scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of *Decorus tungting*, an endemic cyprinid from China

Xuan Xie[1,2,8], Li Zou[1,2,8], Fan Chen[3,8], Zaiquan Zhang[4], Linmei Han[1,2], Xing Xiang[5], Siqi Li[1,2,6], Zhonggui Xie[2,7] ✉ & Zhiqiang Liang[1,2] ✉

Decorus tungting **is a vulnerable, endemic cyprinid species of conservation significance, historically abundant in Dongting Lake but now suffering severe population declines due to overfishing and river modifications. However, no genome assembly has been reported for *D. tungting*, limiting insights into its genetic diversity, adaptive potential, and effective conservation management. Here, we assembled the first chromosome-level genome of *D.tungting* using PacBio HiFi long-read sequencing, Illumina short-read polishing, and Hi-C chromatin interaction mapping. The 1.14 Gb assembly has a contig N50 of 6.04 Mb and a scaffold N50 of 44.81 Mb, with 99.31% of sequences anchored to 25 chromosomes. Repetitive elements span 574.36 Mb, accounting for 50.28% of the genome. We annotated 24,835 protein-coding genes and 7,312 non-coding RNAs, with BUSCO completeness score of 97.1% in genome mode and 92.7% in protein mode, supporting assembly accuracy and annotation integrity. This high-quality genome assembly of *Decorus tungting* provides a foundational resource for genetic diversity, adaptive evolution, and evidence-based conservation.**

Background & Summary

The global decline in freshwater biodiversity has become a pressing concern. A recent assessment by the International Union for Conservation of Nature (IUCN) reported that 3,086 out of 14,898 assessed freshwater fish species are currently threatened with extinction (IUCN, 2023). This loss of biodiversity is driven by multiple factors, including global climate change[1,2], human-induced impacts such as river exploitation and hydroelectric development, and the spread of invasive species[3,4]. Invasive taxa have been associated with negative ecological consequences, such as predation on native small fish, habitat destruction, and substantial reductions in both biodiversity and biomass[5–9]. In China, findings from the third national assessment of the Red List of freshwater fishes indicate that 22.3% of species are facing extinction risks[10]. In response, various conservation strategies have been introduced, including the designation of freshwater fish conservation zones, implementation of fishing moratoriums, and execution of restocking programs[11,12]. Nonetheless, the overall efficacy of these measures remains to be systematically assessed.

Decorus tungting, also known as *Bangana tungting* (Nichols, 1925), classified as a vulnerable species by both the Red List and FishBase, exemplifies the conservation challenges associated with declining freshwater biodiversity. This large-bodied indigenous cyprinid, endemic to China, historically supported an annual catch exceeding 160 tons in the 1940s and 1950s[13]. However, field investigations conducted between 2006 and 2010 revealed a drastic population collapse throughout its native range, mainly due to overfishing and river system modifications[13]. Following this sharp decline, the species was classified as endangered.

The species is now largely restricted to the Yuan River National Aquatic Germplasm Resources Reserve (YUFRR), established in 2010 with strict no-fishing policies. Substantial progress has been made in artificial breeding and restocking efforts, with 8.65 million juveniles bred and released into native rivers between 2007

[1]Hunan Fisheries Research Institute and Aquatic Products Seed Stock Station, Hunan Academy of Agricultural Sciences, Changsha, 410153, China. [2]Yuelushan Laboratory, Changsha, 410128, China. [3]Department of Neurosurgery, Tangdu Hospital of Fourth Military Medical University, Xi'an, 710038, China. [4]Huaihua Tianling Nature Protection Company, Huaihua, 418200, China. [5]Animal Husbandry and Aquatic Products Center of Hongjiang District, Huaihua, 418200, China. [6]College of Fisheries, Hunan Agricultural University, Changsha, 410128, China. [7]Hunan Academy of Agricultural Sciences, Changsha, 410153, China. [8]These authors contributed equally: Xuan Xie, Li Zou, Fan Chen. ✉e-mail: 23692569@qq.com; liangzhiqiang@163.com

