scientific data

OPEN

DATA DESCRIPTOR

A telomere-to-telomere gap-free genome assembly of the striped catfish (*Pangasianodon hypophthalmus*)

Xinhui Zhang[1,2,6], Yu Huang[3,6], Wenchuan Zhou[4], Jieming Chen[2], Zhifei Zhu[5], Yun Xu[5], Xinwen Wang[5], Junmin Xu[2,5] & Qiong Shi[1,2,5] ✉

The well-known striped catfish (*Pangasianodon hypophthalmus*), belonging to the order Siluriformes and the family Pangasiidae, has become an important freshwater economic fish species due to its outstanding growth performance, environmental adaptability and disease resistance. In this study, we reported the first telomere-to-telomere (T2T) gap-free genome assembly of striped catfish, generated by integrating MGI short-read, ONT ultra-long read, PacBio HiFi long read and Hi-C data. A total of 772.03 Mb genome sequences were successfully anchored onto 30 chromosomes, demonstrating exceptional contiguity with a high contig N50 of 26.47 Mb. A comprehensive annotation revealed precise structure of telomeric repeats and centromeric region within each chromosome. In the assembled genome, repetitive elements accounted for 39.53% (305.21 Mb), with DNA transposons (20.77%) as the predominant repeat type. A total of 24,596 protein-coding genes were predicted (BUSCO: 96.21%), and 99.22% of these genes were functionally annotated. Our combined results about identification of telomeric repeats and centromeric regions, BUSCO assessment (99.45%), mapping coverage (99.55%), and Clipping information for Revealing Assembly Quality (CRAQ: 99.62%) support the high quality of this genome assembly. These complete chromosomal sequences will serve as a valuable genetic resource for in-depth biological investigations, evolutionary studies, comparative genomics, and molecular breeding to improve economic value of the striped catfish.

Background & Summary

Aquaculture represents the fastest-growing sector within global agro-food industries, with Asian production accounting for over 70% of the worldwide annual output, positioning it as a critical component for addressing international food security challenges. Notably, many commercially important aquaculture species in Asia exhibit significant sexual dimorphism in key production traits, including growth rates and age at sexual maturation. In these species, one sex typically demonstrates superior performance, creating substantial impetus for development of sex control technologies in commercial aquaculture operations[1].

Striped catfish (*Pangasianodon hypophthalmus*), a well-known fish species under the order Siluriformes and the family Pangasiidae, is primarily cultured in Southeast Asian countries. In recent years, its global aquaculture production and market demand have continued to rise remarkably due to its various advantages, including rapid growth, high yield, tender and boneless flesh, and affordable pricing[2,3]. Investigating sexual growth dimorphism in *Pangasius* fishes holds great significance for optimizing aquaculture strategies. Given the superior growth performance of male, mono-sex aquaculture achieved through biotechnological interventions such as hormonal

[1]Laboratory of Aquatic Genomics, College of Life Sciences and Oceanography, Shenzhen University, Shenzhen, 518057, China. [2]Shenzhen Key Lab of Marine Genomics, Guangdong Provincial Key Lab of Molecular Breeding in Marine Economic Animals, BGI Academy of Marine Sciences, Shenzhen, 518081, China. [3]State Key Laboratory of Breeding Biotechnology and Sustainable Aquaculture, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan, 430072, China. [4]Shenzhen Ocean Development and Promotion Center, Shenzhen, 518000, China. [5]Shenzhen China Fishery Technology Co. Ltd., Shenzhen, 518119, China. [6]These authors contributed equally: Xinhui Zhang, Yu Huang. ✉e-mail: shiqiong@szu.edu.cn; shiqiong@genomics.cn

sex reversal or other sex control techniques could substantially enhance production efficiency and economic returns[3]. Artificial sex manipulation techniques, such as hormone-induced sex change or molecular genetic approaches, may enable large-scale production of mono-sex fry. Furthermore, elucidating molecular mechanisms for sex-related growth differences can provide a theoretical foundation for practical genetic breeding, facilitating development of novel strains or varieties with improved growth performance and genetic stability[4].

With the rapid advancement of high-throughput sequencing technologies and the significant reduction in sequencing costs, whole-genome sequencing and re-sequencing have been increasingly applied to decipher key economic traits in various animals and plants, enabling identification of valuable molecular markers associated with these traits. In recent years, maturation of various high-throughput technologies has accelerated discovery of sex-determining genes and development of sex-specific molecular markers in diverse fishes, leading to substantial progress in this field[5,6]. Therefore, investigations on fish sex-determination mechanisms can integrate fundamental theoretical exploration with practical applications. Sex-controlled breeding, particularly production of mono-sex populations, offers numerous advantageous traits such as accelerated growth and improved feed conversion efficiency. Notably, the successful commercialization of all-male yellow catfish (*Pelteobagrus fulvidraco*) has spurred extensive research on mono-sex breeding in many other economically important fish species, demonstrating promising application prospects[7].

