scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A chromosome-level reference genome assembly of the giant pangasius *(Pangasius sanitwongsei)*

Baojiang Gan[1], Lingjing Wei[1], Yuanxiong Ma[1], Feilong Mo[1], Shan Xiao[1], Yudian Lu[1], Kang Liu[1], Binlan Yang[1], Sheng Zhang[1] ✉ & Haiyang Liu[2] ✉

The giant pangasius (*Pangasius sanitwongsei*) is a critically endangered freshwater species with considerable ecological and economic significance in Southeast Asia and southern China. However, the lack of a high-quality reference genome has limited studies on its genetic adaptation and conservation strategies. Here, we present a chromosome-scale genome assembly of *P. sanitwongsei*, generated using PacBio HiFi long-read sequencing and Hi-C chromatin conformation capture. The final assembly spans 808.85 Mb, with a contig N50 of 18.70 Mb and a scaffold N50 of 28.10 Mb, achieving a BUSCO gene completeness of 99.15%. A total of 23,469 protein-coding genes were annotated, and 96.97% of the genes were functionally characterized. This high-quality genomic resource provides crucial insights into the adaptive evolution of Pangasiidae catfishes and offers a valuable foundation for conservation genomics, adaptive evolution, population genetics, and sustainable aquaculture development.

Background & Summary

Pangasius sanitwongsei, commonly known as the giant pangasius or giant freshwater catfish, belongs to the order Siluriformes, family Pangasiidae, and genus *Pangasius*[1]. This species is endemic to freshwater ecosystems in Southeast Asia, with a distribution concentrated in the Mekong and Chao Phraya River basins[2]. *Pangasius sanitwongsei* is one of the largest freshwater fish in the world, with individuals reaching lengths of over 3 meters and weights of up to 300 kilograms[3]. It inhabits deep-water environments and exhibits an omnivorous but predominantly carnivorous diet, feeding mainly on fish, shrimp, crabs, and other aquatic organisms[4]. Sexual maturity is typically reached at 3~4 years of age, and spawning migrations occur between April and May each year[5]. A distinctive morphological feature of *P. sanitwongsei* is the filamentous extension of the first soft ray in the dorsal, pectoral, and pelvic fins, making it a highly recognizable flagship species within the *Pangasiidae* family[6].

Pangasius sanitwongsei possesses substantial economic value in aquaculture due to its tender flesh, appealing taste, and rapid growth rate[7]. Additionally, its broad head, graceful swimming behavior, and distinctive fin morphology contribute to its ornamental appeal, making it a favored species among aquarium enthusiasts[8]. These attributes underscore the species' potential for both commercial aquaculture and ornamental fish development. However, in recent years, wild populations of *P. sanitwongsei* have undergone a dramatic decline, primarily due to overfishing, environmental pollution, habitat degradation, and the construction of hydropower infrastructure[9–11]. As a result, the species is currently classified as Critically Endangered (CR) on the IUCN Red List of Threatened Species and was designated as a National First-Class Protected Aquatic Wildlife Species in China in 2021.

Previous research on *P. sanitwongsei* has primarily focused on resources conservation[12], breeding[5,13], taxonomy[14,15], and genetic diversity[7]. Molecular resources such as the complete mitochondrial genome and microsatellite markers have been developed[16–18]. In recent years, substantial progress has been made in the genomic research of the Pangasiidae family, including species such as *Pangasianodon hypophthalmus*[19–21]. These genomic studies have provided a solid foundation for investigating phylogenetic evolution, developmental genes, and fatty acid-binding proteins. However, despite its ecological and economic importance, genomic resources for *P. sanitwongsei* remain extremely limited. The only available assembly was generated from short-read Illumina

[1]Aquatic Species Introduction and Breeding Center of Guangxi, Nanning, 530031, China. [2]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. ✉e-mail: 20749078@qq.com; hyliu@prfri.ac.cn

Sequencing Type	Method	Raw data (Gb)	Read Length
Short-reads	DNA	89.96	150 bp PE
Long-reads	PacBio HiFi	43.79	20.94 kb (N50)
Short-reads	Hi-C	138.52	150 bp PE
Short-reads	RNA	11.74	150 bp PE

Table 1. Statistics of sequencing data.

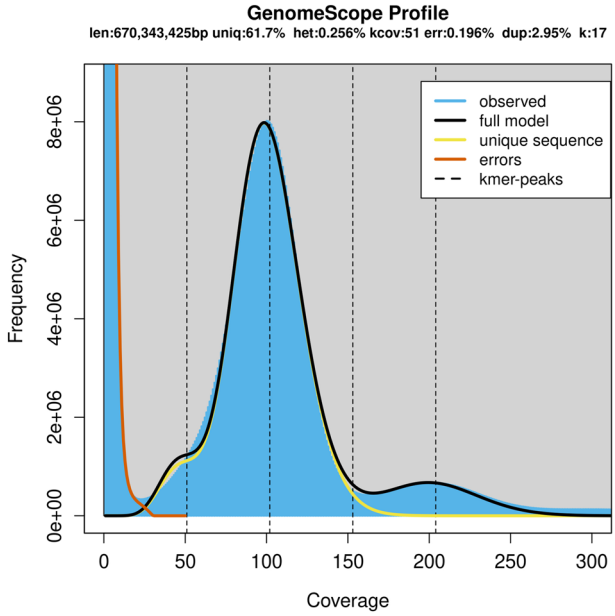

Fig. 1 The K-mer (k = 17) distribution and genome size estimation of *Pangasius sanitwongsei* genome.

data and scaffolded into pseudochromosomes through reference-guided alignment to the *P. hypophthalmus* genome[22]. As a result, a true chromosome-level genome assembly has not yet been available for this species.

