scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level assembly of spotted steed (*Hemibarbus maculatus* Bleeker, 1871) genome

Meng Zhang[1,2,3,5], Yue Xu[1,5], Xiao Ma[1,4], Zhilin Jia[1,2], Dongcai Chen[1], Huifen Liu[1], Hongxia Jiang[1], Lei Wang[1], Miao Yu[1], Yongjing Li[1], Zhigang Qiao[1,2] ✉ & Yawei Shen[1,2,3] ✉

Spotted steed (*Hemibarbus maculatus* Bleeker, 1871), a small and medium-sized benthic fish, is widely distributed between the Yangtze and Heilongjiang River basins and is considered to be one of the most widely distributed freshwater fish in East Asia. It is also an economically valuable aquaculture fish and has become a commercial freshwater aquaculture species in China. Here, a high-quality chromosome-level genome of spotted steed was produced by combining PacBio single molecule sequencing technique and high-throughput chromosome conformation capture technologies. Ultimately, the genome was assembled into 1098.66 Mb with a contig N50 of 33.57 Mb and a scaffold N50 of 40.40 Mb. We constructed a chromosome-level genome assembled with 25 chromosomes, whose total lengths accounted for 97.49%, and the assembled genome represents 95.64% completeness (BUSCO). We also identified 285.16 Mb (25.96%) of repetitive genome sequences and 23233 predicted genes. A total of 23021 genes were functionally annotated, representing 99.09% of the predicted genes. These results will provide valuable genomic resources for subsequent study of the genetic, evolutionary, and biological characteristics of spotted steed.

Background & Summary

Spotted steed (*Hemibarbus maculatus* Bleeker, 1871) belongs to the class Actinopterygii, the order Cypriniformes, and the family Cyprinidae[1], inhabits the middle and lower layers of water[2], and prefers benthic drilling holes. Spotted steed is an omnivorous fish and a small and medium-sized benthic fish widely distributed in the freshwater of China[1]. Spotted steed is widely accepted in the Chinese market because of its fresh meat[3], fewer intermuscular bones[4], fast growth rate, and high nutritional value[2], and has become a commercial freshwater aquaculture species in China[5]. Spotted steed is a precious fish in history, but its number began to decline seriously in the 1970s and 1980s due to environmental pollution, long-term heavy fishing pressure, and destruction of spawning grounds[6]. Despite this population decline, market demand for spotted steed remains high. Consequently, interest in spotted steed farming has increased.

Although the method of artificial propagation of spotted steed has been developed[7], there is no genomic resource of spotted steed so far, which seriously hinders the study of its genetics, development, reproduction, and evolution. A high-quality genome assembly is essential for understanding the genetic mechanisms underlying the adaptation of spotted steed to environmental stressors such as pollution and habitat degradation, and for enhancing its aquaculture production[8].

In the present study, a high-quality chromosome-level genome of spotted steed was generated by combining PacBio single-molecule real-time sequencing (SMRT) and high-throughput chromosome conformation capture (Hi-C) technologies (Fig. 1). Ultimately, the genome was assembled into 1098.66 Mb with the contig N50 of 33.57 Mb and scaffold N50 of 40.40 Mb. We constructed a chromosome-level genome with 25 chromosomes, and the assembled genome represents 95.64% completeness (BUSCO). We also identified 285.16 Mb (25.96%)

[1]College of Fisheries, Henan Normal University, Xinxiang, 453007, China. [2]Jiyuan Industrialized Recirculating Aquaculture Engineering Technology Research Center, Jiyuan Hongji Sturgeon and Giant Salamander Breeding Co., Ltd., Jiyuan, 459018, China. [3]Observation and Research Station on Water Ecosystem in Danjiangkou Reservoir of Henan Province, Nanyang, 474450, China. [4]College of Life Science, Nanjing Forestry University, Nanjing, 210037, China. [5]These authors contributed equally: Meng Zhang, Yue Xu. ✉e-mail: 13503800008@126.com; shenyawei1229@163.com; shenyawei@htu.edu.cn