Compared with mammals and birds, the genomic divergence between males and females is typically smaller in fish. Therefore, obtaining high-quality genome assemblies is crucial for the accurate characterization of sexual dimorphism in examined fishes. Currently, two striped catfish genome assemblies have been reported[8,9]. For a genome designed to serve as a good reference, particularly for detailed research on sex determination, selection of a heterogametic sex individual (e.g., XY individual for the XX/XY sex-determination system, and ZW individual for the ZZ/ZW sex-determination system) as the representative for whole-genome sequencing is of great strategic significance. Furthermore, previous studies have confirmed that striped catfish exhibits an XX/XY sex-determination system[9]. Here, to improve these genetic resources, we constructed a telomere-to-telomere (T2T) genome assembly of a male striped catfish after integrating MGI short-read, PacBio HiFi long-read, ONT (Oxford Nanopore Technologies) ultra-long and Hi-C sequencing data. This assembly was then rigorously assessed for quality, and its key genomic features were systematically characterized. In fact, this gap-free reference assembly represents a substantial improvement over any previously assembled genome of this species. It will not only facilitate research in population genetics and evolutionary studies, but also provide a critical genetic resource for the molecular breeding of mono-sex fry in this economically important fish species.

Methods

Sample collection. A one-year-old male striped catfish (Fig. 1a; sex was identified through gonadal dissection) was collected from an aquaculture facility of the Pearl River Fisheries Research Institute, which is located in Guangzhou City, Guangdong Province, China. Muscle tissue was sampled for DNA sequencing. Additionally, eight distinct tissues including gill, liver, spleen, eye, muscle, testis, kidney and skin were collected for RNA sequencing (Table 1). These samples were rapidly frozen in liquid nitrogen and subsequently maintained at −80 °C before use. The sampling procedure and experimental workflow were performed in accordance with the guidelines and approval of the Animal Ethics Committee of Institute of Hydrobiology, Chinese Academy of Sciences (Number: IHB1LL12024044).

DNA extraction and genome sequencing. Genomic DNA (gDNA) was extracted from muscle tissue using the DNeasy Blood & Tissue Kit (Qiagen, Valencia, CA, USA) following the manufacturer's protocols[10]. Quantification, fragment size and purity of the extracted gDNA were evaluated via a Qubit Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA), 0.75% agarose gel electrophoresis and an Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA), respectively. High-quality gDNA was prepared for subsequent library construction.

For MGI short-read sequencing, paired-end sequencing of a library with an insert size of 350 bp was executed on the DNBSEQ-T7 platform (MGI, Shenzhen, China), yielding 91.17 Gb of raw data. The raw data underwent quality control using fastp v0.12.6[11] with the following parameters: -n 0 -f 5 -F 5 -t 5 -T 5. Consequently, a total of 84.55 Gb of clean data were obtained, which were subsequently utilized for data error correction and genome-size estimation (Table 1).

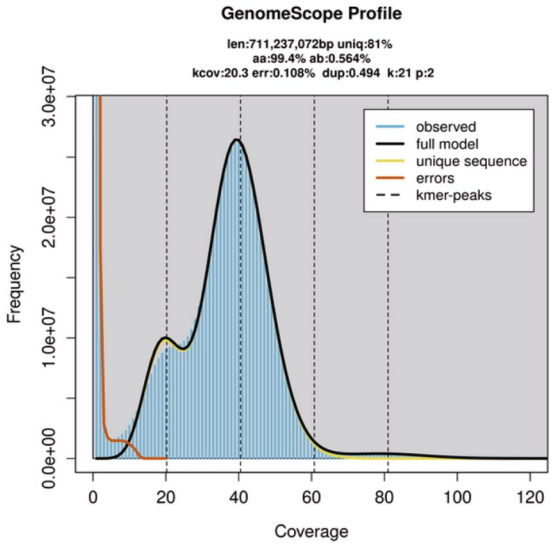
For the PacBio HiFi sequencing, two SMRTbell libraries were constructed and sequenced on a PacBio Sequel II system (Pacific Biosciences, Menlo Park, CA, USA) using the circular consensus sequencing (CCS) technology. A total of 41.99 Gb of HiFi reads with an N50 of 20,907 bp were obtained (Table 1) using the CCS v6.0.0[12] software with the optimized parameter "-min-passes 3".

Three ultra-long libraries were constructed using Oxford Nanopore Technologies (ONT) protocols, which were sequenced on a PromethION platform (Oxford Nanopore Technologies Co., Littlemore, Oxford, UK). Raw reads were initially processed to eliminate those with a quality value (QV) lower than 7 by using the NanoFilt v2.8.0[13] software. Finally, 27.27 Gb of sequencing data were generated (Table 1), with an N50 of 68,067 bp.

For the high-throughput chromosome conformation capture (Hi-C) sequencing, two Hi-C libraries were constructed using a GrandOmics Hi-C kit (GrandOmics, Wuhan, Hubei, China) following the manufacturer's protocol. Hi-C Libraries were then sequenced using the DNBSEQ-T7 platform (MGI) with a 150-bp paired-end mode. A total of 109.09 Gb raw data were generated. Subsequently, fastp v0.12.6[11] was used to trim adaptor sequences and low-quality reads. Finally, 99.87 Gb of Hi-C clean data were retained for chromosome assembly (Table 1).