Here, we present a first high-quality chromosome-level genome assembly of *P. sanitwongsei*, generated using a combination of PacBio HiFi long-read sequencing, MGI short-read sequencing, and Hi-C scaffolding. The assembled genome spans 808.85 Mb with a scaffold N50 of 28.10 Mb, and a BUSCO completeness score of 97.77%. This high-quality genome provides a valuable resource for evolutionary study, genetic analysis, and molecular breeding of this species.

Methods

Sample collection and DNA extraction. A mature male *Pangasius sanitwongsei* was obtained from the Aquatic Species Introduction and Breeding Center of Guangxi. Muscle tissue was collected for genomic DNA extraction using the DNeasy Blood and Tissue Kit (Qiagen, USA), following the manufacturer's instructions. The integrity and concentration of the extracted DNA were assessed by 1% agarose gel electrophoresis and quantified 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-2024-026).

Genome sequencing. For short-read sequencing, a 350 bp paired-end library was prepared using the MGIEasy PCR-Free DNA Library Prep Set (BGI, China) and sequenced on the DNBSEQ-T7 platform (MGI, China), generating 89.96 Gb of paired-end raw sequencing data (Table 1). Long-read sequencing was conducted using the PacBio Sequel II system with the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, USA), yielding 43.79 Gb of high-fidelity (HiFi) reads with an average read length of 20.94 kb (Table 1). For Hi-C sequencing, approximately 1 g of muscle tissue was used for *in situ* chromatin proximity ligation with the DpnII restriction enzyme, following the manufacturer's protocol (GrandOmics, China). The resulting Hi-C library was sequenced on the DNBSEQ-T7 platform (MGI, China), producing 138.52 Gb of raw Hi-C data (Table 1).

RNA extraction and transcriptome sequencing. Total RNA was extracted from ten tissues (muscle, liver, spleen, kidney, intestine, heart, brain, swim bladder, testis, and skin) using TRIzol reagent. RNA integrity was assessed using an Agilent 2100 Bioanalyzer, and concentrations were measured with a NanoDrop 2000 spectrophotometer. Equal amounts of high-quality RNA from each tissue were pooled to construct a strand-specific cDNA library using the MGIEasy PCR-Free DNA Library Prep Set (BGI, China). The library was sequenced on

Feature	Value
Assembly genome length (Mb)	808.85
GC (%)	39.36
contigs number	160
contig N50 (Mb)	18.70
contig N90 (Mb)	5.02
scaffolds number	53
scaffold N50 (Mb)	28.10
scaffold N90 (Mb)	20.09
Longest scaffold (Mb)	46.06
Average scaffold length (Mb)	15.26
Number of chromosomes	30

Table 2. Summary statistics of *Pangasius sanitwongsei* genome assembly.