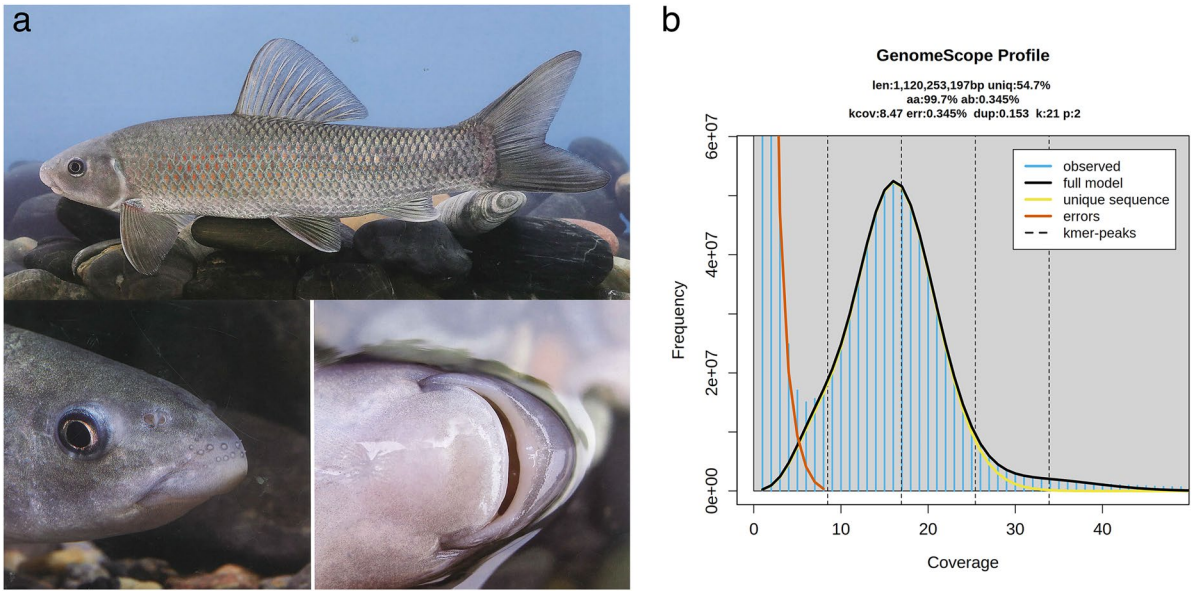

Fig. 1 Overview of the *Decorus tungting* genome assembly. (**a**) Representative morphological features of *D.tungting*, including its streamlined fusiform body, subterminal mouth with thick lips, and prominent sensory pores on the snout[50]. (**b**) 17-mer frequency distribution of Illumina short reads estimated using GenomeScope. The main peak corresponds to the homozygous peak, used to estimate genome size (~1.12 Gb), heterozygosity, and repeat content.

and 2021 through designated release sites[14]. A recent survey confirmed the effectiveness of protective measures in maintaining juvenile populations but also noted the rarity of natural recruitment, indicating that the wild population remains non-self-sustaining[15]. Given the precarious status of *D. tungting*, there is an urgent need to characterize its genetic diversity, adaptive capacity, and germplasm characteristics at the molecular level, so as to support conservation strategies such as genetically informed restocking, assessment of inbreeding and genetic load, and conservation genetic management[16]. A high-quality reference genome would provide an important foundation for conservation-oriented molecular analyses, as it would substantially facilitate downstream genetic assessments essential for implementing these strategies[17].

Genomics approaches have become increasingly integrated into conservation biology, providing powerful tools to assess demographic history, population structure, and adaptive potential. Advances in high-throughput sequencing technologies have facilitated the genome sequencing of various fish species. However, within the Cyprinidae family, genomic efforts have largely focused on model organisms or species of economic significance, leaving ecologically significant but genetically understudied species like *D. tungting* largely neglected.

The taxonomy of the genus *Bangana sensu lato* was recently revised by Zheng[18] and Song *et al.*[19] based on molecular phylogenetic and morphological evidence, leading to the establishment of the genus *Decorus*. This genus currently includes *D. decora*, *D. lemassoni*, *D. rendahli*, *D. tungting*, *D. xanthogenys*, and an undescribed species (D. sp.). Despite their ecological and taxonomic importance, no genome assemblies have been published for any *Decorus* species, leaving a critical gap in the molecular resources needed for evolutionary and conservation studies. The lack of genomic resources has constrained studies in *D. tungting*, making it challenging to evaluate genetic diversity, identifying adaptive signals and deleterious variants, or delineating evolutionarily significant units (ESUs), which limits the development of effective conservation strategies.

A high-quality reference genome can enable a suite of genomic applications under the framework of genomics-assisted conservation, including population genomic analysis, adaptive variation mapping, and marker-assisted conservation breeding. Such approaches have shown promise in improving conservation outcomes in endangered or commercial fishes such as *Oncorhynchus masou formosanu*[20], *Anguilla japonica*[21], and *Nannoperca australis*[22], and are increasingly adopted in conservation planning for non-model species[23].

In the present study, we generated a high-quality chromosome-level genome assembly of *Decorus tungting* using a combination of Illumina NovaSeq. 6000, PacBio long-read sequencing, and high-throughput chromosome conformation capture (Hi-C). Gene annotation was performed through a comprehensive approach integrating de novo prediction, homology-based methods, and RNA-Seq evidence. We also conducted comparative genomic analyses with closely related Cyprinidae species to investigate the evolutionary history of *D. tungting*. This work represents the first chromosome-level genome assembly for the genus Decorus, providing a fundamental genomic resource for studies on genetic diversity, adaptive evolution, and conservation of this vulnerable species.

Methods

Ethics statement. The methods involving animals in this study adhered to the Laboratory Animal Management Principles of China. All experimental protocols were approved by the Ethics Committee of Hunan Fisheries Science Institute.

Statistic	Value
Contig level	
Total length (bp)	1,142,381,163
N50 (bp)	6,036,726
N70 (bp)	3,711,826
N90 (bp)	1,411,969
L50 (count)	57
Longest contig (bp)	23,481,319
GC content (%)	36.43
Scaffold level	
Total length (bp)	1,142,691,163
N50 (bp)	44,812,555
N70 (bp)	41,472,674
N90 (bp)	37,518,612
L50 (count)	11
Longest scaffold (bp)	74,207,019
Total number of scaffolds	117
Number of gaps	620
Total gap length (bp)	310,000

Table 1. Statistics of the *Decorus tungting* genome assembly.

Sample collection. Living specimens of *Decrous tungting* were collected from Tianling Company Limited, Huaihua, China (Fig. 1a). Fish were placed under anesthesia using tricaine methanesulfonate (MS222), after which white muscle tissue was dissected, immediately immersed in liquid nitrogen, and stored at −80 °C until DNA extraction.

