scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of the Leopard Wrasse *Macropharyngodon Meleagris*

Haiyan Yu[1,2,3,4,6], Meng Qu[2,3,4,5,6] ✉, Chao Li[2,3,4], Qiang Lin[2,3,4,5] & Dazhi Wang[1] ✉

Macropharyngodon meleagris **is a coral reef-dwelling, benthic predatory fish renowned for its striking coloration and distinctive body patterns. It exhibits pronounced sexual dimorphism and is characterized by the presence of prominent canine-like teeth. Here, we employed PacBio HiFi sequencing combined with Hi-C assembly technology to generate a high-quality, chromosome-level genome assembly of *M. meleagris*. The final assembly spans 666 Mb across 24 chromosomes, with high contiguity (a contig N50 of 20.57 Mb and a scaffold N50 of 29.79 Mb). Approximately 27.63% of the genome is composed of repetitive elements. A total of 21,940 protein-coding genes were predicted, with 98.76% successfully assigned functional annotations. The assembled genome exhibits high completeness (98.7% BUSCO completeness) and accuracy (98.05% for WGS short reads, 99.86% for HiFi long reads and 92.64% for RNAseq reads).**

Background & Summary

Wrasses (family Labridae) are increasingly recognized for advanced cognitive abilities, including problem-solving and social interaction. Cleaner wrasses, for instance, cooperate with client fish and track their reputations, preferring clients that offered beneficial interactions previously[1]. Other species, such as tuskfish (*Choerodon* spp.) and some *Halichoeres* wrasses, even use rocks as anvils to crack hard-shelled prey - a clear example of tool use[2].

The genus *Macropharyngodon*, a member of the family Labridae, currently comprises 12 recognized species[3]. The Indo-Pacific genus *Macropharyngodon* (family Labridae) comprises small reef fishes (up to 125 mm SL) characterized by a large molar-like tooth on the lower pharyngeal plate[4]. Renowned for its vivid coloration and intricate patterning[4], the leopard wrasse (*Macropharyngodon meleagris*) - also known as the blackspotted wrasse or Guineafowl wrasse - inhabits subtidal reef flats, outer lagoons, and seaward reefs, typically favoring areas with a mix of sand, rubble, and coral[5,6]. Sexual dimorphism is pronounced: females are pale with large black spots, while males display bright orange-red coloration with green spots on each scale, a black shoulder patch dotted with yellow, and green stripes across the head[4]. Using specialized dentition, it crushes and consumes durable prey including gastropods, hard-shelled invertebrates, and foraminiferans[7,8].

In this study, we present a high-quality, chromosome-level genome assembly of *M. meleagris*, using PacBio HiFi long-read sequencing and Hi-C chromatin conformation capture technology. The final assembly spans 666.65 Mb with a scaffold N50 of 29.79 Mb, and 24 pseudo-chromosomes were anchored. This chromosome-scale genome provides a valuable genomic resource for investigating the ecological adaptation and evolutionary history of Labridae species.

[1]College of the Environment and Ecology, Xiamen University, Xiamen, 361102, China. [2]State Key Laboratory of Tropical Oceanography, South China Sea Institute of Oceanology, Guangdong Provincial Key Laboratory of Applied Marine Biology, Chinese Academy of Sciences, Guangzhou, 510301, China. [3]University of Chinese Academy of Sciences, Beijing, 100101, China. [4]Marine Biodiversity and Ecological Evolution Research Center, South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou, 510301, China. [5]Sanya Institute of Ocean Eco-Environmental Engineering, Sanya, 572000, China. [6]These authors contributed equally: Haiyan Yu, Meng Qu. ✉e-mail: qumeng@scsio.ac.cn; dzwang@xmu.edu.cn

Library type	Insert Size (bp)	Clean Data (Gb)	Coverage (×)
Short reads	350	38.91	60.65
PacBio reads	20,000	31.78	49.54
Hi-C reads	350	77.15	120.26

Table 1. Statistics of sequencing data of the *M. meleagris* genome.