RNA extraction and transcriptome sequencing. Total RNA was extracted from eight tissue samples using TRIZOL Reagent (Invitrogen, Carlsbad, CA, USA) following the manufacturer's protocol. RNA

a



b

Fig. 1 Striped catfish and its whole-genome sequence distribution. (**a**) A morphological image of the sequenced striped catfish. (**b**) A k-mer (21-mer) distribution curve for estimation of the genome size.

Type	Library type	Raw data (Gb)	Clean data (Gb)	Read N50/ length (bp)	Coverage of the genome (×)
DNA	MGI	91.17	84.55	150	118
	PacBio HiFi	/	41.99	20,907*	59
	ONT Ultra-long	/	27.27	68,067*	38
	Hi-C	109.09	99.87	150	153
RNA	Eye	13.38	11.99	150	/
	Muscle	9.84	8.81	150	/
	Skin	9.69	8.86	150	/
	Liver	6.54	5.81	150	/
	Gill	7.23	6.43	150	/
	Kidney	7.28	6.62	150	/
	Testis	7.95	7.31	150	/
	Spleen	5.52	5.08	150	/

Table 1. Sequencing data of the striped catfish genome and transcriptomes. *For the PacBio HiFi and ONT Utra-long sequencing, this number is N50 of reads; for others, it denotes read length.

concentration and purity were determined by spectrophotometric analysis (NanoDrop 8000 Spectrophotometer; Thermo Fisher Scientific, Waltham, MA, USA), while RNA integrity was assessed using capillary electrophoresis (Agilent 2100 Bioanalyzer; Agilent Technologies, Santa Clara, CA, USA). To ensure high-quality data for downstream transcriptomic analysis, only RNA samples meeting stringent quality criteria (OD260/280 ratio $\geq$ 1.8 and RNA integrity number [RIN] $\geq$ 7.0) were subjected to RNA sequencing (RNA-seq).

The purified RNA was used to construct a cDNA library according to the manufacturer's protocol. High-throughput sequencing was performed on an Illumina HiSeq X Ten platform (Illumina, San Diego, CA, USA), yielding 67.43 Gb of raw data. To ensure data quality, low-quality reads and adapter sequences were filtered out using fastp v0.12.6[11], resulting in 60.91 Gb of high-quality clean data (Table 1). These clean reads were subsequently used for downstream gene structure prediction.

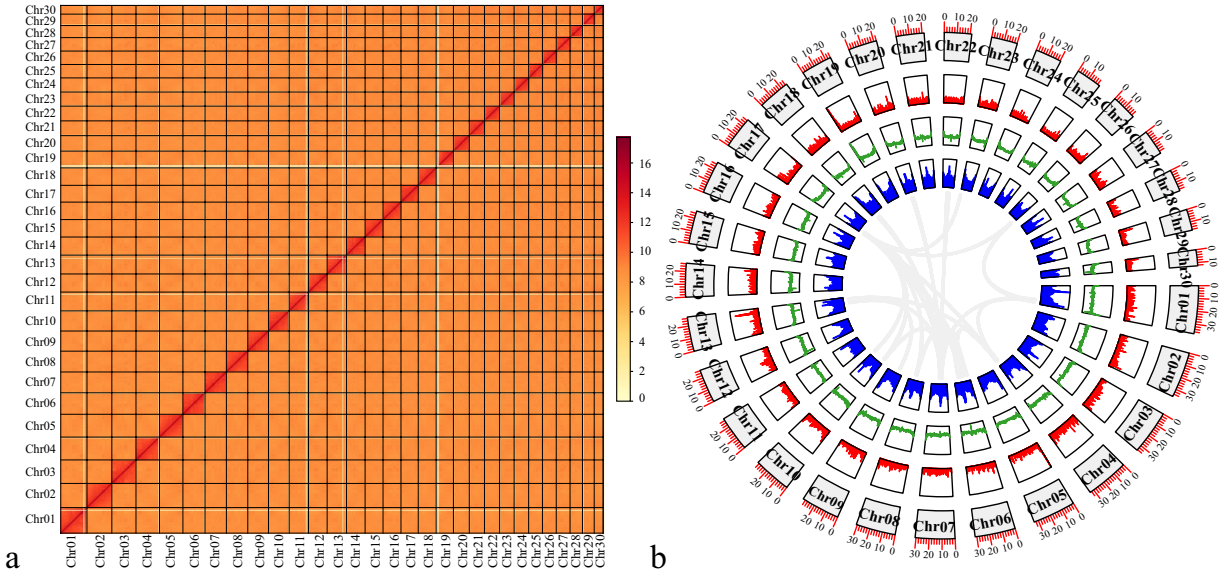


Fig. 2 The first T2T genome assembly of striped catfish. (**a**) Genome-wide chromatin interactions at a 500-kb resolution. Color blocks represent corresponding interactions, with various strengths from white (low) to red (high). (**b**) A Circos plot of the main genome features. From outside to inside include the 30 chromosomes, gene density, GC content, repetitive sequences density, and a colinear relationship among chromosomes of the striped catfish genome assembly. Note that the density calculation window is set as 100 kb.