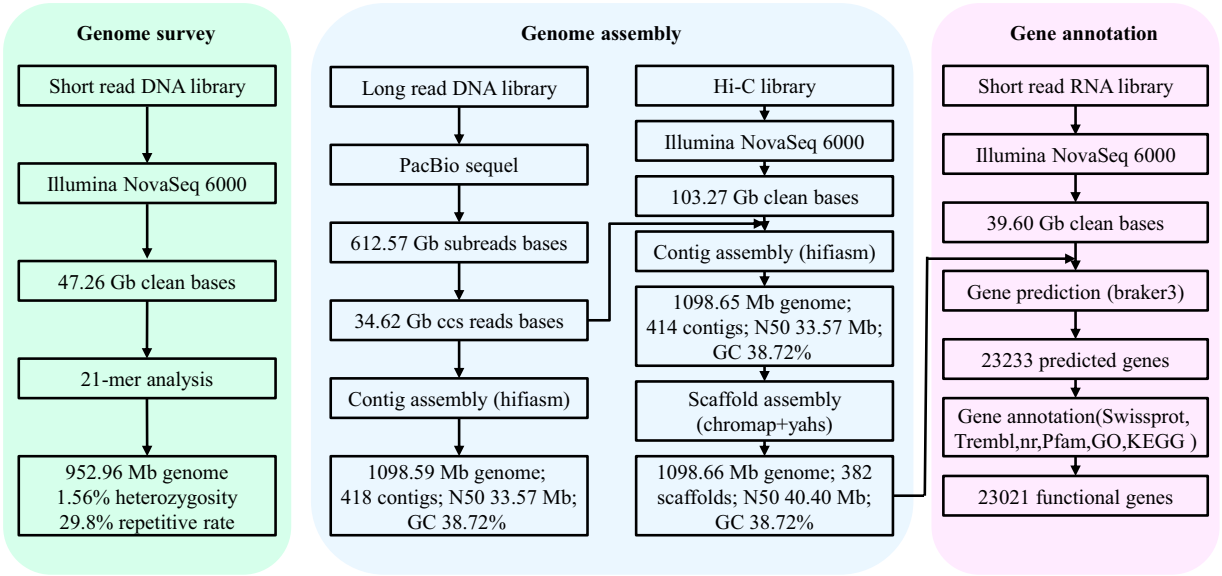

Fig. 1 The pipeline overview of the spotted steed chromosome-level genome assembly and annotation.

Library		Sequencing platform	Raw data				Clean data				Depth (×)	Genome size (Mb)	Contig N50(Mb)
			Total bases	Q20 (%)	Q30 (%)	GC (%)	Total bases	Q20 (%)	Q30 (%)	GC (%)			
RNA	Brain	Illumina NovaSeq 6000	7.42 Gb	96.20	91.26%	45.89%	6.29 Gb	97.05%	92.73%	45.83%	—	—	—
	Liver	Illumina NovaSeq 6000	7.59 Gb	96.61	91.97%	47.96%	6.46 Gb	97.34%	93.27%	47.92%	—	—	—
	Ovary	Illumina NovaSeq 6000	7.13 Gb	95.57	92.35%	48.15%	6.05 Gb	97.35%	93.62%	48.12%	—	—	—
	Mix tissues	Illumina NovaSeq 6000	24.07 Gb	97.59	92.80%	47.02%	20.80 Gb	97.73%	93.06%	47.06%	—	—	—
DNA	Short reads	Illumina NovaSeq 6000	47.56 Gb	97.77	92.68%	38.92%	47.26 Gb	97.78%	92.70%	38.88%	49.59	952.96	—
	Long reads	PacBio Sequel	614.44 Gb	—	—	—	612.57 Gb	—	—	38.72%	557.60	1098.59	33.57
	Hi-C	Illumina NovaSeq 6000	104.11 Gb	96.15%	88.19%	39.56%	103.27 Gb	96.17%	88.20%	39.46%	93.99	1098.66	40.40

Table 1. Sequencing statistics of spotted steed.

of repetitive genome sequences and a total of 23233 protein-coding genes, and 23021 genes were functionally annotated, representing 99.09% of the total predicted protein-coding genes.

This genomic data is the first chromosome-level genome of spotted steed[9], which will not only increase our understanding of its genetic structure, but also help us explore the evolution and unique biological characteristics[10]. This high-quality chromosome-level genome will provide valuable resources for subsequent studies while serving as the basis for investigating environmental adaptation mechanisms in spotted steed[11].

Methods

Sample collection and Sequencing. The muscle tissue of a two-year-old wild female spotted steed, collected from Suyahu Reservoir (Zhumadian, China), was used to construct the genomic DNA sequencing library. Meanwhile, its twelve tissues, including brain, gill, kidney, ovary, liver, spleen, muscle, heart, eye, blood, pectoral fin, and intestine, were used to construct the transcriptome sequencing libraries. All the tissues were stored in liquid nitrogen until use.

Genomic DNA was extracted from muscle tissue by the phenol chloroform method, and a short read DNA library was constructed according to the TruSeq™ DNA Sample Preparation Guide (Illumina, 15005180 Rev. A), including genomic DNA fragmentation, end repair, A tail and adaptor addition, target fragment selection, PCR amplification, and library quality control. The qualified short-read DNA library was sequenced on the Illumina NovaSeq 6000 (PE150. Illumina, San Diego, USA) and 47.56 Gb raw bases were obtained. After filtering using fastp v0.20.1 (parameters: --length_required 75)[12], 47.26 Gb clean bases (Table 1) were retained to perform the genome survey.

For long-read sequencing, high-quality genomic DNA from muscle was sheared using g-Tubes (Covaris, Woburn, MA, USA), and concentrated with AMPure PB magnetic beads. Each SMRT bell library was constructed using the Pacific Biosciences SMRT bell template prep kit 1.0[13]. The constructed library was size

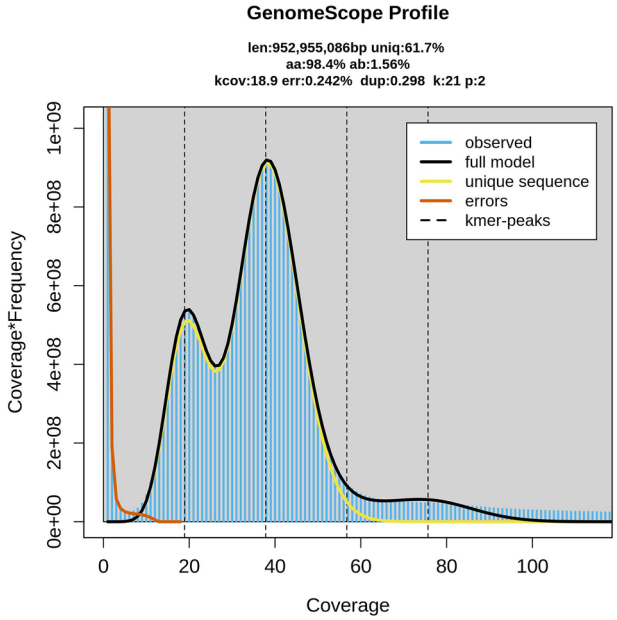

Fig. 2 The 21-mer distribution of spotted steed. The 21-mer distribution was calculated by GenomeScope2 based on 47.26 Gb short reads data. *K*-mer coverages (x axis) were plotted against the value of coverage multiplying frequency (y axis).

selected by Sage ELF for molecules 11~15 Kb, followed by primer annealing and the binding of SMRT bell templates to polymerases with the DNA Polymerase Binding Kit. Sequencing was carried out on the PacBio Sequel II platform for 30 h, yielding 614.44 Gb polymerase read bases and 612.57 Gb subread bases (Table 1).