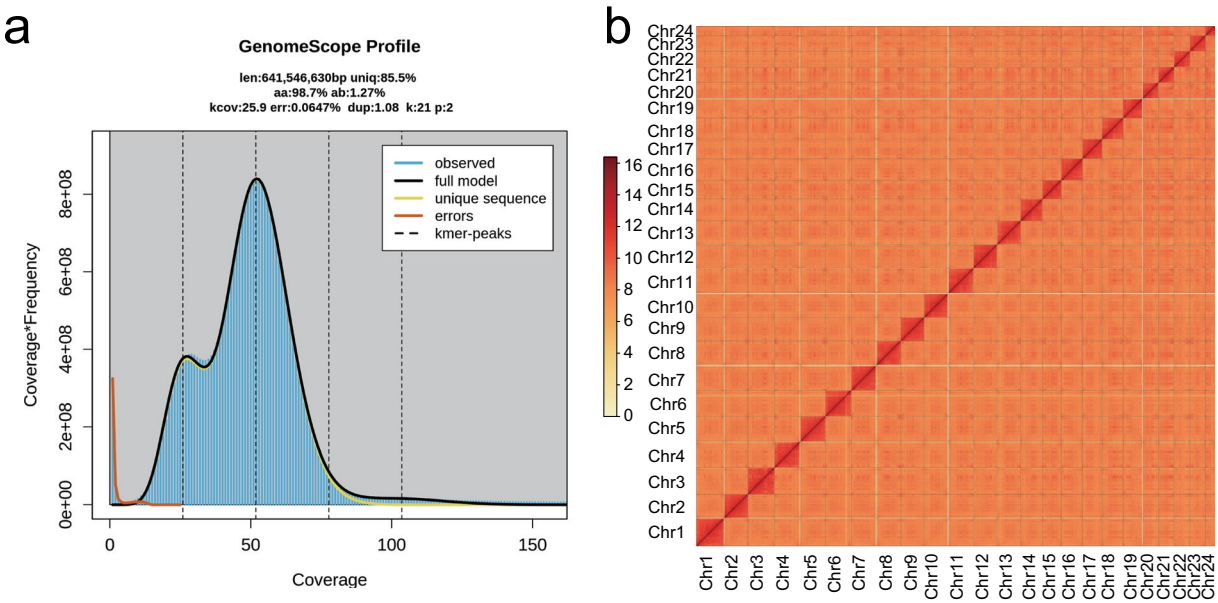

Fig. 1 Overview of *M. meleagris* genome assembly: (**a**) GenomeScope plot of the 21-mer frequency distribution in the *M. meleagris* genome. (**b**) The Hi-C interaction heatmap for the *M. meleagris*.

Methods

Sample collection, sequencing and genome size estimation. A juvenile male was collected and genomic DNA was isolated from muscle tissue using the standard phenol-chloroform extraction method. DNA quality and concentration were assessed via 1% agarose gel electrophoresis and quantified using the Pultton DNA/Protein Analyzer (Plextech). Sample's sex could not be definitively determined at the time of sampling. Given that this species exhibits protogynous hermaphroditism (female-to-male sex change), the individual could develop into either sex.

The genome was sequenced using three strategies: Illumina, PacBio HiFi, and Hi-C. For short-read DNA sequencing, paired-end library with an insert size of 350 bp were constructed and sequenced on an Illumina NovaSeq 6000 platform. FastQC v0.11.9 was used to check the quality of the sequencing data. The short-read data were used for k-mer analysis (for genome size and heterozygosity estimation) and for polishing the draft assembly. For long-read DNA sequencing, a Single Molecule Real-Time (SMRT) bell library with a fragment size of 20 kb was constructed following the manufacturer's protocol and sequenced on a PacBio Sequel II system to generate circular consensus sequences (CCS). For Hi-C sequencing, muscle tissue was fixed with 2% formaldehyde to cross-link DNA and proteins. The cross-linked DNA was digested with DpnII.

To obtain proximity-ligated DNA fragments, sticky ends were labeled with biotin and they were then circularized, sheared to a size of 300–700 bp, and enriched using biotin pull-down. The size-selected fragments were then used to construct Hi-C libraries and sequenced using the Illumina Novaseq 6000 sequencing platform.

For RNA-seq sequencing, total RNA was extracted from the brain, gill, and kidney tissues using TRIzol reagent (Invitrogen), and then Nanodrop 2000C and Qubit 3.0 were used to evaluate the purity and quantity. The total RNA was mixed and sequenced on the Illumina NovaSeq 6000 platform. The resulting data were used for gene prediction.

A total of 38.91 Gb (estimated sequencing coverage of approximately 60.65-fold) of Illumina short reads, 31.78 Gb (estimated sequencing coverage of approximately 49.54-fold) of HiFi long reads and 77.15 Gb (estimated sequencing coverage of approximately 120.26-fold) of Hi-C data were obtained (Table 1).