DNA Extraction and genome sequencing. Genomic DNA was extracted from muscle tissue using the proteinase-K–chloroform method. DNA quality and integrity were assessed by 1% agarose gel electrophoresis, and fragment size distribution was evaluated using a Bioanalyzer or PFGE to confirm the presence of high-molecular-weight DNA for long-read sequencing. High-quality DNA was used to construct libraries following the 20 kb PacBio template protocol with BluePippin size selection. DNA was sheared into ~20 kb fragments using a g-TUBE, then end-repaired and ligated to SMRTbell adapters. Libraries were purified, quality-checked, and mixed with primers and polymerase before sequencing on the PacBio Sequel II platform. Sequencing was conducted in CCS mode, generating high-fidelity HiFi reads for accurate genome assembly.

RNA Extraction and transcriptome sequencing. Total RNA was extracted from white muscle tissue using TRIzol. RNA integrity and contamination were assessed by 1% agarose gel electrophoresis, and concentration was measured using a Qubit fluorometer. Poly(A) + mRNAs were fragmented using fragmentation buffer and reverse-transcribed into cDNA with random hexamer primers. The cDNA was end-repaired, A-tailed, and ligated to sequencing adapters to prepare libraries with an average insert size of 350 bp. Libraries were sequenced in paired-end mode on the Illumina NovaSeq. 6000 platform, generating transcriptome data for genome annotation.

***De Novo* Genome assembly and quality assessment.** De novo genome assembly of *D.tungting* was performed using HiFi reads generated on the PacBio Sequel II platform. Genomic DNA was prepared with the SMRTbell Express Template Prep Kit 2.0 according to standard protocols. Sequencing was performed using Circular Consensus Sequencing (CCS v6.0.0) with a minimum of three passes, yielding 20.4 Gb of HiFi reads with an N50 of 13,539 bp, representing approximately 102 × sequencing coverage of the estimated genome size.

Prior to assembly, genome size, heterozygosity, and repeat content were estimated via 17-mer analysis using 9.5 Gb of quality-trimmed Illumina reads (TrimGalore v0.6.1[24], $Q < 20$). Jellyfish v2.3.0[25] and GenomeScope v1.0[26] were used to analyze k-mer distributions. HiFi reads were then assembled using hifiasm v0.16.1[27] with diploid optimization (-D 50 -N 500), yielding a final genome assembly size of 1.14 Gb, which is in good agreement with the k-mer-based estimate of ~1.12 Gb (Fig. 1b). Two rounds of polishing were performed with Pilon v1.23[28], using BWA v0.7.17-aligned Illumina[29] reads to correct indels and base errors. Assembly completeness was evaluated using BUSCO v5.2.2[30] with the Actinopterygii_odb10 dataset.

The final assembly resulted in a total genome size of 1.14 Gb across 117 scaffolds (Table 1). At the contig level, the assembly exhibited high continuity, with an N50 of 6.04 Mb, an N70 of 3.71 Mb, and an N90 of 1.41 Mb, while the longest contig reached 23.48 Mb. Following scaffolding, the scaffold-level assembly was further improved, achieving an N50 of 44.81 Mb, an N70 of 41.47 Mb, and an N90 of 37.52 Mb, with the longest scaffold measuring 74.21 Mb. The GC content of the assembled genome was 36.43%, and a total of 620 gaps corresponding to 0.31 Mb were detected. The 17-mer frequency analysis validated the high accuracy of the assembly (Fig. 1b).

Hi-C Library preparation and chromosome scaffolding. Muscle tissues were collected for Hi-C library construction. Genomic DNA was extracted and crosslinked with formaldehyde to stabilize chromatin interactions. Crosslinked DNA was digested using DpnII, end-labeled with biotinylated nucleotides, ligated to form

Chromosome	Length(bp)	GC content (%)	Number of Genes
Chr1	74,207,019	37.74	1702
Chr2	60,313,871	36.78	1304
Chr3	58,075,797	36.48	1358
Chr4	52,683,674	36.26	1176
Chr5	50,262,350	36.55	994
Chr6	48,817,928	35.69	932
Chr7	46,853,508	37.13	922
Chr8	46,836,827	36.59	1076
Chr9	46,519,164	36.31	1251
Chr10	45,297,846	36.18	1032
Chr11	44,812,555	36.28	867
Chr12	42,514,782	36.4	906
Chr13	42,273,426	36.68	867
Chr14	42,120,326	36.6	878
Chr15	41,896,208	36.18	952
Chr16	41,688,398	36.5	761
Chr17	41,472,674	36.28	920
Chr18	40,935,770	36.27	904
Chr19	40,161,905	36.26	791
Chr20	39,540,543	36.08	864
Chr21	38,854,554	36.13	892
Chr22	38,680,321	36.08	849
Chr23	37,518,612	36	943
Chr24	36,891,096	35.94	864
Chr25	35,547,276	35.76	753

Table 2. Statistics of chromosome assembly in *Decorus tungting*.