In order to assemble the genome of spotted steed at the chromosome-level, the Hi-C libraries were prepared according to the standard procedure[14,15]. Briefly, the input Hi-C DNA was first fragmented, and the end of the fragmented DNA was repaired, phosphorylated at the 5' end, and the A tail at the 3' end. Subsequently, the adapter was connected to the end of the fragmented final product in the previous step, and streptavidin immunomagnetic beads were introduced to capture and retrieve the biotin-labeled site. Finally, the captured Hi-C DNA was subjected to PCR amplification to obtain Hi-C libraries. After the Hi-C library construction was completed, the library was quantified and insert fragments were detected. A high-quality Hi-C library was sequenced on the Illumina NovaSeq 6000 sequencing platform (Illumina, San Diego, USA), generating a total of 104.11 Gb raw bases and 103.27 Gb clean bases (Table 1), which were used for chromosome-level assembly[16].

Total RNA of twelve tissues, including brain, gill, kidney, ovary, liver, spleen, muscle, heart, eye, blood, pectoral fin, and intestine, was extracted following the protocol of our previous study[17]. Then the mixed RNA formed by equal mixing of twelve tissues and the RNA of three single tissues (brain, liver, and ovary) were generated to produce cDNA using the SMARTer PCR cDNA Synthesis Kit and sequenced on the Illumina NovaSeq 6000 (PE150. Illumina, San Diego, USA). In total, the transcriptome sequencing yielded 46.21 Gb raw data. After strict quality control and filtering by fastp v0.20.1 (parameters:–trim_front1 10 |–trim_tail1 10 |–trim_front2 10 |–trim_tail2 10)[12], a total of 39.60 Gb clean data (Table 1) was obtained and used to assist in gene prediction.

Genome survey. We evaluated the genome size, heterozygosity, and repetitive rate of the spotted steed genome by *k*-mer frequencies calculated from short-read data. Jellyfish v2.2.10[18] was used to obtain a 21-mer count histogram, and then GenomeScope2 v2.0.1[19] was used to evaluate the genomic characteristics of the above 21-mer frequency results. A total of 47.26 Gb high-quality data was generated from the short-read DNA library, with 49.59× genome coverage, Q20 of 97.78%, and Q30 of 92.70%, respectively (Table 1). The genome size was estimated at 952.96 Mb, with 1.56% heterozygosity and 29.8% repetitive rate, respectively (Fig. 2).

Chromosome-level genome assembly. The 34.62 Gb ccs reads data, generated by 612.57 Gb subreads through circular consensus sequencing, was used for the first draft genome assembly using hifiasm v0.21.0[20] with default parameters. Using QUAST (v5.2.0)[21], we evaluated the first draft assembly. The results indicated a genome size of 1098.59 Mb, comprising 418 contigs with an N50 length of 33.57 Mb and a GC content of 38.72% (Table 1).

The second draft genome was assembled with 103.27 Gb clean Hi-C data and 34.62 Gb ccs reads data using hifiasm[20] with default parameters. QUAST evaluation of the second draft assembly showed a genome size of 1098.65 Mb, comprising 414 contigs with an N50 of 33.57 Mb and a GC content of 38.72%. Subsequently, the Hi-C data were mounted on the second draft genome using chromap v0.2.7[22] and yahs v1.2.2[23] to generate the final chromosome-level genome. The Hi-C data were first mapped to the second draft genome using chromap v0.2.7[22], and the format of the mapped data was converted using samtools v1.21[24]. Then the visualized temporary HIC file was generated using yahs v1.2.2[23] and juicertools v1.19.02[25]. After adjusting the above temporary