The genome size was estimated using K-mer method with a total of 38.91 Gb Illumina DNA short reads (60.65x of the estimated genome size). Jellyfish v2.1.4 was used to count the number of 21-mer sequences (jellyfish count -C -m 21 -s 5 G -t 10) and generate K-mer histogram (jellyfish histo -h 3000000 -o test_21mer.histo test_21mer)[9]. Genomescope was used to fit a Bayesian model to the k-mer frequency histogram for genome profiling and then estimate genome size, heterozygosity and repeat content. K-mer analysis estimated the genome size of *M. meleagris* to be 641.55 Mb, with a heterozygosity rate of 1.27% and a repeat content of 14.51% (Fig. 1a).

	Contig	Scaffold
Total Number	150	112
Total Length (bp)	666,633,134	666,652,134
N50 Length (bp)	20,571,900	29,785,295
N90 Length (bp)	6,954,216	19,848,471
Max Length (bp)	31,965,816	34,477,392
Min Length (bp)	13,810	13,810
Gap Number (bp)	0	19,000
GC Content (%)	41.34	41.34

Table 2. The statistics of the *de novo* assembled *M. meleagris* genome.

Organism Name	Accession Link
Cetoscarus ocellatus	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/037/902/135/GCA_037902135.1_ASM3790213v1/GCA_037902135.1_ASM3790213v1_genomic.fna.gz
Cheilinus chlorourus	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/050/084/645/GCA_050084645.1_ASM5008464v1/GCA_050084645.1_ASM5008464v1_genomic.fna.gz
Cheilinus undulatus	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/049/815/525/GCA_049815525.1_ASM4981552v1/GCA_049815525.1_ASM4981552v1_genomic.fna.gz
Choerodon schoenleinii	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/047/301/625/GCA_047301625.1_ASM4730162v1/GCA_047301625.1_ASM4730162v1_genomic.fna.gz
Halichoeres trimaculatus	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/033/239/565/GCA_033239565.1_ASM3323956v1/GCA_033239565.1_ASM3323956v1_genomic.fna.gz
Labroides dimidiatus	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/030/710/495/GCA_030710495.1_ASM3071049v1/GCA_030710495.1_ASM3071049v1_genomic.fna.gz
Labrus bergylta	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/963/930/695/GCF_963930695.1_fLabBer1.1/GCF_963930695.1_fLabBer1.1_genomic.fna.gz
Labrus mixtus	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/963/584/025/GCF_963584025.1_fLabMix1.1/GCF_963584025.1_fLabMix1.1_genomic.fna.gz
Notolabrus celidotus	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/009/762/535/GCF_009762535.1_fNotCel1.pri/GCF_009762535.1_fNotCel1.pri_genomic.fna.gz
Scarus forsteni	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/036/419/115/GCA_036419115.1_ASM3641911v1/GCA_036419115.1_ASM3641911v1_genomic.fna.gz
Semicossyphus pulcher	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/022/749/685/GCF_022749685.1_fSemPul1.0.p/GCF_022749685.1_fSemPul1.0.p_genomic.fna.gz
Symphodus melops	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/947/650/265/GCA_947650265.1_fSymMel2.1/GCA_947650265.1_fSymMel2.1_genomic.fna.gz
Symphodus ocellatus	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/965/170/635/GCA_965170635.1_Socellatus_v1/GCA_965170635.1_Socellatus_v1_genomic.fna.gz
Symphodus tinca	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/965/165/685/GCA_965165685.1_Stinca_v1/GCA_965165685.1_Stinca_v1_genomic.fna.gz
Tautogolabrus adspersus	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/020/745/685/GCA_020745685.1_fTauAds1.pri.cur/GCA_020745685.1_fTauAds1.pri.cur_genomic.fna.gz
Thalassoma bifasciatum	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/008/086/565/GCA_008086565.1_UOO_Thbifa_1.2/GCA_008086565.1_UOO_Thbifa_1.2_genomic.fna.gz
Thalassoma lucasanum	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/046/269/465/GCA_046269465.1_ASM4626946v1/GCA_046269465.1_ASM4626946v1_genomic.fna.gz
Xyrichtys novacula	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/962/446/985/GCA_962446985.1_fXyrNov1.1/GCA_962446985.1_fXyrNov1.1_genomic.fna.gz

Table 3. Data sources for the publicly available Labridae genomes used in this study.