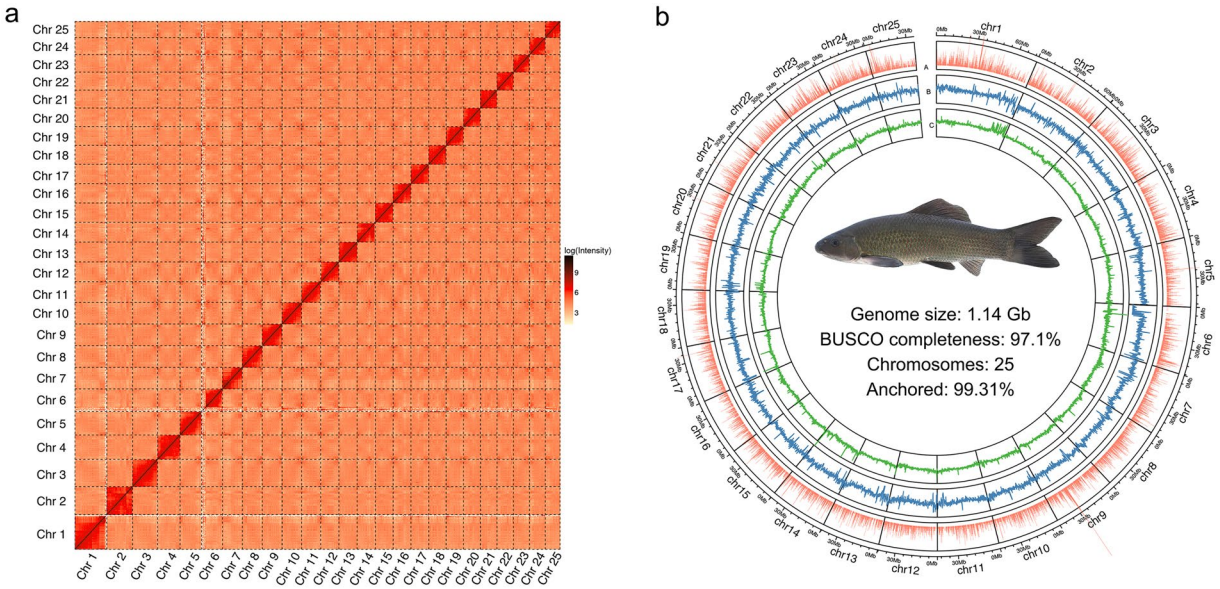

Fig. 2 The diagonal interaction signals indicate scaffold continuity and chromosomal integrity. (**a**) Hi-C contact heatmap showing chromosome-level assembly of *D. tungting*. (**b**) Circos plot of distribution of the genomic elements in *Decorus tungting*. For the outer to inner regions, each circle represents the gene density (A), repeative element density (B), GC content (C).

chimeric junctions, then purified and fragmented to ~350 bp. Hi-C libraries were prepared using the GrandOmics Hi-C kit and sequenced on the Illumina NovaSeq. 6000 platform in paired-end mode.

Raw Hi-C reads were processed using Juicer v1.6[31]. Contigs from the primary assembly were clustered, ordered, and oriented into chromosome-scale scaffolds using 3D-DNA v180922[32]. Manual curation was conducted with Juicebox v1.8.8[33] to refine scaffold structure. After Hi-C scaffolding, a total of 25 chromosomes were

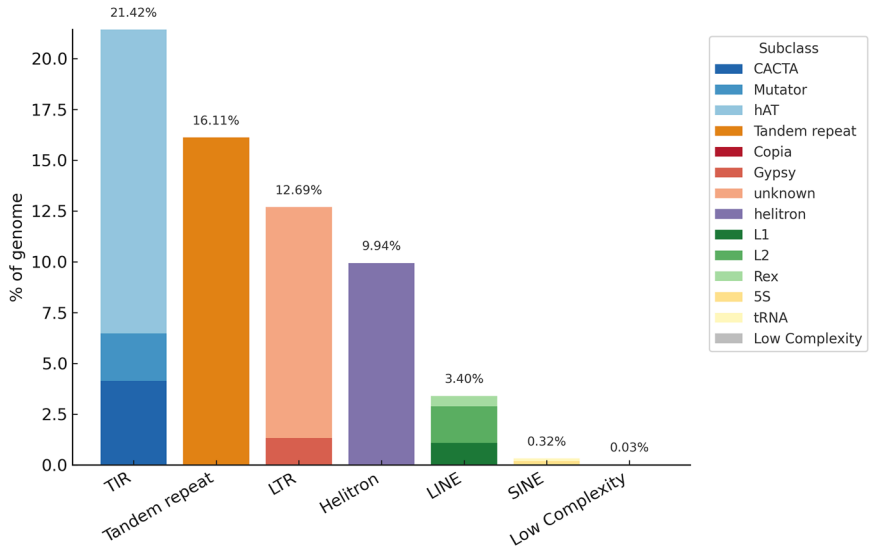

Fig. 3 Abundance and classification of repetitive elements in the *D. tungting* genome.

Class		Count	Size (bp)	% in genome
LTR	Gypsy	22,886	15,118,153	1.32%
	Copia	146	110,360	0.01%
	unknown	302,313	129,784,243	11.36%
TIR	hAT	436,429	170,821,733	14.95%
	CACTA	172,623	47,312,533	4.14%
	Mutator	80,003	26,655,006	2.33%
LINE	L2	36,284	20,625,783	1.81%
	L1	57,510	12,367,880	1.08%
	Rex	8,948	5,806,617	0.51%
SINE	5S	18,061	2,145,436	0.19%
	tRNA	6,949	1,536,941	0.13%
Helitron		347,312	113,549,279	9.94%
Tandem repeat		868,134	184,139,949	16.11%
Low Complexity		285	293,814	0.03%
Total Repeats		1,591,805	574,363,605	50.28%

Table 3. Prediction statistics of genome repeat sequence of *Decorus tungting*.

reconstructed, anchoring 96.42% of the genome assembly (Table 2), which is consistent with a previous cytogenetic study reporting the diploid karyotype of *D. tungting* as 2n = 50[34].The assembled chromosomes ranged in length from ~35.5 Mb to 74.2 Mb, containing between 753 and 1,702 predicted genes. Chromosomal GC content ranged from 35.76% to 37.74%. The high-resolution Hi-C contact map confirmed successful chromosomal scaffolding, with well-defined diagonal interaction patterns indicative of chromosome-scale continuity and integrity (Fig. 2a). These genome-wide features, including gene density, repeat content, and GC composition, were visualized using a Circos plot (Fig. 2b), providing an overview of structural organization.