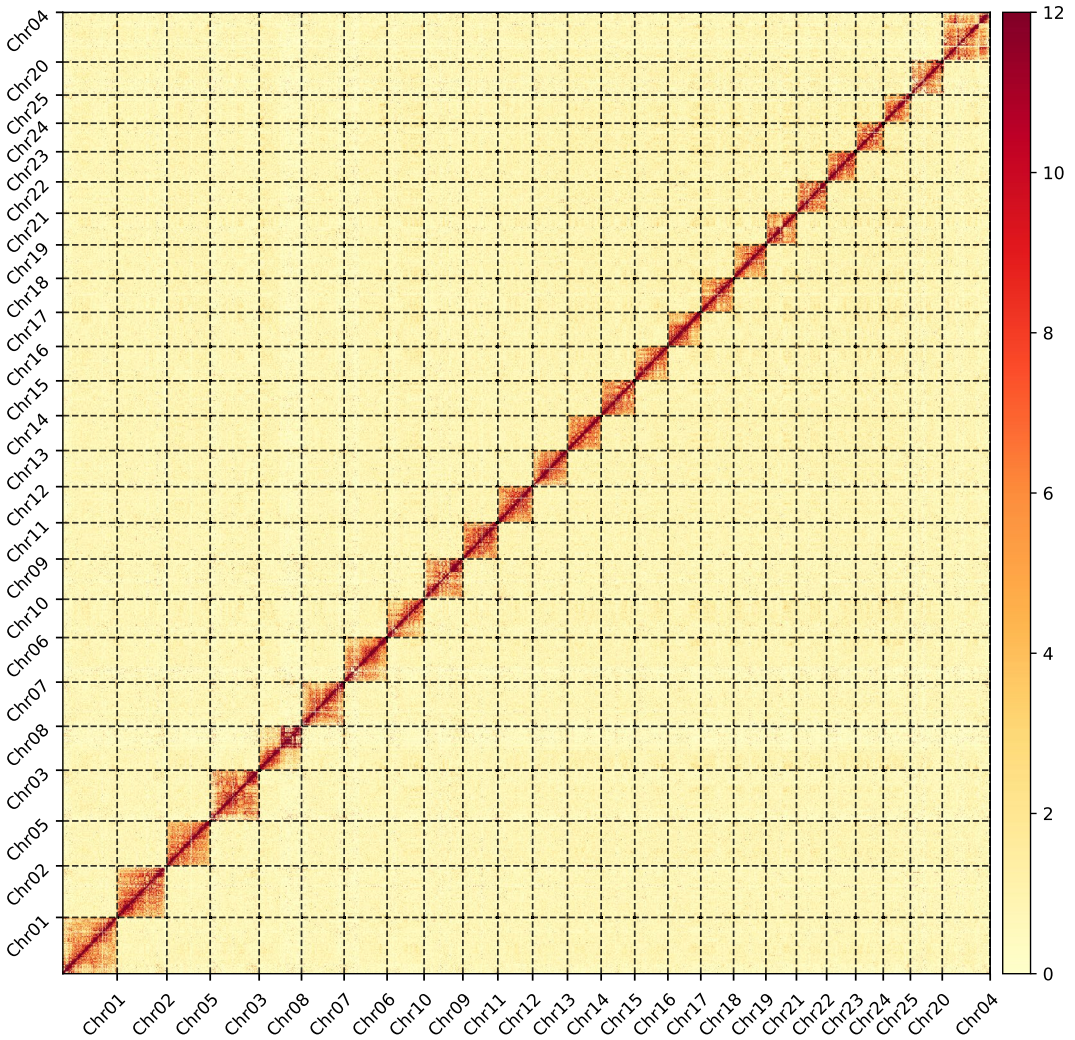

Fig. 3 Hi-C intra-chromosome contact map of the chromosome-level spotted steed genome assembly. Chr01 to Chr25 were the abbreviations of 25 chromosomes, which numbered in order of chromosome length from largest to smallest. The color block illuminated the intensity of interaction from yellow (low) to red (high).

HIC file in juicebox v1.11.08[26], the chromosome-level genome was finally generated. The plothic v1.0.0[27] was used to display the final chromosome-level genome, which contained 25 chromosomes (Fig. 3).

The QUAST v5.2.0[21] and BUSCO v5.8.3[28] were used to evaluate the integrity of the final chromosome-level spotted steed genome. The final chromosome-level genome assembly was assessed with QUAST. The assembly spans 1098.66 Mb, contains 382 scaffolds with an N50 of 40.40 Mb, and has a GC content of 38.72% (Table 1). The length of 25 chromosomes ranged from 31.38 Mb to 62.77 Mb, and the total length of 25 chromosomes accounted for 97.49% of the final chromosome-level genome. The BUSCO evaluation calculated the genome module benchmark value to be C: 95.64% (including S: 94.74% and D: 0.90%), F: 1.64%, M: 2.72%, and n = 7207 (C: complete BUSCOs, S: complete and single-copy BUSCOs, D: complete and duplicated BUSCOs, F: fragmented BUSCOs, M: missing BUSCOs, and n: total BUSCO groups of actinopterygii_odb12 data set) (Table 2).

Repeats prediction. Before annotating the repetitive sequences, the genome was pre-processed, including base substitution in uppercase and reordering by scaffold length. Then, using zebrafish (*Danio rerio*) as a reference species, the RepeatMasker v4.1.5[29] (parameters: -nolow -no_is -norna -species danio) was used to predict and mask repeat sequences in chromosome-level spotted steed genome. We identified the repetitive sequence of 285.16 Mb, accounting for 25.96% of the chromosome-level spotted steed genome, composed mainly of class II transposable elements (Table 3). Subsequently, the genomic file, masked by repetitive sequences, was used for subsequent gene prediction and annotation.

Gene prediction and annotation. The protein-coding genes in the chromosome-level spotted steed genome were predicted by using the working pipeline of BRAKER3 v3.0.8[30] combined with the evidence of transcript mapping and homologous gene alignment. Briefly, the evidence based on the transcript was the mapping result of the 39.60 Gb clean short read RNA data and the genomic file masked by repetitive sequences using hisat2 v2.2.1[31]. For homologous gene alignment, the protein database consists of seven fish species, including

BUSCO genome completeness score	Chromosome-level genome		23233 predicted genes	
	Data	Ratio	Data	Ratio
Complete BUSCOs (C)	6893	95.64%	6277	87.10%
Complete and single-copy BUSCOs (S)	6828	94.74%	6182	85.78%
Complete and duplicated BUSCOs (D)	65	0.90%	95	1.32%
Fragmented BUSCOs (F)	118	1.64%	120	1.66%
Missing BUSCOs (M)	196	2.72%	810	11.24%
Total BUSCO groups searched in Actinopterygii_odb12	7207	100%	7207	100%

Table 2. BUSCO evaluated the integrity of the chromosome-level spotted steed genome and predicted genes.