Genome assembly and quality assessment. CCS v4.2.0 (https://github.com/PacificBiosciences/ccs) was used to generate Highly Accurate Single-Molecule Consensus Reads (HiFi Reads) from raw PacBio subreads, with a minimum of 3 passes (–min-passes 3) to ensure high base accuracy. The resulting HiFi reads were filtered for length (>5 kb) and quality (QV > 20) before being assembled *de novo* with Hifiasm v0.16.0 under default settings[10]. NextPolish v1.4.1 was employed for polishing the assembly with Illumina short-read sequencing data to improve base-level accuracy[11]. Purge_dups v1.2.5 was used to remove haplotigs and heterozygous contig overlaps from the initial genome based on read depth cutoffs calculated from aligned HiFi reads[12]. Hi-C read was used to further scaffold the genome assembly to the chromosome level. HiC-Pro v3.1.0 was used to process Hi-C data, transforming raw paired-end Illumina FASTQ files into normalized contact maps[13]. Juicer v1.5 was used for quality control and mapping of Hi-C reads[14]. 3d-dna v180922 was used to anchor contigs into chromosomes[15]. Finally, Juicebox v1.11.08 was used to visualize the Hi-C contact map and manually correct potential assembly errors[16].

HiFi long reads were *de novo* assembled into a draft genome of approximately 666.63 Mb, comprising 150 contigs with an N50 of 20.57 Mb and a maximum contig length of 31.97 Mb (Table 2). We consider the assembly to be reasonable and biologically accurate for two main reasons. First, it closely matches the k-mer–based estimate of 641.55 Mb obtained from GenomeScope analysis (Fig. 1a), differing by only 3.91%. Second, the assembly shows high completeness (BUSCO score: 98.7%). Using the Hi-C data, 62 contigs containing 649.72 Mb draft sequences (97.46% of the entire draft assembly) were anchored to 24 pseudochromosomes, with individual chromosome length ranging from 11.94 Mb to 34.48 Mb (Fig. 1b), consistent with the chromosome numbers previously reported in several closely related species[17–20]. The final genome assembly comprised 24 chromosomes and 88 contigs, totaling 666.65 Mb in length, with a contig N50 of 20.57 Mb and a scaffold N50 of 29.79 Mb (Table 2). The final chromosome-level genome assembly was assessed using Benchmarking Universal Single-Copy Ortholog BUSCO (v5.2.2) with the actinopterygii_odb10 database[21]. The BUSCO assessment revealed a completeness of 98.7% for the genome assembly. All available Labridae fish genomes (*Labrus bergylta* (GCF_963930695.1), *Symphodus ocellatus* (GCA_965170635.1), *Tautogolabrus adspersus* (GCA_020745685.1), *Notolabrus celidotus* (GCF_009762535.1), *Symphodus melops* (GCA_947650265.1), *Choerodon schoenleinii* (GCA_047301625.1), *Labrus mixtus* (GCF_963584025.1), *Cheilinus chlorourus*

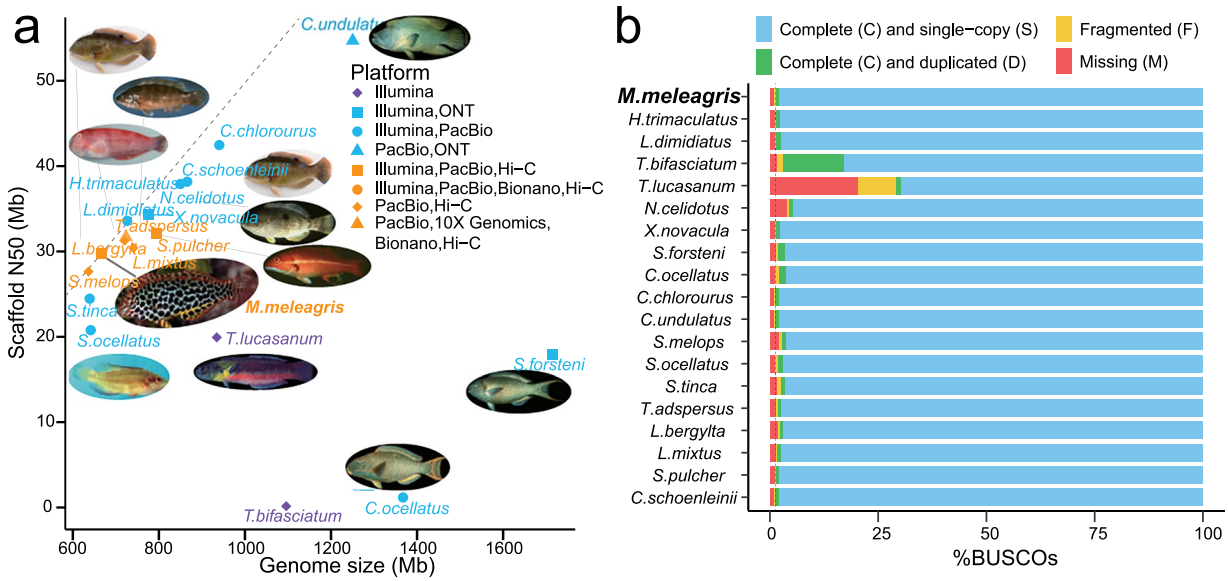

Fig. 2 Comparison of genome information between *M. meleagris* and other Labridae species. (**a**) Scatter plot of genome sizes versus scaffold N50 of 19 Labridae fishes. (**b**) BUSCO (Benchmarking set of Universal Single-Copy Orthologues) result for 19 Labridae fish genomes.

