs in the widest sampling of organismal diversity. *Nucleic Acids Res* **51**, D445–D451 (2023).
39. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology* **9**, 1–22 (2008).
40. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* **18**, 366–368 (2021).
41. Bairoch, A. *et al*. The universal protein resource (UniProt). *Nucleic Acids Res* **33**, D154–D159 (2005).
42. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27–30 (2000).
43. Tatusov, R. L. *et al*. The COG database: an updated version includes eukaryotes. *BMC bioinformatics* **4**, 1–14 (2003).
44. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. *Nat Genet* **25**, 25–29 (2000).
45. Kanz, C. *et al*. The EMBL nucleotide sequence database. *Nucleic Acids Res* **33**, D29–D33 (2005).
46. *NCBI Sequence Read Archive*. https://identifiers.org/ncbi/insdc.sra:SRP598015 (2025).
47. *NCBI GenBank*. https://identifiers.org/ncbi/insdc.gca:GCA_049996615.1 (2025).
48. Liu, H. Chromosome-level genome assembly of the Siniperca obscura. *Figshare*. https://doi.org/10.6084/m9.figshare.29641379 (2025).
49. 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).
50. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome biology* **21**, 245 (2020).
51. 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. *Nat Commun* **14**, 6556 (2023).
52. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
53. Okonechnikov, K., Conesa, A. & Garcia-Alcalde, F. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics*. **32**, 292–294 (2016).
54. Tang, H. *et al*. Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).
55. Tang, H. *et al*. JCVI: A versatile toolkit for comparative genomics analysis. *iMeta*, e211 (2024).

Acknowledgements

We acknowledge financial support from the National Modern Agriculture Industry Technology System Special Project (CARS-46), Monitoring of Aquatic Resources in Key Waters of Anhui Province (2024BFAFZ02936), Special Fund for Anhui Agriculture Research System (2021-711), Central Public-interest Scientific Institution Basal Research Fund, CAFS (2025XK01, 2025SJHX1, 2023TD37), China-ASEAN Maritime Cooperation Fund (CAMC-2018F), Guangdong Province Rural Revitalization Strategy Special Fund (2023-SJS-00-001).

Author contributions

All authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-026-06678-6.

Correspondence and requests for materials should be addressed to H.L., Y.L. or Y.L.