Type		Number	Length (bp)	Rate (%)
Class I: Retroelements	SINEs	22253	4060686	0.37%
	LINEs: L2/CR1/Rex	45898	17333366	1.58%
	LINEs: R1/LOA/Jockey	2820	1296193	0.12%
	LINEs: R2/R4/NeSL	530	325973	0.03%
	LINEs: RTE/Bov-B	1463	666078	0.06%
	LINEs: L1/CIN4	11050	2154927	0.20%
	LINEs: others	12146	7458830	0.68%
	LTRs: BEL/Pao	4112	2816096	0.26%
	LTRs: Ty1/Copia	66	116707	0.01%
	LTRs: Gypsy/DIRS1	57373	30531605	2.78%
	LTRs: Retroviral	34096	6851824	0.62%
	LTRs: others	21939	4285892	0.39%
Class II: DNA transposons	hobo-Activator	273473	32726370	2.98%
	Tc1-IS630-Pogo	95627	45810516	4.17%
	MULE-MuDR	10626	1065093	0.10%
	PiggyBac	14226	2307844	0.21%
	Tourist/Harbinger	74545	12996195	1.18%
	Other (Mirage, P-element, Transib)	10438	1748914	0.16%
	Unknown	723997	88609002	8.07%
Rolling-circles		32145	7969426	0.73%
Satellites:		63347	11138186	1.01%
Unclassified:		14470	2894171	0.26%
Total		1526640	285163894	25.96%

Table 3. The repetitive sequence statistics of the chromosome-level spotted steed genome. Note: Type: the type of repetitive sequence; Number: the number of repetitive sequences; Length: the total length of predicted repetitive sequences; Rate (%): the proportion of repetitive sequences in the total genome.

Pseudorasbora parva (NCBI Accession: GCF_024679245.1), *Danio rerio* (Ensembl: GRCz11, GCA_000002035.4), *Cyprinus carpio* (Ensembl: Cypcar_WagV4.0, GCA_905221575.1), *Carassius auratus* (Ensembl: ASM336829v1, GCA_003368295.1), *Oryzias latipes* (Ensembl: ASM223471v1, GCA_002234715.1), *Ictalurus punctatus* (Ensembl: ASM400665v3, GCA_004006655.3), and *Takifugu rubripes* (Ensembl: fTakRub1.2, GCA_901000725.2). After the above prediction, the initial 25850 predicted genes were obtained and filtered, including removing genes with ORFs less than 100 (1398), filtering incomplete genes (221), and redundant genes (998). Finally, a total of 23233 protein-coding genes were predicted, with a total gene length of 372353550 bp, exon length of 36092805 bp, and intron length of 336260745 bp, respectively (Fig. 4, Table 4).

Firstly, the protein sequences encoded by the 23233 predicted genes were compared with three protein databases, including Swiss-Prot (https://www.uniprot.org/uniprotkb?query=reviewed:true), TrEMBL (uniprot_trembl_human.dat.gz, uniprot_trembl_mammals.dat.gz, uniprot_trembl_rodents.dat.gz, and uniprot_trembl_vertebrates.dat.gz, https://ftp.ebi.ac.uk/pub/databases/uniprot/current_release/knowledgebase/taxonomic_divisions/), and NR protein databases (https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/), using diamond v2.1.11[32]. And their protein family annotation (Pfam) was performed by InterProScan v5.59-91.0[33]. Subsequently, Gene Ontology (GO) annotation was performed on the best alignment results of the predicted protein sequences with the above four protein databases, and Kyoto Encyclopedia of Genes and Genomes (KEGG) annotation was performed by the online website (https://www.kegg.jp/kegg/). Finally, a total of 23,021 genes, representing 99.10% of all predicted genes, were successfully functionally annotated in at least one protein database (Fig. 5)[34].