Annotation of repeat elements. Repeat elements (REs) in the genome were annotated using the Extensive de-novo TE Annotator (EDTA, v1.9.9) pipeline[35,36],which integrates multiple de novo transposable element (TE) identification tools. Known repeats from the Repbase database, including cyprinid-specific entries, were also incorporated. The identified repeats were masked using RepeatMasker v4.1.1[37], and annotations were cross-validated with Repbase.

Repetitive elements comprised approximately 50.28% of the *D. tungting* genome, corresponding to ~574.36 Mb (Fig. 3). DNA transposons were the predominant repeat class. Among them, the hAT superfamily contributed the largest fraction, representing 14.95% of the genome. The CACTA and Mutator families accounted for 4.14% and 2.33%, respectively. Helitrons were also prevalent, occupying 9.94% of the genome. Tandem repeats, identified independently using TRF, accounted for 16.11% of the genome. LINEs and SINEs constituted 3.52% and 0.41%, respectively. Within LINEs, the L2, L1, and Rex subfamilies represented 1.81%, 1.08%, and 0.51% of the genome, respectively. SINEs were primarily derived from 5S rRNA and tRNA sequences, contributing 0.19% and 0.13%. A full summary of repeat composition, including LTRs, TIRs, Helitrons, LINEs,

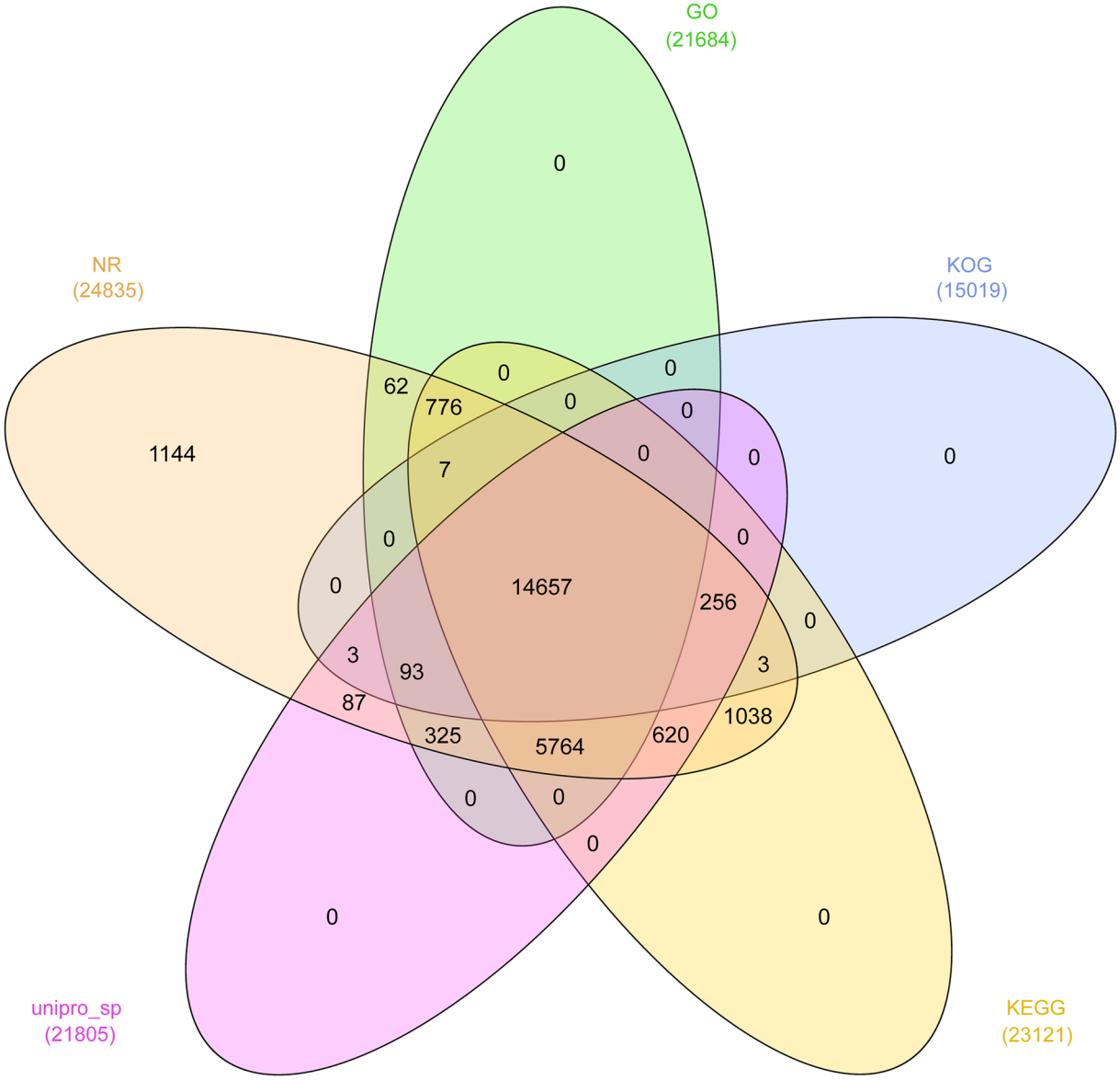

Fig. 4 Venn diagram for functional annotations of *Decorus tungting* protein-coding genes with the public database, including NR, KEGG, Interpro, Uniprot, and KOG.