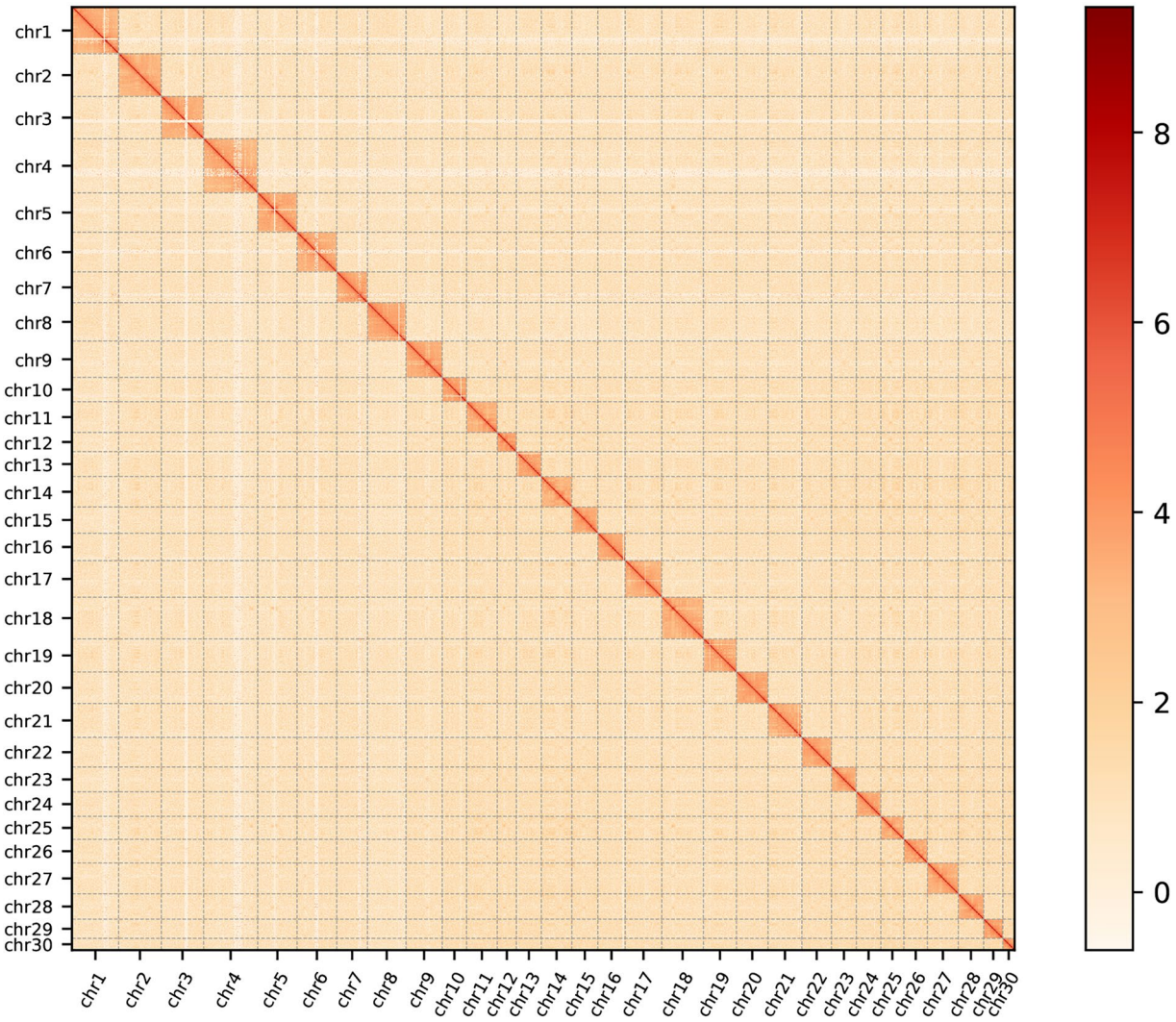

Fig. 2 Hi-C interaction map of the genome at 250 kb resolution.

the DNBSEQ-T7 platform (MGI, China), generating 11.74 Gb of transcriptomic data for genome annotation (Table 1).

Genome size and heterozygosity estimation. To estimate the genome size of *P. sanitwongsei*, a k-mer analysis was performed using high-quality MGI short-read data. First, Jellyfish v2.3.0[23] was used to calculate the frequency of 17-mers and generate the k-mer frequency distribution. The resulting data were then analyzed with