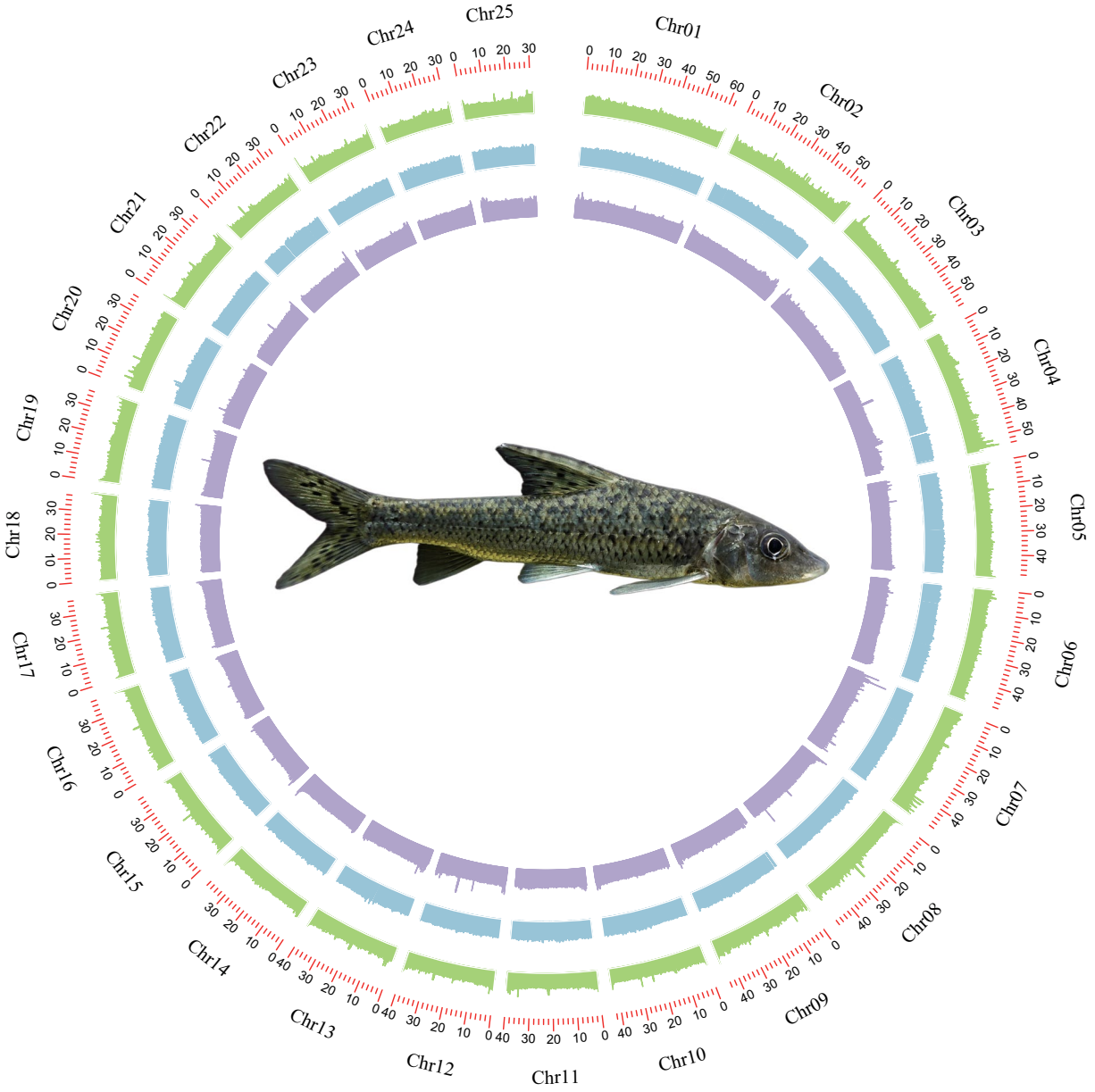

Fig. 4 Circos plot of the chromosome-level spotted steed genome. The tracks from outside to inside were 25 chromosomes, GC content, gene density, and repeat density, respectively.

Data Record

The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive (Genomics, Proteomics & Bioinformatics 2021) in National Genomics Data Center[35,36] (Nucleic Acids Res 2022), China National Center for Bioinformation (CNCB) / Beijing Institute of Genomics, Chinese Academy of Sciences (GSA: CRA028615[37], CRA028617[38], and CRA028635[39]). The whole genome sequence data reported in this paper have been deposited in the Genome Warehouse in National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation, under accession number (GWH: GWHGGFU00000000.1[40]). In addition, the final assembly of the genome sequence, gene structure annotation and functional annotation, and repeat sequence annotation files, have been uploaded and shared publicly in the figshare database[41].

Technical Validation

The degradation and contamination of DNA and RNA were detected by 1% agarose gel. The purity was checked by using the Nano Photometer Spectrophotometer. The absorbance of DNA and RNA was 1.82 and 1.80~2.01 at 260/280, respectively. DNA concentration was checked by using Qubit 2.0 Flurometer (Life Technologies, CA, USA), and RNA concentration was measured using Qubit RNA Assay Kit in Qubit 2.0 Flurometer (Life

Chr	Length/bp	Total gene length/bp	Gene number	Total exon length/bp	Exon number	Total intron length/bp	Intron number
Chr01	62769734	20244103	1279	1962327	10613	18281776	9334
Chr02	57504200	17655951	1147	1706517	9396	15949434	8249
Chr03	56569232	16750318	1338	1988352	10518	14761966	9180
Chr04	55238984	16930638	1192	1825569	10220	15105069	9028
Chr05	49968698	20327196	1239	1935417	11573	18391779	10334
Chr06	49663467	16244892	904	1451169	8386	14793723	7482
Chr07	49247539	15159700	845	1379838	7354	13779862	6509
Chr08	48908492	11633154	932	1402086	7072	10231068	6140
Chr09	44808216	11258342	976	1259898	7175	9998444	6199
Chr10	42687005	16552757	810	1361175	7820	15191582	7010
Chr11	40402732	13020331	783	1238547	6269	11781784	5486
Chr12	40230286	15878086	805	1384257	7567	14493829	6762
Chr13	40149866	14679995	891	1318407	7663	13361588	6772
Chr14	38911018	15352264	919	1522584	8975	13829680	8056
Chr15	38743376	13662125	892	1335489	7367	12326636	6475
Chr16	38328704	15437802	921	1383402	8253	14054400	7332
Chr17	37959718	13457156	727	1221936	6627	12235220	5900
Chr18	37903642	15077707	979	1533039	8472	13544668	7493
Chr19	37362408	14171970	862	1386015	7563	12785955	6701
Chr20	36836062	12001734	806	1101024	6348	10900710	5542
Chr21	35238100	12586261	850	1414371	7887	11171890	7037
Chr22	34877536	13376833	751	1225800	7222	12151033	6471
Chr23	33614000	15629395	754	1254294	7255	14375101	6501
Chr24	31785300	14039980	759	1260582	7135	12779398	6376
Chr25	31377721	11081505	841	1203906	6918	9877599	6077
Others	27574985	143355	31	36804	144	106551	113
Sum	1098661021	372353550	23233	36092805	201792	336260745	178559

Table 4. Characteristics of 23233 predicted protein-coding genes in spotted steed genome.