Genome-size estimation and construction of a T2T genome assembly. We performed genome size estimation for striped catfish using a k-mer-based approach. Initial k-mer counting was conducted by Jellyfish v2.2.10[14] with a k-mer size of 21 and the optimized parameters "-m 21 -s 10 G -C" to ensure comprehensive k-mer coverage. The resulting k-mer frequency distribution curve (Fig. 1b) was then analyzed using GenomeScope v2.0[15] to estimate key genomic characteristics. Our analysis showed that the genome size of striped catfish is approximately 711.23 Mb, with an estimated heterozygosity of 0.56% (see Fig. 1b).

Primary contigs were initially generated by assembling PacBio HiFi and ONT sequencing data using Hifiasm v0.19.8[16] with default parameters. Then, purge_dups v1.2.5[17] was employed to remove haplotypic and heterozygous duplications from the *de novo* assembly, yielding a final assembly of 772.01 Mb in total length.

We performed chromosome-level scaffolding of the primary genome assembly using Hi-C data. First, high-quality Hi-C paired-end reads were aligned to the draft contigs using Bowtie2 v2.2.5 with default parameters[18]. Valid chromatin ligation products were then identified and filtered using the HiC-Pro v2.8.1[19] pipeline, retaining only properly paired reads for subsequent analysis. Based on these valid reads, the contigs were clustered, ordered, and oriented into chromosomes using the Juicer v1.5[20] and 3D-DNA v3.0[21] software with parameters -m haploid -r 2 -c 30. The preliminary chromosomal scaffolds were visually inspected and manually corrected using Juicebox v1.11.08[22] to correct any misjoins and optimize chromosome structures. This approach successfully anchored and oriented the assembled contigs into 30 chromosome-scale scaffolds (Fig. 2), representing a complete chromosome-level assembly of the striped catfish genome.

To fill the remaining gaps, those corrected ultra-long ONT reads were applied to generate a gap-free genome assembly using TGS-GapCloser v1.2.1[23] with optimized parameter "–min_match 1000–min_nread 3" and LR_Gapcloser v1.0[24] with the parameter "-t 35 -m 1000000 -v 500". The final gap-free genome assembly was 772.03 Mb in length with a contig/scaffold N50 of 26.47 Mb (Table 2), and each chromosome is represented by a single contig (Fig. 2).

Identification of the centromere and telomere sequences. Telomeres were identified by scanning for the conserved (CCCTAA/TTAGGG > 100) repeat sequence at both ends of each chromosome using Telomere-to-Telomere Toolkit quarTeT v1.1.1[25]. Centromeres, as specialized DNA sequences connecting sister chromatids, exhibit complex structures in most animals and plants with highly repetitive satellite DNA and scattered retrotransposon sequences. In this study, after identifying repeat sequences according to TRF v4.0.4[26] (with default parameters) and RepeatMasker v4.0.6[27] (optimized parameter: nolow -no_is -gff -norna -engine abblast -lib) and obtaining a TE annotation file, quarTeT v1.1.1 (parameters: CentroMiner -n 100 -m 200 -g 50000 -l 100000–TE) was applied to identify centromeres, and the candidate interval range of every centromere was predicted. Our data revealed that the striped catfish genome contains a complete complement of 30 centromeres and 60 telomeres (Fig. 3), which are consistent with the haplotypic chromosome number (Fig. 2a and Fig. 4).

Annotation of repeat elements. We performed comprehensive annotation of repetitive elements (REs) using a combination of computational approaches. Simple sequence repeats (SSRs) were identified using TRF v4.0.4[26], while genome-wide tandem repeats were detected with GMATA v2.2[28].

Category	This study	NCBI GCA_027382375.1
Sex	Male	Unknown
Genome survey (Mb)	711.23	/
Genome length (bp)	772,037,129	764,974,446
Longest scaffold (bp)	37,557,520	36,215,093
Number of scaffolds	30	58
Contig N50 (bp)	26,472,860	13,007,371
Scaffold N50 (bp)	26,472,860	26,128,932
GC content	39.0%	39.0%
BUSCO	99.45% (S:99.01%; D:0.44%)	99.10% (S:98.10%; D:1.0%)
Number of chromosomes	30	30
Chromosome length (bp)	772,037,129	764,406,574
Repetitive sequence	39.5%	/

Table 2. Comparison of the available genome assemblies for striped catfish. Abbreviations: S, single copy complete genes; D, duplicated complete genes.