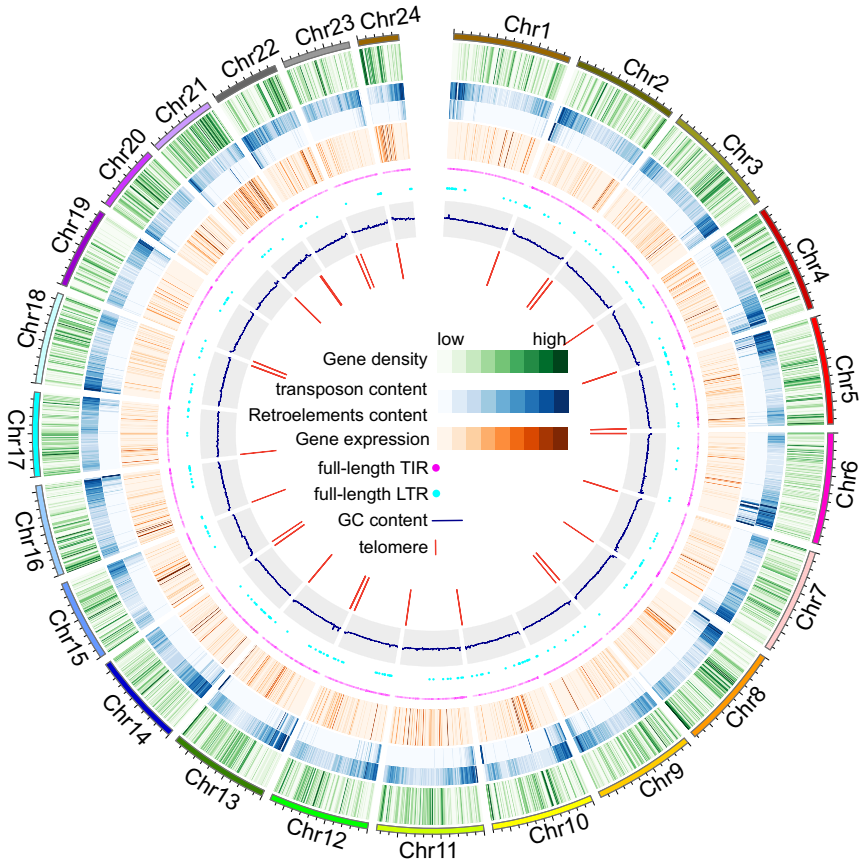

Fig. 3 Circos plot of *M. meleagris* genomic features. From the outer to the inner track: Gene density, Transposon coverage, Retroelements coverage, Gene expression, Distribution of full-length TIRs, Distribution of full-length LTRs, GC content and Telomeric regions.

(GCA_050084645.1), *Xyrichtys novacula* (GCA_962446985.1), *Labroides dimidiatus* (GCA_030710495.1), *Symphodus tinca* (GCA_965165685.1), *Semicossyphus pulcher* (GCF_022749685.1), *Halichoeres trimaculatus*