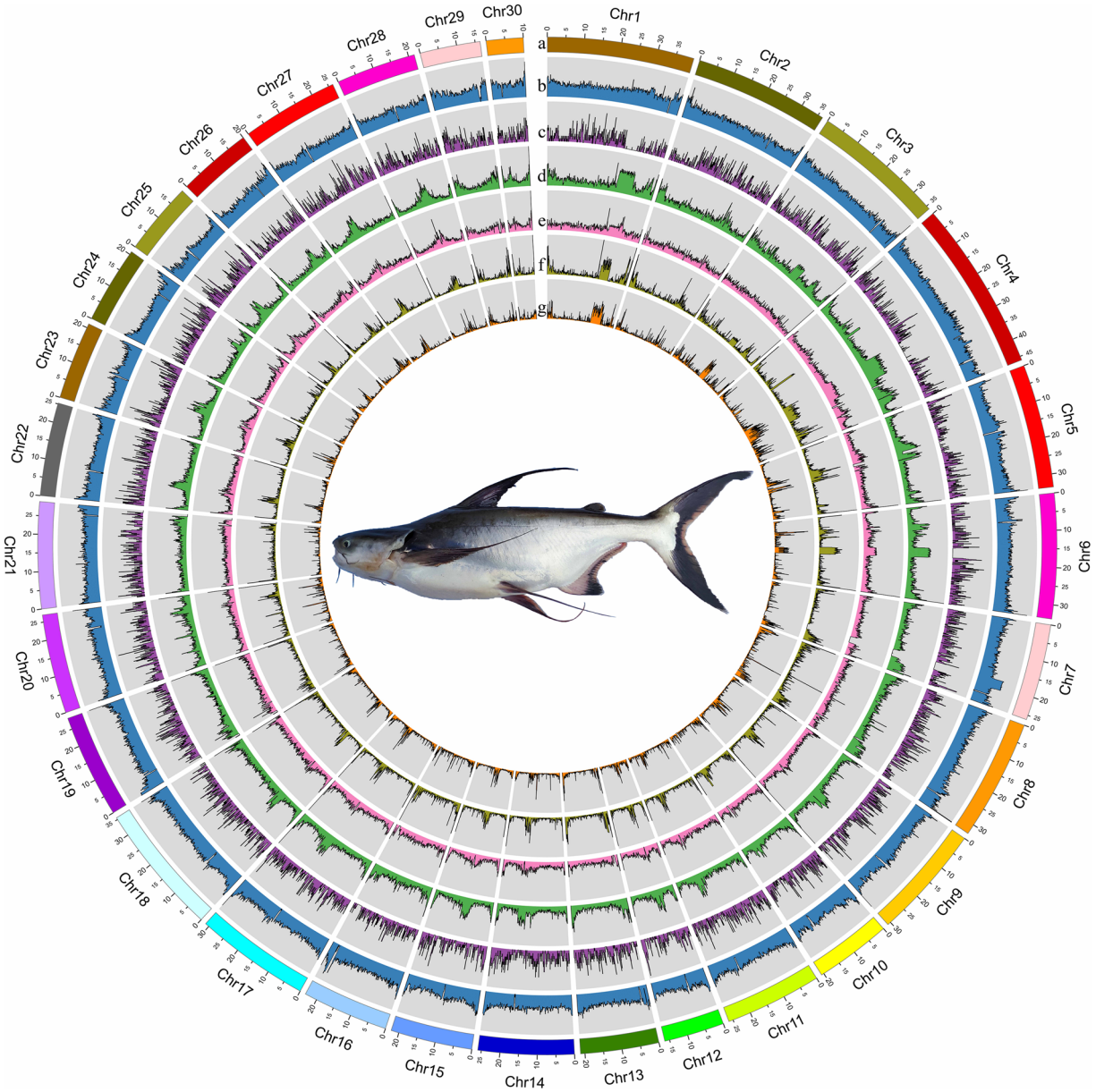

Fig. 3 Features of the *Pangasius sanitwongsei* genome. From outside to inside: (**a**) The 30 pseudo-chromosomes, (**b**) GC content, (**c**) Gene density, (**d**) Repeats content, (**e**) LTR content, (**f**) LINE content and (**g**) DNA-TE content.

GenomeScope v2.0[24], which estimated the genome size to be 670,343,425 bp, with a heterozygosity rate of 0.256% and 61.7% unique sequences (Fig. 1).

Genome assembly. The genome was de novo assembled using Hifiasm v0.19.5[25] with default parameters, resulting in 160 contigs with a total length of 808.85 Mb and a contig N50 of 18.70 Mb (Table 2). For chromosome-level scaffolding, a hybrid strategy integrating Juicer v1.6[26] and 3D-DNA v201008[27] was employed. The workflow began with indexing of the contig-level assembly using BWA v0.7.17[28], followed by Juicer analysis to identify restriction enzyme cutting sites. Clean Hi-C paired-end reads were mapped to the contigs using Juicer, and 3D-DNA was applied according to standard protocols to generate the initial chromosome-scale assembly. Manual curation was performed in two rounds using Juicebox v1.11.08[29] to adjust chromosome boundaries and correct scaffold misassembles, ultimately yielding a high-quality assembly with 30 chromosomes (Figs. 2, 3). The curated assembly was subsequently reprocessed with 3D-DNA to generate per-chromosome scaffolds Fig. 4. The final genome assembly comprised 53 scaffolds, with a maximum scaffold length of 46.06 Mb and a scaffold N50 of 28.10 Mb (Tables 2, 3).

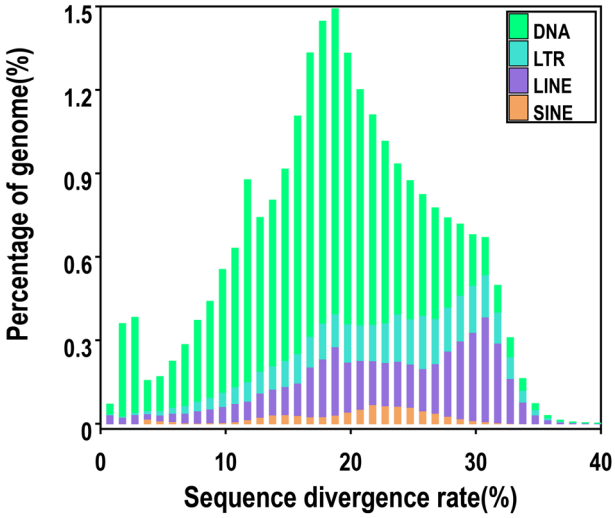

Fig. 4 Four types of TE sequence divergence distribution diagram annotated by RepeatMasker.

Pseudomolecule	Length (bp)	Contig number	GC content(%)
Chr01	39752580	4	39.51
Chr02	36355055	6	39.26
Chr03	36017024	3	39.24
Chr04	46057073	7	39.30
Chr05	33517209	5	39.30
Chr06	33803822	6	39.17
Chr07	26188897	2	39.96
Chr08	32738081	4	39.66
Chr09	30900743	3	39.11
Chr10	20683005	4	39.26
Chr11	26177216	5	39.37
Chr12	16334559	4	39.33
Chr13	21189289	4	39.52
Chr14	25957842	4	39.25
Chr15	22153762	4	39.14
Chr16	23321922	3	39.57
Chr17	31116038	6	39.22
Chr18	35524820	9	39.29
Chr19	28100061	7	39.33
Chr20	26921474	3	39.10
Chr21	28771072	4	39.16
Chr22	25267160	3	39.24
Chr23	21137666	3	39.12
Chr24	20857103	11	39.41
Chr25	19637615	5	39.27
Chr26	20085813	6	39.53
Chr27	26242051	2	39.00
Chr28	21599743	4	39.02
Chr29	16349843	4	39.37
Chr30	10008409	2	39.78
Total	802766947	137	39.36

Table 3. Pseudo-chromosome length statistics after Hi-C assisted assembly of *Pangasius sanitwongsei*.