SINEs, and low-complexity regions, is provided in Table 3, offering a comprehensive view of the transposable element landscape in the *D. tungting* genome.

Gene annotation and functional assignment. Quality control of raw RNA-Seq reads was performed using FastQC v0.11.9[38], followed by adapter trimming and filtering with Trimmomatic v0.39[39] ($Q < 20$, length ≥ 50 bp). Cleaned reads were used for transcript assembly to support gene model prediction. Gene prediction was carried out using the MAKER pipeline v3.01.04[40], which integrates transcriptomic evidence, homologous protein sequences, and ab initio gene models. HISAT2 v2.2.0[41] was used to align RNA-Seq reads to the genome to generate transcript alignments. Protein sequences from related species, e.g., *Ctenopharyngodon idella*, *Danio rerio*, *Gasterosteus aculeatus*, *Labeo rohita*, and *Megalobrama amblycephala*, were retrieved from public databases and included as homologous evidence. In the first round of MAKER, transcript and protein evidence were used with parameters est2genome = 1 and protein2genome = 1. Gene models with Annotation Edit Distance (AED) < 0.1 were selected to train ab initio predictors (AUGUSTUS[42] and SNAP[43]), which were used in the second round of MAKER to generate additional gene predictions. Genes potentially associated with hAT transposons were further predicted using GENSCAN[44]. Functional annotation of predicted protein-coding genes was conducted using BLASTP (e-value < 1e−5) against SwissProt, TrEMBL, and KEGG. InterProScan v5.52[45] was used to identify conserved domains and assign Gene Ontology (GO) terms.

A total of 24,835 protein-coding genes were predicted in the *D. tungting* genome (Table 4), comprising 251,015 exons and spanning approximately 73.59 Mb in total. The average number of exons per gene was 10.11,

Type	Number
Protein-coding genes (mRNA)	24,835
Coding sequences (CDS)	251,015
Exons	251,015
Single exon genes	2,129
Mean mRNAs per gene	1
Mean CDS per gene (bp)	1,751.58
Mean exons per mRNA	10.11
Total length of CDS/exons(bp)	43,500,516
Mean CDS length (bp)	49.317
Percentage of genome covered by genes	64.39%

Table 4. Summary of structural features of predicted protein-coding genes in *Decorus tungting*.

Functional database		Gene Number	Length 300~1000	Length ≥ 1000	% in genome
NR		24,835	13,439	3,397	100%
KEGG		23,121	13,112	3,347	93.10%
KOG		15,019	9,222	2,542	60.48%
Interpro	All	21,978	12,532	3,271	88.50%
	Pfam	21,978	12,532	3,271	88.50%
	GO	21,684	12,389	3,255	87.31%
uniprot_sp		21,805	12,510	3,328	87.80%
Annotated		24,835	13,439	3,397	100%

Table 5. Functional annotation statistics of predicted genes across major public databases in *Decorus tungting*.

Platform	Library insert size (bp)	Total reads (pairs for Illumina/Hi-C)	Total bases (Gb)	Mean read length (bp)
PacBio HiFi	–	1,486,672	20.43	13,741
Illumina PE150	350	64,525,708 pairs	19.36	150
Hi-C PE150	400	732,982,884 pairs	109.95	150

Table 6. Summary of raw sequencing data from three sequencing platforms used for the *Decorus tungting* genome project.

and 2,129 genes were identified as single-exon genes. Functional annotation was successfully assigned to all predicted genes based on at least one public database. Specifically, 93.10% of genes were linked to KEGG pathways, 88.50% matched InterPro entries, 87.31% had GO terms, and 87.80% showed similarity to UniProt records. Additionally, 60.48% of genes were classified into KOG categories. A Venn diagram summarizing annotation overlap across NR, KEGG, InterPro, UniProt, and KOG is shown in Fig. 4. Most genes were annotated in three or more databases, reflecting broad functional characterization. Complete annotation statistics are summarized in Tables 4, 5, detailed results for each annotation database are available in the public repository Figshare[46].

Data Records

The genome sequencing and assembly data for *Decorus tungting* have been deposited in the China National Center for Bioinformation (CNCB) under the BioProject accession PRJCA040001[47]. The PacBio HiFi sequencing reads are available under the BioSample accession SAMC5089076[48]. The assembled genome has been deposited in the Genome Warehouse (GWH) under accession number GWHGDIL00000000.1[49]. Functional annotation results, including KEGG, GO, Pfam, UniProt/Swiss-Prot, and KOG classifications, are available from the Figshare repository[46]. All datasets are publicly accessible through the NGDC portal and Figshare. These resources serve as foundational references for comparative genomics, evolutionary analysis, and molecular breeding studies of cyprinid fish.