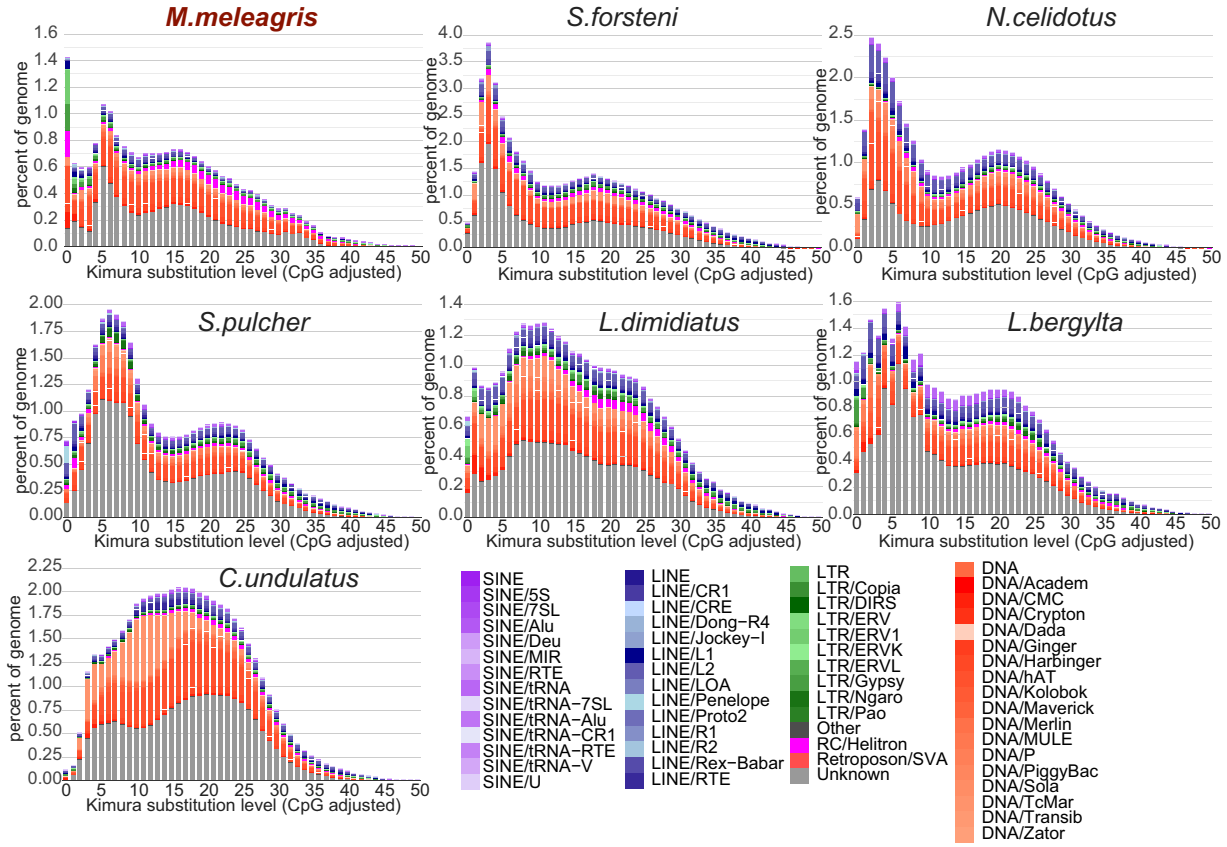

Fig. 4 Repeat landscape showing the abundance (Y axis) and Kimura-2-divergence (X axis) profiles for seven Labridae fishes.

(GCA_033239565.1), *Thalassoma bifasciatum* (GCA_008086565.1), *Scarus forsteni* (GCA_036419115.1), *Cetoscarus ocellatus* (GCA_037902135.1), *Thalassoma lucasanum* (GCA_046269465.1), *Cheilinus undulatus* (GCA_049815525.1)), listed in Table 3, were also assessed using BUSCO[17–20,22–28]. BUSCO analysis showed that the *M. meleagris* genome exhibited high completeness and continuity among all available Labridae genomes, with 98.7% Complete genes and 0.9% fragmented genes. Among the 19 analysed Labridae fish genomes, only two species - *T. bifasciatum* and *T. lucasanum* - which were assembled solely using Illumina short-read data, exhibited BUSCO completeness scores below 90% (Fig. 2a and b).

Genome annotation. Genome annotation primarily involves three components: repeat identification, gene structure prediction, and functional annotation. Repetitive elements were annotated using a combination of homology-based and *de novo* methods. RepeatModeler v1.0.11 and EDTA v2.2.2 were used to construct *de novo* repeat libraries[29,30]. Then, these two libraries were combined with Repbase database v21.12[31]. RepeatMasker v4.0.7 (http://www.repeatmasker.org) was then used to identify repetitive elements using the combined repeat library. EDTA can identify intact TIR repeats (Terminal Inverted Repeat transposons) with TIR (Terminal Inverted Repeat sequence) and TSD (Target Site Duplication) segments on both ends. Using the same pipeline, we also predicted repetitive sequences in six other species of Labridae (*C. undulatus*, *L. bergylta*, *L. dimidiatus*, *N. celidotus*, *S. forsteni* and *S. pulcher*). These six species were chosen for their evolutionary representativeness and the availability of high-quality, chromosome-level genome assemblies. The Kimura distances between genomic TE copies and their corresponding consensus sequences from the combined repeat library were calculated using the calcDivergenceFromAlign.pl, and createRepeatLandscape.pl scripts (located in the RepeatMasker utility directory), based on the aligned file (cat.gz) generated by RepeatMasker.