Repeat annotation. Given the biological importance of tandem repeats, a genome-wide survey of repetitive elements was conducted using GMATA v2.2.1[30] and Tandem Repeats Finder (TRF) v4.10.0[31], both with default parameters. GMATA was employed to specifically detect simple sequence repeats (SSRs) with short repeat

Type	Length (Bp)	% in genome
DNA	189440298	23.43
LINE	63918510	7.91
LTR	38704290	4.79
SINE	8910003	1.10
Satellite	6977313	0.86
RC	6877667	0.85
Unknown	5593553	0.69
Total	321165151	39.71

Table 4. Repetitive sequences in the genome of *Pangasius sanitwongsei*.

Item	Number	Average length(bp)
Gene	23,469	14,818.67
Exon	9.42	173.59
Intron	8.42	1,565.06
Database	Number	Percentage (%)
Swissprot	19,768	84.20
KEGG	22,291	84.23
KOG	18,193	94.98
GO	16,135	77.52
NR	22,758	68.75
ALL	22,807	96.97

Table 5. Gene structures and function annotation of *Pangasius sanitwongsei*.

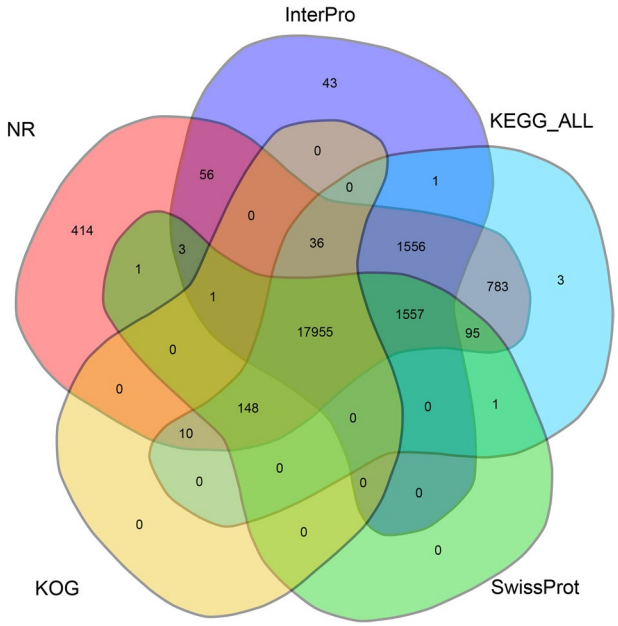

Fig. 5 Venn diagram of function annotations from five databases.

units, while TRF was used to identify a broader range of tandem repeats. For dispersed repetitive elements, the pipeline began with MITE-Hunter[32] to identify miniature inverted-repeat transposable elements (MITEs), generating a dedicated MITE library. The genome was then hard-masked (repetitive regions were replaced with "N"), and RepeatModeler[33] was used for de novo identification of additional repeat elements to construct a RepMod.lib library. Unclassified elements within RepMod.lib were subsequently categorized using TEclass v2.4[34]. Comprehensive repeat library was compiled by integrating MITE.lib, RepMod.lib, and Repbase[35]. This combined library was then used to annotate repetitive sequences across the genome using RepeatMasker[36].

Type	Number	Percent (%)
Complete BUSCOs (C)	3609	99.15
Complete and single-copy BUSCOs (S)	3594	98.74
Complete and duplicated BUSCOs (D)	15	0.41
Fragmented BUSCOs (F)	11	0.30
Missing BUSCOs (M)	20	0.55
Total BUSCO groups searched	3,640	100.00

Table 6. BUSCO analysis of the genome assembly.

The repeats annotation results revealed that dispersed repeats accounted for 38.94% of the genome (Table 4). Among the transposable elements (TEs), DNA transposons were the most abundant (23.43%), followed by long interspersed nuclear elements (LINEs, 7.91%), long terminal repeat (LTR) retrotransposons (4.79%), and short interspersed nuclear elements (SINEs, 1.10%). In total, repetitive sequences spanned 321165151 bp, representing 39.71% of the genome (Table 4).

Gene prediction and function assignment. Gene annotation was performed using a comprehensive three-tiered evidence integration pipeline, incorporating transcriptomic data, homology-based evidence, and ab initio predictions. Transcriptomic evidence was obtained by aligning Illumina RNA-seq reads to the genome assembly using HISAT2 v2.2.1[37], followed by transcript assembly with StringTie v2.2.3[38]. Putative coding sequences (CDS) were predicted using TransDecoder v5.7.1[39] with default parameters. Homology-based annotation utilized protein sequences from five evolutionarily conserved species, with protein-to-genome alignments conducted using miniprot[40]. For ab initio gene prediction, BRAKER v2.1.5[41] was employed, integrating Augustus v3.5.0[42] and GeneMark-ETP v1[43]. The three sources of evidence were integrated using EVidenceModeler v2.1.0[44], resulting in 23,469 high-confidence protein-coding genes. These genes had an average length of 14,818.67 bp, an average coding sequence length of 173.59 bp, and contained an average of 9.42 exons per gene (Table 5).