Technical Validation

To ensure assembly quality, we evaluated the raw sequencing data from PacBio HiFi, Illumina PE150, and Hi-C PE150 before assembly. Illumina and Hi-C datasets had Q30 values above 91%, and sequencing depth reached ≥ 16 × for PacBio and Illumina and ~96 × for Hi-C, indicating high base quality and adequate coverage for reliable contig construction and chromosome scaffolding (Table 6). Assembly completeness was assessed using BUSCO v5.2.2 with the actinopterygii_odb10 dataset (n = 3,640 conserved single-copy orthologs). In genome mode, 97.1% of BUSCOs were complete, comprising 95.0% single-copy and 2.1% duplicated genes, while 1.2% were fragmented and 1.7% were missing. In protein mode, 92.7% of BUSCOs were complete (89.4%

BUSCO (genome mode)	Count	Proportion
Complete BUSCOs (C)	3534	97.1%
Complete and single-copy BUSCOs (S)	3457	95%
Complete and duplicated BUSCOs (D)	77	2.1%
Fragmented BUSCOs (F)	43	1.2%
Missing BUSCOs (M)	63	1.7%
BUSCO (protein mode)	**Count**	**Proportion**
Complete BUSCOs (C)	3374	92.7%
Complete and single-copy BUSCOs (S)	3253	89.4%
Complete and duplicated BUSCOs (D)	121	3.3%
Fragmented BUSCOs (F)	23	0.6%
Missing BUSCOs (M)	243	6.7%

Table 7. Genome quality assessment statistics of the *Decorus tungting* genome.

Parameters	value
Percentage of Assembly Anchored to Chromosomes	99.31%
Chromosome N50 (bp)	44812555
Chromosome N70 (bp)	41472674
Chromosome L50 (bp)	11
Chromosome L70 (bp)	17
Repeat Content Distribution	48.31%

Table 8. chromosome construction assessment statistics of *Decorus tungting*.

single-copy, 3.3% duplicated), and 6.7% were missing (Table 7). These results support the completeness of the genome and the accuracy of the annotated gene set.

Chromosome-scale scaffolding was achieved using Hi-C data, which anchored 99.31% of the genome to 25 chromosomes. The scaffold N50 reached 44.81 Mb, with the longest chromosome measuring 74.21 Mb. The L50 and L70 values were 11 and 17, respectively, indicating high contiguity and efficient chromosomal anchoring (Table 8). Additionally, repeat analysis revealed that 48.31% of the genome is composed of repetitive sequences, consistent with values reported in other cyprinid genomes. These combined metrics indicate a high-quality assembly suitable for downstream comparative and functional genomic analyses.

Data availability

The genome sequencing and assembly data for Decorus tungting have been deposited in the China National Center for Bioinformation (CNCB) under BioProject accession PRJCA04000148. The PacBio HiFi sequencing reads are available under BioSample accession SAMC508907649. The assembled genome is deposited in the Genome Warehouse (GWH) under accession GWHGDIL00000000.1. Functional annotation datasets are available at Figshare[46].

Code availability

All software and pipelines used for data processing were executed according to the manuals and protocols of the bioinformatics software cited above, and the parameters are clearly described in the Methods section. If no detailed parameters are mentioned for a software, the default parameters were used. The version of the software has been described in Methods. No custom code was used in this study for the curation or validation of the dataset.

Received: 20 June 2025; Accepted: 23 September 2025;
Published online: 04 November 2025

References

1. Aziz, M. S. *et al.* Decline in fish species diversity due to climatic and anthropogenic factors in Hakaluki Haor, an ecologically critical wetland in northeast Bangladesh. *Heliyon* **7** (2021).
2. Barbarossa, V. *et al.* Threats of global warming to the world's freshwater fishes. *Nat Commun* **12**, 1701 (2021).
3. Barbarossa, V. *et al.* Impacts of current and future large dams on the geographic range connectivity of freshwater fish worldwide. *Proceedings of the National Academy of Sciences* **117**, 3648–3655 (2020).
4. Su, G. *et al.* Human impacts on global freshwater fish biodiversity. *Science (1979)* **371**, 835–838 (2021).
5. Yalçın Özdilek, Ş., Partal, N. & Jones, R. I. An invasive species, Carassius gibelio, alters the native fish community through trophic niche competition. *Aquat Sci* **81**, 29 (2019).
6. Lusk, S., Lusková, V. & Hanel, L. Alien fish species in the Czech Republic and their impact on the native fish fauna. *Folia Zool Brno* **59**, 57–72 (2010).