Repetitive elements accounted for 27.63% of *M. meleagris* genome, consisting primarily of long terminal repeats DNA transposons (15.96%) and (LTRs, 3.17%) (Fig. 3). We identified 5,346 and 276 intact TIRs and LTR-RTs, indicating recent transposable element mobility (Fig. 3). Multiple DNA transposon and LTR families displayed Kimura distances of $< 4\%$ from their consensus sequences, a threshold indicative of recent evolutionary activity[32]. This indicates that many DNA and LTR elements are active and recently inserted into *M. meleagris* genome. DNA elements are active among other six Labridae fish genomes (*S. pulcher*, *S. forsteni*, *N. celidotus*, *L. dimidiatus*, *L. bergylta*, and *C. undulatus*) while LTR elements are active only among *M. meleagris*, *L. dimidiatus*, and *L. bergylta*, the three species belong to the Julidines of Libridae (Fig. 4). The differential activity of DNA and LTR elements likely arises from distinct evolutionary dynamics. DNA transposons sustained activity across deeper evolutionary timescales, as observed in all six compared species. LTR retrotransposons, however, show activity restricted to the three Julidines (*M. meleagris*, *L. dimidiatus* and *L. bergylta*), suggesting

Chromosome	Start	End	Telomeric copies
Chr1	34460000	34477392	891
Chr2	30950000	30959100	975
Chr3	10000	30000	620
Chr4	1	10000	241
Chr4	31670000	31680000	161
Chr5	31540000	31547129	692
Chr6	10000	20000	856
Chr7	10000	20000	796
Chr8	1	10000	252
Chr8	30670000	30678957	632
Chr9	1	10000	421
Chr10	30000000	30010000	146
Chr11	31190000	31196406	822
Chr12	29440000	29453578	635
Chr13	10000	50000	240
Chr13	29610000	29630000	137
Chr14	27320000	27322552	265
Chr15	1	10000	407
Chr16	1	10000	758
Chr17	1	10000	623
Chr18	26900000	26919273	566
Chr19	10000	20000	540
Chr20	19360000	19362728	417
Chr21	17180000	17190000	301
Chr22	19570000	19572090	949
Chr23	1	10000	296
Chr24	10000	30000	910

Table 4. Genomic coordinates and telomeric repeat copy numbers at chromosomal termini.

Genome annotation	*M. meleagris*
Number of protein-coding genes	21,940.00
Average gene length (bp)	9253.99
Average coding length of each gene(bp)	1592.99
Average exon length (bp)	339.89
Average exon number per gene	8.92
Average CDS length (bp)	178.46
Average intron length (bp)	959.39
Repeat sequence proportion (%)	27.63
LTR (%)	3.17
LINE (%)	1.57
SINE (%)	0.19
DNA transposons (%)	15.96

Table 5. Statistics of the genome-wide annotations in *M. meleagris*.

a lineage-specific event. Differences in repeat activity among species can shape genome plasticity, regulatory divergence, and species-specific adaptation.

TIDK v0.2.65 (a telomere identification toolKit) was used to identify and find telomeric repeats (TTAGGG)n in *M. meleagris* genome using the 'find' command with the parameter -c Labriformes[33]. Telomeric repeats (TTAGGG)n were detected at the termini of all 24 chromosomes in *M. meleagris*, with the longest array containing 975 tandem repeats (Table 4).

Protein-coding genes were annotated using a combination of ab initio gene prediction, homology-based prediction and transcriptome evidence. Softwares Augustus v3.3.2 (with the zebrafish parameters), GlimmerHMM v3.0.4 and Genscan v1 were used to predict *de novo* gene models[34,35]. Exonerate v2.2.0 was used for homology-based predictions with proteins of *Danio rerio* (assembly GCF_000002035.6; downloaded from https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/002/035/GCF_000002035.6_GRCz11/GCF_000002035.6_GRCz11_genomic.fna.gz), *Oreochromis niloticus* (assembly GCF_001858045.2;

Type	Gene number	Percentage (%)
Total	21,940	100
SWISSPROT	18,698	85.22
NR	21,215	96.69
KEGG	17,725	80.78
GO	13,917	63.43
TREMBL	21,120	96.26
IPR	18,736	85.39
KOG	13,799	62.89
Annotated in at least one database	21,670	98.76
Annotated genes excluding NR-only hits	21,419	97.63

Table 6. The number of genes annotated in databases.