Functional annotation of the predicted proteins was conducted using Diamond v2.1.10[45] against the Swiss-Prot[46], KEGG[47], KOG[48], GO[49], and NR[50] databases, with an e-value cutoff of 1e-5. A total of 22,807 genes were successfully annotated, accounting for 96.97% of all predicted genes (Fig. 5 and Table 5).

Data Records

All sequencing and assembly data generated in this study have been deposited in public repositories. The raw sequencing datasets, including MGI short-read, PacBio HiFi, Hi-C, and RNA-Seq data, are available in the NCBI Sequence Read Archive (SRA) under the accession number SRP579478[51]. The chromosome-level genome assembly has been deposited in NCBI GenBank under the accession GCA_051225755.1[52]. In addition, the genome annotation files including the GFF annotation file, predicted coding sequences, and protein sequences can be obtained from Figshare[53].

Technical Validation

Assessment of genome assembly. The accuracy of the *P. sanitwongsei* genome assembly was evaluated by assessing its completeness using the conserved metazoan gene set *actinopterygii_odb10* from BUSCO v5.8.3[54]. The analysis revealed a high level of completeness, with an overall score of 99.15%. Specifically, 98.74% of the genes were complete and single-copy, 0.41% were complete and duplicated, 0.30% were fragmented, and 0.55% were missing. These results indicate that the *P. sanitwongsei* genome assembly is of high quality (Table 6). The earlier genome for *P. sanitwongsei* was assembled exclusively from Illumina short reads and scaffolded through reference-guided alignment to *Pangasianodon hypophthalmus*[22]. Although reported as a chromosome-scale assembly, it was constructed without long-read sequencing or Hi-C data, and the resulting assembly contains 29 chromosomes with a BUSCO completeness of 97.58%. These limitations reduce its reliability for structural genomic analyses, evolutionary studies, and high-quality gene annotation. In contrast, our newly generated assembly integrates PacBio HiFi long-read and Hi-C technologies, anchors 30 chromosomes, and achieves a markedly higher BUSCO completeness of 99.15%, providing a substantially improved and more robust genomic resource for downstream applications.

To further assess assembly accuracy, mapping rates of PacBio HiFi, Illumina short reads, and RNA-seq data were calculated by aligning them to the final assembly using Minimap2 v2.24[55], BWA v0.7.17[28] and HISAT2 v2.2.1[37], respectively. The mapping rates were 99.71% for PacBio, 99.06% for Illumina, and 97.23% for RNA-seq reads, further validating the high quality and completeness of the assembled genome.

Genome synteny analysis. To investigate whole-genome synteny, we performed MCscan v0.8[56] analysis using the CDS sequences of annotated protein-coding genes from *Pangasianodon hypophthalmus*[19] and *P. sanitwongsei*. Gene-level syntenic relationships were subsequently visualized with JCVI v1.1.12[57]. The collinearity analysis revealed a substantial level of conserved synteny between the two genomes, supporting the accuracy and completeness of our assembly (Fig. 6A). Furthermore, we conducted a whole-genome alignment between *P. sanitwongsei* and *P. hypophthalmus* using MUMmer v4.0.0[58], and visualized structural variation and syntenic relationships with SyRI[59], which revealed extensive DNA-level syntenic blocks alongside discrete regions exhibiting pronounced sequence divergence (Fig. 6B).