7. Strayer, D. L. Alien species in fresh waters: ecological effects, interactions with other stressors, and prospects for the future. *Freshw Biol* **55**, 152–174 (2010).
8. Castaldelli, G. *et al*. Introduction of exotic fish species and decline of native species in the lower Po basin, north-eastern Italy. *Aquat Conserv* **23**, 405–417 (2013).
9. Wang, J., Chen, L., Tang, W., Heino, J. & Jiang, X. Effects of dam construction and fish invasion on the species, functional and phylogenetic diversity of fish assemblages in the Yellow River Basin. *J Environ Manage* **293**, 112863 (2021).
10. Cao, L., Shao, W., Yi, W. & Zhang, E. A review of conservation status of freshwater fish diversity in China. *J Fish Biol* **104**, 345–364 (2024).
11. Zheng, Y., Zhang, H., Niu, Z. & Gong, P. Protection efficacy of national wetland reserves in China. *Chinese Science Bulletin* **57**, 1116–1134 (2012).
12. Liang, Y. & Zhuang, S. Biodiversity in China: challenges, efforts and prospects. *China Economic J* **17**, 26–39 (2024).
13. Bian, W., Li, C. & Liang, Z. Biological characteristic and resource dynamic of Sinilabeo decorus tungting. *Journal of Hydroecology* **32**, 67–73 (2011).
14. Li, S. *et al*. Histological Study of the Development of the Digestive System of Sinilabeo decorus Tungting Larvae. *Journal of Fishery Sciences of China* 1033–1043 (2022).
15. Tian, L. *et al*. Investigation on the Distribution of Bangana Tungting in Yuanshui Unique Fish Species National Aquatic Germplasm Resources Reserve Using Environmental DNA Technology. *Ecol Evol* **14**, 1–9 (2024).
16. Arthington, A. H., Dulvy, N. K., Gladstone, W. & Winfield, I. J. Fish conservation in freshwater and marine realms: status, threats and management. *Aquat Conserv* **26**, 838–857 (2016).
17. Ali, I. Biotechnology in Environmental Conservation: Genetic Tools for Biodiversity Preservation. *Frontiers in Biotechnology and Genetics* **1**, 147–165 (2024).
18. Zheng, L. P., Chen, X. Y. & Yang, J. X. Molecular phylogeny and systematic revision of Bangana sensu lato (Teleostei, Cyprinidae). *Journal of Zoological Systematics and Evolutionary Research* **57**, 884–891 (2019).
19. Song, W. *et al*. E-cadherin maintains the undifferentiated state of mouse spermatogonial progenitor cells via β-catenin. *Cell Biosci* **12**, 141 (2022).
20. Christensen, K. A. *et al*. Masu salmon species complex relationships and sex chromosomes revealed from analyses of the masu salmon (Oncorhynchus masou masou) genome assembly. *G3: Genes, Genomes, Genetics* **15**, jkae278 (2025).
21. Wang, H. *et al*. A chromosome-level assembly of the Japanese eel genome, insights into gene duplication and chromosomal reorganization. *Gigascience* **11**, giac120 (2022).
22. Pavlova, A. *et al*. Planning and implementing genetic rescue of an endangered freshwater fish population in a regulated river, where low flow reduces breeding opportunities and may trigger inbreeding depression. *Evol Appl* **17**, e13679 (2024).
23. Bernos, T. A., Jeffries, K. M. & Mandrak, N. E. Linking genomics and fish conservation decision making: a review. *Rev Fish Biol Fish* **30**, 587–604 (2020).
24. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* **17**, 10–12 (2011).
25. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
26. 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).
27. 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).
28. Walker, B. J. *et al*. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
29. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics* **25**, 1754–1760 (2009).
30. 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).
31. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst* **3**, 95–98 (2016).
32. Dudchenko, O. *et al*. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science (1979)* **356**, 92–95 (2017).
33. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst* **3**, 99–101 (2016).
34. Zhang, J., Xiaofang, L. & Zuxiong, W. A comparative study on the karyotypes among the hybrid fish (sinilabeo decorus tungting × cirrhinus molitorella) and its parental fishes. *Acta hydrobiologica sinica* **8**, 313–321 (1984).
35. Su, W., Ou, S., Hufford, M. B. & Peterson, T. A tutorial of EDTA: extensive de novo TE annotator. *Plant Transposable Elements: Methods and Protocols* 55–67 (2021).
36. 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).
37. Chen, N. Using Repeat Masker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **5**, 4–10 (2004).
38. Brown, J., Pirrung, M. & McCue, L. A. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. *Bioinformatics* **33**, 3137–3139 (2017).
39. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
40. Cantarel, B. L. *et al*. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res* **18**, 188–196 (2008).
41. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907–915 (2019).
42. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34**, W435–W439 (2006).
43. Aylor, D. L., Price, E. W. & Carbone, I. SNAP: Combine and Map modules for multilocus population genetic analysis. *Bioinformatics* **22**, 1399–1401 (2006).
44. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**, 78–94 (1997).
45. Jones, P. *et al*. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
46. Xie, X. *et al*. *Gene Annotation and Functional Assignment of Decorus tungting*. https://doi.org/10.6084/m9.figshare.29875199.
47. *China National Center for Bioinformation*. https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA040001.
48. *China National Center for Bioinformation*. https://ngdc.cncb.ac.cn/biosample/browse/SAMC5089076.
49. *China National Center for Bioinformation*. https://ngdc.cncb.ac.cn/gwh/Assembly/95114/show.
50. Liao, F. *et al*. *Color Atlas of Fishes in Hunan*. (2020).

Acknowledgements

The study was financially supported by the Fund for Native fish breeding, the earmarked fund for HARS (HARS-07), Yuelushan Laboratory Talent Program (2024RC2014) and Aquatic Seed Industry Plan of Department of Agriculture and Rural Affairs of Hunan Province (HNAP2023).

Author contributions

X.Xie and L.Zou conceived and designed the experiments; L.Zou, Z.Zhang and X.Xiang conducted the experiments; X.Xie and F.Chen performed the genome assembly and repeat elements detection; X.Xie and Z. Liang conducted the gene homologues and genomic synteny analysis; L.Han and S.Li performed the gene annotation; Z.Xie and Z. Liang recruited animal resources; X.Xie, L.Zou and F.Chen wrote the paper; and all authors read, edited 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.X. or Z.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.