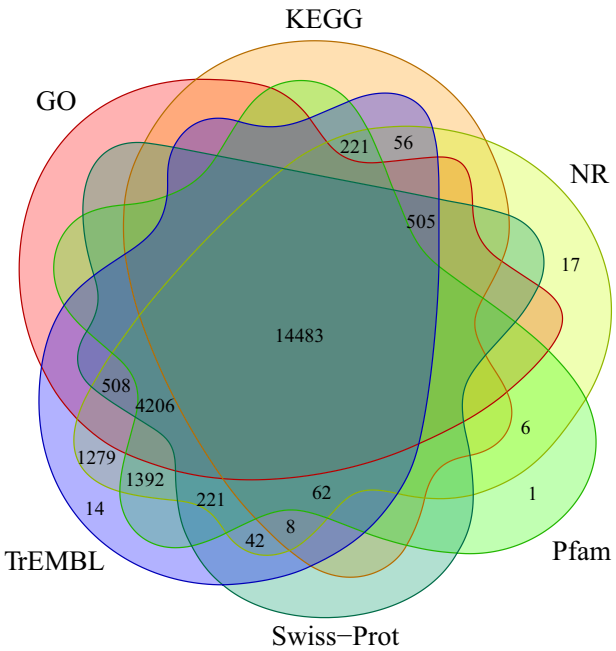

Fig. 5 Venn diagram of the number of genes with functional classification by each database.

Technologies, CA, USA). The integrity of the genome assembly was evaluated using BUSCO v5.8.3. Finally, 95.64% of the complete BUSCOs were included in the assembled genome.

Data availability

The raw sequence data of PacBio data, Hi-C data, Illumina short reads data, and genome assembly data reported in this paper have been deposited in the National Genomics Data Center, China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences under BioProject accession number PRJCA044140 (GSA Accession No.: CRA028615[37], CRA028617[38], and CRA028635[39]; GWH Accession No.: GWHGGFU00000000.1[40]). The final assembly of the genome sequence, gene structure annotation and functional annotation, and repeat sequence annotation files, have been uploaded and shared publicly in the figshare database[41].

Code availability

All commands and pipelines used in data processing were executed according to the manual and protocols of the corresponding bioinformatics software.

Received: 11 August 2025; Accepted: 19 March 2026;
Published online: 27 March 2026

References

1. Liu, H. F. *et al*. Development of SNP in spotted steed (*Hemibarbus maculatus* Bleeker) using GBS sequencing. *Conservation Genetics Resources* **12**(3), 379–383, https://doi.org/10.1007/s12686-019-01123-w (2020).
2. Wang, H. *et al*. Modeling oxygen consumption of the spotted steed larvae, *Hemibarbus maculatus* (Bleeker, 1871). *Aquaculture* **519**, 734914, https://doi.org/10.1016/j.aquaculture.2019.734914 (2020).
3. Wang, L. *et al*. Identification of genes and signaling pathways associated with immune of *Hemibarbus maculatus* (Bleeker, 1871) to ammonia stress. *Aquaculture* **524**, 735265, https://doi.org/10.1016/j.aquaculture.2020.735265 (2020).
4. Gu, R. B., Wen, H. B. & Xu, G. C. Evaluation of nutritive quality and nutritional components in *Hemibarbus maculates* Bleeker muscle. *Journal of Dalian Fisheries University* **21**(4), 378–382, https://doi.org/10.16535/j.cnki.dlhyxb.2006.04.017 (2006).
5. Liu, H. F. *et al*. Effects of recombinant AMH during oocyte maturation in spotted steed *Hemibarbus maculatus*. *Aquaculture* **543**, 736961, https://doi.org/10.1016/j.aquaculture.2021.736961 (2021).
6. Sun, L., Liu, J. D. & Yang, Y. J. Primarily study on fish community structure of *Hemibarbus maculatus* in Tonglu area of Qiantang river. *Journal of Biology* **31**(6), 46–50, https://doi.org/10.3969/j.issn.2095-1736.2014.06.046 (2014).
7. Liu, H. F. *et al*. The insulin-like growth factor 1 stimulates ovarian steroidogenesis and oocyte maturation in spotted steed *Hemibarbus maculatus*. *The Journal of Steroid Biochemistry and Molecular Biology* **224**, 106159, https://doi.org/10.1016/j.jsbmb.2022.106159 (2022).
8. Zhang, D. C. *et al*. Chromosome-level genome assembly of golden pompano (*Trachinotus ovatus*) in the family Carangidae. *Scientific data* **6**(1), 216, https://doi.org/10.1038/s41597-019-0238-8 (2019).
9. Zhang, Y. *et al*. Transcriptome sequencing and molecular markers discovery in the gonads of *Portunus sanguinolentus*. *Scientific data* **5**(1), 180131, https://doi.org/10.1038/sdata.2018.131 (2018).
10. Xu, D. D. *et al*. Chromosome-scale assembly and high-density genetic map of the yellow drum, *Nibea albiflora*. *Scientific Data* **8**(1), 268, https://doi.org/10.1038/s41597-021-01045-z (2021).
11. He, K. *et al*. Chromosome-level genome assembly of largemouth bass (*Micropterus salmoides*) using PacBio and Hi-C technologies. *Scientific Data* **9**(1), 482, https://doi.org/10.1038/s41597-022-01601-1 (2022).
12. Chen, S. F., Zhou, Y. Q., Chen, Y. R. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**(17), i884–i890, https://doi.org/10.1093/bioinformatics/bty560 (2018).
13. Korlach, J. *et al*. Real-time DNA sequencing from single polymerase molecules. *Methods in Enzymology* **472**, 431–455, https://doi.org/10.1016/S0076-6879(10)72001-2 (2010).
14. Rao, S. S. P. *et al*. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**(7), 1665–1680, https://doi.org/10.1016/j.cell.2014.11.021 (2014).
15. Gong, G. R. *et al*. Chromosomal-level assembly of yellow catfish genome using third-generation DNA sequencing and Hi-C analysis. *Gigascience* **7**(11), 1–9, https://doi.org/10.1093/gigascience/giy120 (2018).
16. Burton, J. N. *et al*. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. *Nature Biotechnology* **31**(12), 1119–1125, https://doi.org/10.1038/nbt.2727 (2013).
17. Shen, Y. W. *et al*. Short time-series expression transcriptome data reveal the gene expression patterns and potential biomarkers of blood infection with LPS and poly (I:C) in Mandarin fish (*Siniperca chuatsi*). *Fish & Shellfish Immunology* **153**, 109806, https://doi.org/10.1016/j.fsi.2024.109806 (2024).
18. Marcais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of *k*-mers. *Bioinformatics* **27**(6), 764–770, https://doi.org/10.1093/bioinformatics/btr011 (2011).
19. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature Communications* **11**(1), 1432, https://doi.org/10.1038/s41467-020-14998-3 (2020).
20. Cheng, H. Y., Concepcion, G. T., Feng, X. W., Zhang, H. W. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Method* **18**(2), 170–175, https://doi.org/10.1038/s41592-020-01056-5 (2021).
21. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**(8), 1072–1075, https://doi.org/10.1093/bioinformatics/btt086 (2013).
22. Zhang, H. W. *et al*. Fast alignment and preprocessing of chromatin profiles with Chromap. *Nature Communications* **12**(1), 6566, https://doi.org/10.1038/s41467-021-26865-w (2021).
23. Zhou, C. X., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **39**(1), btac808, https://doi.org/10.1093/bioinformatics/btac808 (2023).
24. Li, H. *et al*. & 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**(16)), 2078–2079, https://doi.org/10.1093/bioinformatics/btp352 (2009).
25. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems* **3**(1), 95–98, https://doi.org/10.1016/j.cels.2016.07.002 (2016).
26. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Systems* **3**(1), 99–101, https://doi.org/10.1016/j.cels.2015.07.012 (2016).
27. Jiang, Z. J. *et al*. A deep learning-based method enables the automatic and accurate assembly of chromosome-level genomes. *Nucleic Acids Research* **52**(19), e92, https://doi.org/10.1093/nar/gkae789 (2024).
28. Manni, M., Berkeley, M. R., Seppey, M. & Zdobnov, E. M. BUSCO: Assessing genomic data quality and beyond. *Current Protocols* **1**(12), e323, https://doi.org/10.1002/cpz1.323 (2021).
29. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current protocols in Bioinformatics* **25**, 4.10.1–4.10.14, https://doi.org/10.1002/0471250953.bi0410s25 (2009).