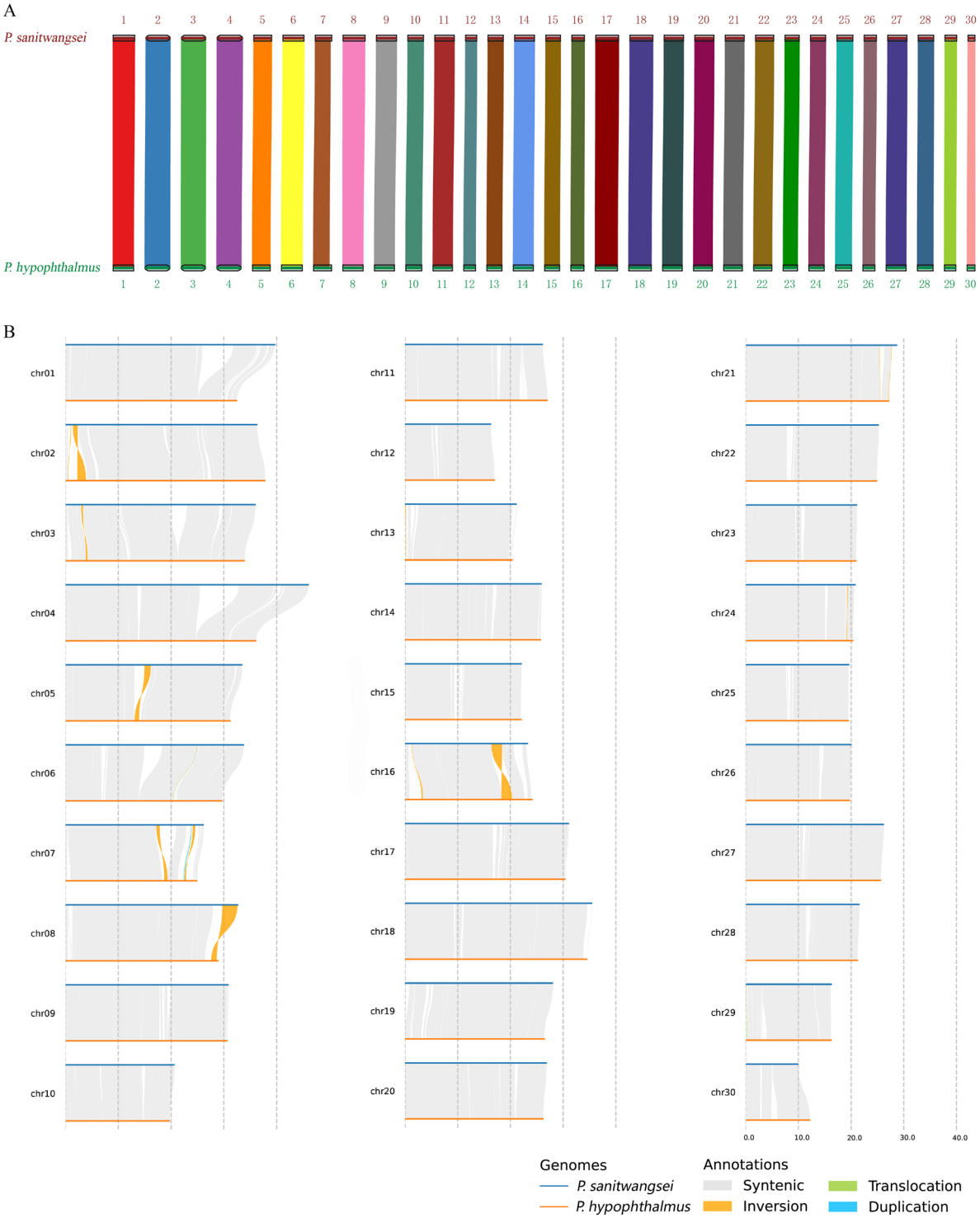

Fig. 6 Gene-level synteny (**A**) and DNA-level synteny (**B**) between *Pangasius sanitwongsei* and *Pangasianodon hypophthalmus*.

Data availability

The raw sequencing datasets have been deposited in the NCBI Sequence Read Archive under accession SRP579478[51]. The assembled *P. sanitwongsei* genome is available in GenBank under accession GCA_051225755.1[52], and the corresponding genome annotation files has been archived on Figshare (https://doi.org/10.6084/m9.figshare.29715215)[53].

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: 1 August 2025; Accepted: 11 December 2025;
Published online: 16 December 2025