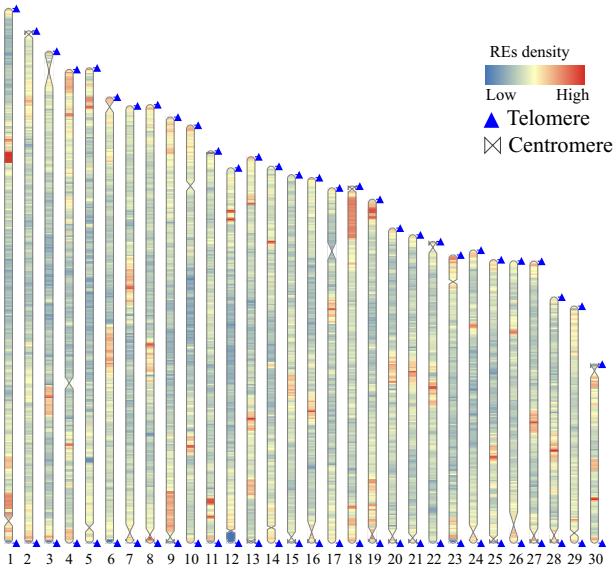


Fig. 3 Genome-wide localization of repetitive elements (REs), telomeres and centromeres. The triangles at both ends of each chromosome represent the telomere regions, and the gully area within each chromosome stands for the centromere region.

Transposable elements (TEs) were characterized through both homology-based and *de novo* prediction strategies. For homology-based annotation, we employed RepeatMasker v4.0.6 (with RepBase library) and RepeatProteinMask v4.0.6[27] to identify known TEs. For *de novo* prediction, we constructed a custom repeat library using RepeatModeler v1.0.8[29] and LTR_FINDER v1.0.6[30], followed by genome-wide annotation using RepeatMasker v4.0.6 against this library. The results from both approaches were integrated to generate a comprehensive RE dataset.

Repetitive elements constituted 39.53% of the assembled striped catfish genome (Table 3). DNA transposons represented the most abundant class (20.77%), followed by long terminal repeats (LTRs; 5.77%) and long interspersed nuclear elements (LINEs; 5.35%). In contrast, short interspersed nuclear elements (SINEs) were relatively scarce (1.35%). This distribution suggests differential evolutionary dynamics among TE classes in this species.

Prediction and functional annotation of protein-coding genes. Prior to gene prediction, repetitive regions in the assembled genome were soft-masked using RepeatMasker to improve annotation accuracy. Protein-coding genes were annotated through an integrated approach combining three complementary methods: For RNA-Seq annotation, the RNA-seq data from eight tissues were assembled into contigs using Trinity v2.5.1[31] with optimized parameters "–min_contig_length 200–min_kmer_cov 4–min_glue 4–bfly_opts '-V 5–edge-thr = 0.1–stderr'–genome_guided_max_intron 10000", and then gene structures were identified using PASA v2.3.3[32] and TransDecoder v5.5.0 (https://github.com/TransDecoder/TransDecoder). For *de novo* prediction, AUGUSTUS v3.2.1[33] and GlimmerHMM v3.0.4[34] were employed to predict *ab inito* gene structures. For

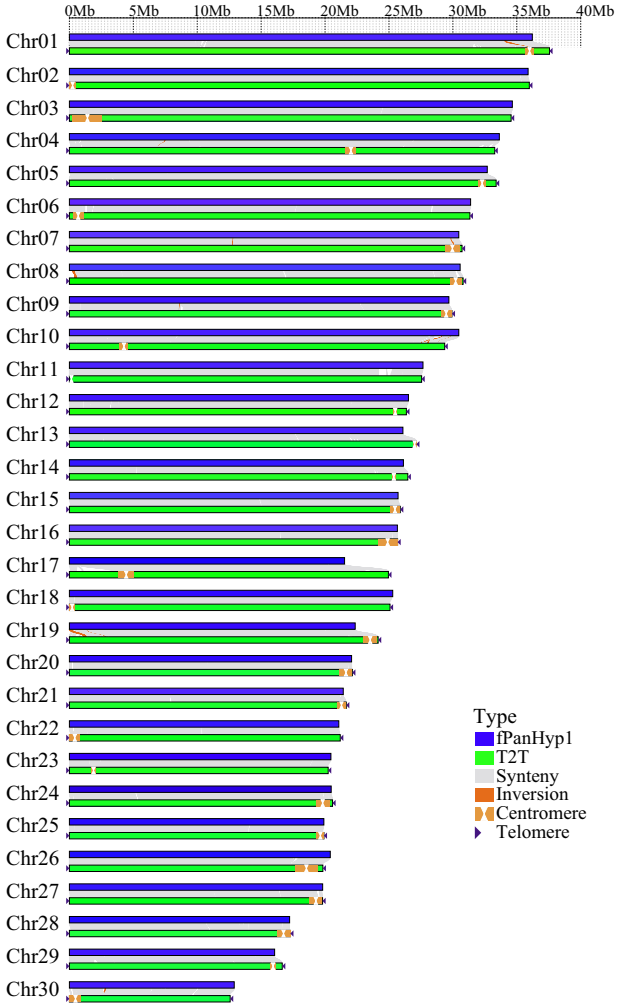


Fig. 4 Good synteny of chromosomes between our T2T assembly in this study and the previously published genome assembly (GCA_027382375.1) of the striped catfish.