30. Gabriel, L. *et al*. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Research* **34**(5), 769–777, https://www.genome.org/cgi/doi/10.1101/gr.278090.123 (2024).
31. 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**(8), 907–915, https://doi.org/10.1038/s41587-019-0201-4 (2019).
32. Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods* **18**(4), 366–368, https://www.nature.com/articles/s41592-021-01101-x (2021).
33. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**(9), 1236–1240, https://doi.org/10.1093/bioinformatics/btu031 (2014).
34. Zhang, Y. P. *et al*. BioLadder: A bioinformatic platform primarily focused on proteomic data analysis. *iMeta* **3**(4), e215, https://doi.org/10.1002/imt2.215 (2024).
35. Chen, T. T. *et al*. The Genome Sequence Archive Family: Toward explosive data growth and diverse data types. *Genomics, Proteomics & Bioinformatics* **19**(4), 578–583, https://doi.org/10.1016/j.gpb.2021.08.001 (2021).
36. CNCB-NGDC Members and Partners. Database resources of the national genomics data center, China National Center for Bioinformation in 2025. *Nucleic Acids Research* **53**(D1), D30–D44, https://doi.org/10.1093/nar/gkae978 (2025).
37. *CNCB Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA028615 (2026).
38. *CNCB Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA028617 (2026).
39. *CNCB Genome Sequence Archive* https://ngdc.cncb.ac.cn/gsa/browse/CRA028635 (2026).
40. *CNCB Genome Warehouse* https://ngdc.cncb.ac.cn/gwh/Assembly/98818/show (2026).
41. Zhang, M. Genome annotation files of Hemibarbus maculatus. *figshare* https://doi.org/10.6084/m9.figshare.31361515 (2026).

Acknowledgements

This work was supported by the International Science and Technology Cooperation Project in Henan Province (252102520013); the Key Research Project of the Higher Education Institutions in Henan Province (24A240003); the Science and Technology Vice President of Henan Province (HNFZ20251518); and the Key R&D and Promotion Special Project of Henan Province (262102110075).

Author contributions

Meng Zhang, Zhigang Qiao and Yawei Shen managed the grants, supervised the laboratory work, and led the design of this study. Meng Zhang, Yue Xu and Xiao Ma performed bioinformatics and also drafted the manuscript. Zhilin Jia, Dongcai Chen and Huifen Liu participated in the tissue sampling. Meng Zhang, Yue, Xu, Hongxia Jiang, Lei Wang, Miao Yu, Yongjing Li, Zhigang Qiao and Yawei Shen revised the manuscript. 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 Z.Q. or Y.S.