References

1. Froese, R. & Pauly, D. (Fisheries Centre, University of British Columbia Los Baños, Philippines, 2010).
2. Kang, B. & Huang, X. Mekong fishes: Biogeography, migration, resources, threats, and conservation. *Rev Fish Sci Aquac* **30**, 170–194 (2022).
3. Roberts, T. R. & Vidthayanon, C. Systematic revision of the Asian catfish family Pangasiidae, with biological observations and descriptions of three new species. *Proceedings of the Academy of Natural Sciences of Philadelphia*, 97-143 (1991).
4. Roberts, T. R. & Baird, I. G. Traditional fisheries and fish ecology on the Mekong River at Khone Waterfalls in southern Laos. *Natural History Bulletin of the Siam Society* **43**, 219–262 (1995).
5. Chanthasoo, M., Wiwatcharakoset, S. & Lisanga, S. Breeding of Chao Phaya Giant catfish (Pangasius Sanitwongsei) (1990).
6. Sutthi, N., Panase, A., Phinrub, W., Srisuttha, P. & Panase, P. Cold shock and its effect on biochemical indices, cortisol and electrolyte changes in Chao Phraya catfish, Pangasius sanitwongsei smith, 1931. *Comparative Clinical Pathology* **31**, 757–764 (2022).
7. Na-Nakorn, U. *et al*. Genetic Diversity of the Vulnerable Pangasius sanitwongsei using Microsatellite DNA and 16S rRNA. *Journal of Fisheries and Environment* **33**, 24–40 (2009).
8. Makinen, T., Weyl, O. L. F., Van der Walt, K.-A. & Swartz, E. R. First record of an introduction of the giant pangasius, Pangasius sanitwongsei Smith 1931, into an African river. *African Zoology* **48**, 388–391 (2013).
9. Hogan, Z., Na-Nakorn, U. & Kong, H. Threatened fishes of the world: Pangasius sanitwongsei Smith 1931 (Siluriformes: Pangasiidae). *Environmental Biology of Fishes* **84**, 305–306 (2009).
10. Campbell, T., Pin, K., Ngor, P. B. & Hogan, Z. Conserving Mekong megafishes: Current status and critical threats in Cambodia. *Water* **12**, 1820 (2020).
11. Baird, I. G. & Hogan, Z. S. Hydropower Dam development and fish biodiversity in the Mekong River Basin: A review. *Water* **15**, 1352 (2023).
12. Jutagate, T. & Rattanachai, A. Inland fishery resource enhancement and conservation in Thailand. Inland fisheries resource enhancement and conservation in Asia 133 (2010).
13. Kitcharoen, N., Nakkham, K. & Mengumphan, K. A study on growth performance of interspecific crosses-hybrid cat fish spices: Buk Siam hybrid catfish (male Pangasianodon gigas x female P. hypophthalmus) Pangosius larnaudii and Pangasius sanitwongsei (2022).
14. Karinthanyakit, W. & Jondeung, A. Molecular phylogenetic relationships of pangasiid and schilbid catfishes in Thailand. *J Fish Biol* **80**, 2549–2570 (2012).
15. Duong, T. Y. *et al*. Mitophylogeny of Pangasiid catfishes and its taxonomic implications for Pangasiidae and the suborder Siluroidei. *Zoological studies* **62**, e48 (2023).
16. Na-Nakorn, U., Sriphairoj, K., Sukmanomon, S., Poompuang, S. & Kamonrat, W. Polymorphic microsatellite primers developed from DNA of the endangered Mekong giant catfish, Pangasianodon gigas (Chevey) and cross-species amplification in three species of Pangasius. *Molecular Ecology Notes* **6**, 1174–1176 (2006).
17. Wei, L. *et al*. Complete mitochondrial genome and phylogenetic position of Pangasius sanitwongsei (Siluriformes: Pangasiidae). *Mitochondrial DNA Part B* **5**, 945–946 (2020).
18. Sriphairoj, K., Na-Nakorn, U. & Klinbunga, S. Species identification of non-hybrid and hybrid Pangasiid catfish using polymerase chain reaction-restriction fragment length polymorphism. *Agriculture and Natural Resources* **52**, 99–105 (2018).
19. Gao, Z. *et al*. A chromosome-level genome assembly of the striped catfish (Pangasianodon hypophthalmus). *Genomics* **113**, 3349–3356 (2021).
20. Hai, D. M. *et al*. A high-quality genome assembly of striped catfish (pangasianodon hypophthalmus) based on highly accurate long-read hifi sequencing data. *Genes* **13**, 923 (2022).
21. Kim, O. T. P. *et al*. A draft genome of the striped catfish, Pangasianodon hypophthalmus, for comparative analysis of genes relevant to development and a resource for aquaculture improvement. *Bmc Genomics* **19**, 733 (2018).
22. Wen, M. *et al*. An ancient truncated duplication of the anti-Müllerian hormone receptor type 2 gene is a potential conserved master sex determinant in the Pangasiidae catfish family. *Mol Ecol Resour* **22**, 2411–2428 (2022).
23. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
24. 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).
25. 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).
26. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell systems* **3**, 95–98 (2016).
27. Dudchenko, O. *et al*. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
28. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics* **25**, 1754–1760 (2009).
29. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell systems* **3**, 99–101 (2016).
30. 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).
31. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573–580 (1999).
32. 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).
33. 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).
34. Abrusán, G., Grundmann, N., DeMester, L. & Makalowski, W. TEclass—a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* **25**, 1329–1330 (2009).
35. Jurka, J. *et al*. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and genome research* **110**, 462–467 (2005).
36. Hubley, R. *et al*. The Dfam database of repetitive DNA families. *Nucleic Acids Res* **44**, D81–D89 (2016).
37. 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).
38. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology* **33**, 290–295 (2015).

39. 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).
40. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics* **39**, btad014 (2023).
41. 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).
42. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34**, W435–W439 (2006).
43. Bruna, T., Lomsadze, A. & Borodovsky, M. GeneMark-ETP: automatic gene finding in eukaryotic genomes in consistency with extrinsic data. BioRxiv, 2023-2001 (2023).
44. 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).
45. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* **18**, 366–368 (2021).
46. Bairoch, A. *et al*. The universal protein resource (UniProt). *Nucleic Acids Res* **33**, D154–D159 (2005).
47. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27–30 (2000).
48. Tatusov, R. L. *et al*. The COG database: an updated version includes eukaryotes. *BMC bioinformatics* **4**, 1–14 (2003).
49. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. *Nat Genet* **25**, 25–29 (2000).
50. Kanz, C. *et al*. The EMBL nucleotide sequence database. *Nucleic Acids Res* **33**, D29–D33 (2005).
51. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP579478 (2025).
52. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_051225755.1 (2025).
53. Liu, H. Chromosome-level genome assembly of the Siniperca obscura. *Figshare* https://doi.org/10.6084/m9.figshare.29715215 (2025).
54. 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).
55. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
56. Tang, H. *et al*. Synteny and collinearity in plant genomes. *Science* **320**, 486–488 (2008).
57. Tang, H. *et al*. JCVI: A versatile toolkit for comparative genomics analysis. iMeta, e211 (2024).
58. Marçais, G. *et al*. MUMmer4: A fast and versatile genome alignment system. *PLoS computational biology* **14**, e1005944 (2018).
59. Goel, M., Sun, H., Jiao, W.-B. & Schneeberger, K. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome biology* **20**, 277 (2019).

Acknowledgements

We acknowledge financial support from the Guangxi Key R&D Program Agriculture and Rural Areas (AB2506910048), National Modern Agriculture Industry Technology System Special Project (CARS-46), 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

Correspondence and requests for materials should be addressed to S.Z. or H.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.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Cre ative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025