Type			Length (bp)	Count	% of Genome
Dispersed repeats	DNA transposons		160,362,089	1,347,105	20.77
	Retroelements	LINE	41,364,834	253,601	5.35
		LTR	44,588,785	337,748	5.77
		SINE	10,428,269	112,650	1.35
Tandem Repeats	Simple repeats		4,929,305	27,150	0.64
	Satellites		17,863,568	280,226	2.31
Unknown			25,674,555	212,989	3.31
Total			305,211,405	2,571,469	39.53

Table 3. Classification of repetitive sequences in the assembled striped catfish genome.

homology-based gene prediction, we employed GeMoMa v1.6.4[35] using protein sequences from five phylogenetically representative fish species: *Ictalurus punctatus* (channel catfish; GCA_001660625.3), *Tachysurus fulvidraco* (yellow catfish; GCA_022655615.1), *Ictalurus furcatus* (blue catfish; GCA_023375685.2), *Silurus meridionalis* (southern catfish; GCA_014805685.1), and *Clarias gariepinus* (North African catfish; GCA_024256425.2). These reference proteomes were obtained from the NCBI database. The predictions from GeMoMa were combined with other evidence sources (RNA-Seq and *de novo* predictions) using the Evidence Modeler (EVM) v1.0.3[36] pipeline with weighted evidence integration. This comprehensive approach yielded 24,596 high-confidence protein-coding genes (Table 4). The annotated genes exhibited a mean gene length of 15.36 kb and an average coding sequence length of 1,624.48 bp, consistent with typical gene architectures in teleost fishes.

Functional annotation of the protein-coding genes was performed using Blastp v2.2.26[37], which aligned deduced protein sequences against five public databases including NCBI Non-Redundant Protein Sequence

Method	Software/Species	Number	Average length (bp)				Average exon per gene
			gene	CDS	exon	intron	
De novo	Augustus	24,450	17,338.33	1688.48	163.4	1,676.77	10.33
	Glimmer	50617	13,988.32	829.4	151.98	2,952.13	5.46
Homolog	*I. punctatus*	50,816	28,881.59	1,655.72	182.46	3,371.94	9.07
	T. fulvidraco	57,551	30,715.09	1,588.99	183.83	3,810.33	8.64
	I. furcatus	42,560	26,194.41	1,497.17	168.06	2,976.95	9.28
	S. meridionalis	48,112	27,083.98	1,676.64	166.19	3,194.82	9.01
	C. gariepinus	50,705	25,592.55	1,564.96	183.56	2,940.16	9.13
RNA-seq	PASA	20,853	20,105.79	3,494.58	298.93	1,553.86	11.69
Integrated	EVM	24596	15,365.21	1,624.48	175.58	1,665.54	9.25

Table 4. Summary of the predicted gene structures after integration of three annotation methods.

Database	Number	Percentage (%)
Total	24,596	100.00
NR	23,176	94.23
Swissprot	20,704	84.18
KEGG	16,459	66.92
GO	15,716	63.90
KOG	15,188	61.75
Ensembl	23,031	93.63
Overall	24,406	99.22

Table 5. Functional annotation of predicted protein-coding genes. Overall represents the total number of annotated genes with at least one hit from the five searched databases.

(NR), SwissProt[38], Gene Ontology (GO)[39], Kyoto Encyclopedia of Genes and Genomes (KEGG)[40], EuKaryotic Orthologous Groups (KOG)[41] and Ensembl (bony fish genome data)[42], with an E-value cutoff of $<1e-5$. Ultimately, 24,406 protein-coding genes (99.22% of the total predicted genes) were functionally annotated, with at least one hit for each gene in the searched databases (Table 5). These predicted protein-coding genes were evaluated using BUSCO with the actinopterygii_odb10 database as the reference. Finally, more than 96.21% (3,502) of complete BUSCOs were identified (Table 6).

Data Records

All sequencing data generated in this study have been deposited in the NCBI SRA database under BioProject accession PRJNA1281275[43], including MGI short-read data (SRA: SRR34366062), Hi-C data (SRA: SRR34366061 and SRR34366054), Nanopore data (SRA: SRR34366049-SRR34366051), PacBio HiFi data (SRA: SRR34366052 and SRR34366053) and RNA short-read data (SRR34366047, SRR34366048 and SRR34366055-SRR34366060)[44]. The final genome assembly and predicted coding sequences files were stored in Figshare (No: m9.figshare.29493197)[45]. The complete genome assembly has also been submitted at the NCBI/GenBank under the accession number of GCA_051362305.1[46].

Technical Validation

To evaluate the quality of our final genome assembly, we employed the following five approaches. First, BUSCO v5.2.2[47] was employed to examine completeness. A total of 99.45% (single copy complete genes (S): 99.01%, duplicated complete genes (D): 0.44%) of complete BUSCOs in the actinopterygii_odb10 database were identified. Second, Clipping information for Revealing Assembly Quality (CRAQ v1.09)[48] was used to assess the accuracy of our genome assembly based on PacBio HiFi and Illumina reads, resulting in a R-AQI of 94.77% and a S-AQI of 98.01%. Third, Merqury v1.328[49] was applied to estimate the base-level accuracy and completeness on the basis of k-mer counts (generated from Illumina, PacBio HiFi and ONT reads), resulting in a QV of 46.17, 55.66 and 57.48, respectively. Fourth, we mapped the sequencing data to the assembled genome using bwa v0.7.17[50] and minimap2 v2.26[51], which showed mapping rates of 99.55% for the MGI data, 99.99% for the PacBio data and 97.97% for the ONT data. Fifth, the newly assembled T2T genome was compared with the previously published assembly (GCA_027358585.1) using MUMmer v4.0.0beta2[52] with nucmer model (-g 1000 -c 90 -l 40 and delta-filter -l 1000). We used GenomeSyn v1.2.7[53] to visualize collinearity, which revealed one chromosome to one chromosome clearly with high homology (Fig. 4), suggesting good accuracy of chromosome anchoring. These results collectively support high quality of our striped catfish genome assembly in this study (Table 6).

The BUSCO completeness value was calculated to be 96.21% for the predicted protein-coding genes of striped catfish. To further evaluate the quality of these predicted protein-coding genes, we aligned the transcriptome data to the assembled genome using STAR v 2.7.11b[54], and then calculated the exonic coverage rate with bedtools v2.29.2[55]. We observed that 95.11% of the exonic regions had been covered with the RNA-seq reads, indicating high annotation accuracy (Table 6).

Type	Evaluation Methods		Results
Genome accuracy and completeness	Mapping short reads rate		99.55%
	Mapping HiFi reads rate		99.99%
	Mapping ONT reads rate		97.97%
	QV	HiFi reads	55.66
		Short reads	46.17
		ONT reads	57.48
	CRAQ	R-AQI	94.77%
		S-AQI	98.01%
		Coverage rate	99.62%
	BUSCO		99.45%
Annotation quality	Complete BUSCOs		96.21% (3,502)
	Complete and single-copy BUSCOs (S)		95.58% (3,479)
	Complete and duplicated BUSCOs (D)		0.63% (23)
	Fragmented BUSCOs (F)		0.74% (27)
	Missing BUSCOs (M)		3.05% (111)
	RNA-seq coverage ratio of the exonic regions		95.11%

Table 6. Assessment metrics of the genome assembly and annotation.

Data availability

In this study, sequencing data for P. hypophthalmus, including MGI, PacBio, Nanopore, Hi-C, and transcriptome sequencing, have been deposited in the NCBI BioProject database under the accession number PRJNA128127543. The raw sequencing reads are available in the NCBI Sequence Read Archive (SRA) under the accession numbers SRP59824844. The final genome assembly and predicted coding sequences files were stored in Figshare (No: m9.figshare.29493197)45. The complete genome assembly has also been submitted at the NCBI/GenBank under the accession number of GCA_051362305.146.

Code availability

The versions and parameters of bioinformatics tools applied in this study have been described in the Method section. If no parameter is provided, the default is set. No custom code was used.

Received: 23 July 2025; Accepted: 22 October 2025;
Published online: 28 November 2025

References

1. Taslima, K. *et al.* Genomic architecture and sex chromosome systems of commercially important fish species in Asia–Current status, knowledge gaps and future prospects. *Reviews in Aquaculture* **16**(4), 1918–1946 (2024).
2. Anka, I. Z. *et al.* Environmental issues of emerging pangas (*Pangasianodon hypophthalmus*) farming in Bangladesh. *Progressive Agriculture* **24**(1-2), 159–70 (2013).
3. Singh, A. K. *et al.* Culture of *Pangasianodon hypophthalmus* into India: Impacts and present scenario. *Pakistan Journal of Biological Sciences* **15**(1), 19–26 (2012).
4. Budd, A. M. *et al.* Sex control in fish: approaches, challenges and opportunities for aquaculture. *Journal of Marine Science and Engineering* **3**(2), 329–355 (2015).
5. Zhang, X. *et al.* Identification and characterization of a male-specific region in largemouth bass (*Micropterus Salmoides*) by whole-genome sequencing, resequencing and genomics comparison. *Frontiers in Marine Science* **12**, 1586534 (2025).
6. Xu, X. *et al.* A male-specific insert of *Opsariichthys bidens* identified based on genome-wide association analyses and comparative genomics. *Aquaculture Reports* **35**, 101982 (2024).
7. Liu, H. *et al.* Genetic manipulation of sex ratio for the large-scale breeding of YY super-male and XY all-male yellow catfish (*Pelteobagrus fulvidraco* (Richardson)). *Marine biotechnology* **15**(3), 321–328 (2013).
8. 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. *Molecular ecology resources* **22**(6), 2411–2428 (2022).
9. Gao, Z. *et al.* A chromosome-level genome assembly of the striped catfish (*Pangasianodon hypophthalmus*). *Genomics* **113**(5), 3349–3356 (2021).
10. Mei, L. *et al.* Evaluation of QIAamp® DNA Stool Mini Kit for ecological studies of gut microbiota. *Journal of Microbiological Methods* **54**(1), 13–20 (2003).
11. Chen S. *et al.* fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**(17), i884–i890.
12. Rhoads, A. *et al.* PacBio Sequencing and Its Applications. *Genomics Proteomics & Bioinformatics* **13**, 278–289 (2015).
13. De Coster, W. *et al.* NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**(15), 2666–2669 (2018).
14. Marçais, G. *et al.* A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**(6), 764–770 (2011).
15. Vurture, G. W. *et al.* GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**(14), 2202–2204 (2017).
16. Cheng, H. *et al.* Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170–175 (2021).
17. Roach, M. J. *et al.* Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics* **19**, 1–10 (2018).
18. Langmead, B. *et al.* Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**(4), 357–359 (2012).
19. Dekker, J. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology* **16**, 259 (2015).

20. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems* **3**(1), 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. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Systems* **3**, 99–101 (2016).
23. 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**(9), giaa094 (2020).
24. Xu, G. C. *et al*. LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. *Gigascience* **8**(1), giy157 (2019).
25. Lin, Y. *et al*. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Horticulture Research* **10**(8), uhad127 (2023).
26. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**(2), 573–580 (1999).
27. Tarailo-Graovac, M. *et al*. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* **Chapter 4**, 4–10 (2009).
28. Wang, X. & Wang, L. GMATA: an integrated software package for genome-scale SSR mining, marker development and viewing. *Frontiers in Plant Science* **7**, 1350 (2016).
29. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Science of the United States of America* **117**, 9451–9457 (2020).
30. Xu, Z. *et al*. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* **35**, W265–268 (2007).
31. 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).
32. Haas, B. J. *et al*. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31**(19), 5654–5666 (2003).
33. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **34**, W435–439 (2006).
34. Majoros, W. H. *et al*. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**(16), 2878–2879 (2004).
35. Keilwagen, J. *et al*. GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Methods in Molecular Biology* **1962**, 161–177 (2019).
36. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* **9**, R7 (2008).
37. Altschul, S. F. *et al*. Basic local alignment search tool. *Journal of Molecular Biology* **215**(3), 403–410 (1990).
38. Bairoch, A. *et al*. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research* **28**(1), 45–48 (2000).
39. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. *Nature genetics* **25**(1), 25–29 (2000).
40. Kanehisa, M. *et al*. KEGG as a reference resource for gene and protein annotation. *Nucleic acids research* **44**(D1), D457–D462 (2016).
41. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
42. Hubbard, T. *et al*. The Ensembl genome database project. *Nucleic acids research* **30**(1), 38–41 (2002).
43. *NCBI Bioproject*. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1281275 (2025).
44. *NCBI Sequence Read Archive*. https://identifiers.org/ncbi/insdc.sra:SRP598248 (2025).
45. Zhang, X. Genome assembly and predicted coding sequences files of *P. hypophthalmus*. *Figshare*. https://doi.org/10.6084/m9.figshare.29493197 (2025).
46. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_051362305.1 (2025).
47. Simao, F. A. *et al*. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
48. Li, K. *et al*. Identification of errors in draft genome assemblies at single-nucleotide resolution for quality assessment and improvement. *Nature Communications* **14**(1), 6556 (2023).
49. Rhie, A. *et al*. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21**(1), 245 (2020).
50. Li, H. *et al*. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**(14), 1754–1760 (2009).
51. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**(18), 3094–3100 (2018).
52. Marçais, G. *et al*. MUMmer4: A fast and versatile genome alignment system. *PLoS computational biology* **14**(1), e1005944 (2018).
53. Zhou, Z. W. *et al*. GenomeSyn: a bioinformatics tool for visualizing genome synteny and structural variations. *Journal of genetics and genomics* **49**(12), 1174–1176 (2022).
54. Dobin, A. *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**(1), 15–21 (2013).
55. Quinlan, A. R. *et al*. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**(6), 841–842 (2010).

Acknowledgements

This work was supported by Shenzhen Natural Science Foundation (no. JCYJ20241202124511016) and National Key Research and Development Program of China (no. 2022YFE0139700).

Author contributions

Q.S. and X.Z. conceived and designed the study. X.Z., Y.H., Y.X., X.W. and J.X. collected the samples. X.Z., Y.H. and J.C. performed data analysis. Y.H., Z.Z. and W.Z. conducted experiments for species identification. X.Z. and Y.H. wrote the manuscript. Q.S. revised the manuscript. All authors read and approved the final manuscript for publication.

Competing interests

The authors declare no competing interests.

Additional information

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