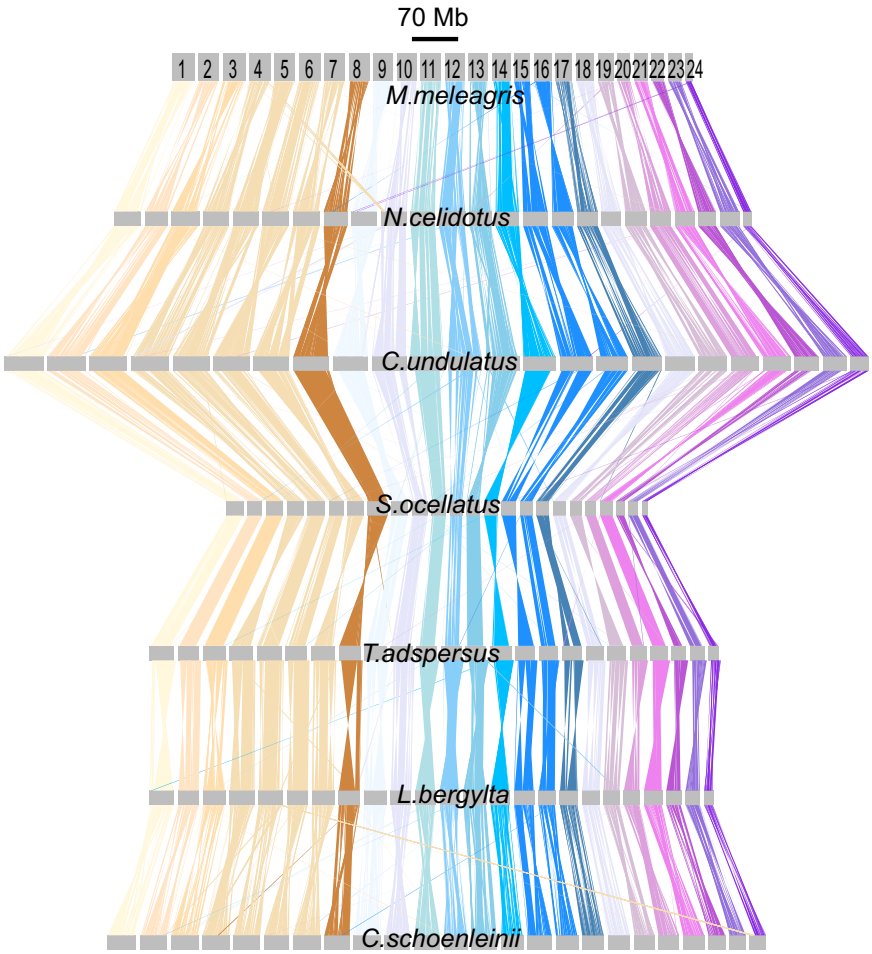

Fig. 5 Chromosomal synteny across seven Labridae fishes.

downloaded from https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/001/858/045/GCF_001858045.2_O_niloticus_UMD_NMBU/GCF_001858045.2_O_niloticus_UMD_NMBU_genomic.fna.gz), *Cheilinus undulatus* (assembly GCF_018320785.1; downloaded from https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/018/320/785/GCF_018320785.1_ASM1832078v1/ GCF_018320785.1_ASM1832078v1_genomic.fna.gz) and *Oryzias latipes* (assembly GCF_002234675.1; downloaded from https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/002/234/675/GCF_002234675.1_ASM223467v1/ GCF_002234675.1_ASM223467v1_genomic.fna.gz) as reference[36]. Transcriptome-based annotation was carried out using two complementary strategies: (i) RNA-seq reads were mapped to the assembled genome using HISAT2 v2.0.4, sorted with SAMtools v1.17 and assembled into transcripts using StringTie v2.1.3b[37,38]. Open reading frame was identified from transcripts using TransDecoder v3.0.1[39]; (ii) Trinity v2.8.5 was used for *de novo* assembly of transcripts from RNA-seq data and then PASA v2.5.2 was used to align assembled transcripts to the genome, obtaining spliced alignments of transcripts and model

gene structure[39,40]. EvidenceModeler (EVM) v2.0.0 was used to integrate these gene models via these different methods and PASA v2.5.2 was used to update protein-coding gene sets[41].

Using a combination of *de novo* prediction, homology-based annotation, and transcriptome evidence, a total of 21,940 protein-coding genes were identified in the *M. meleagris* genome, with an average gene length of 9,253.99 bp and an average coding sequence (CDS) length of 1,592.99 bp of each gene (Table 5). BUSCO assessment (actinopterygii_odb10) indicated that the predicted genes were 95% complete, with 87.3% Complete and single-copy BUSCOs, 7.7% Complete and duplicated BUSCOs, 1.9% Fragmented BUSCOs, and 3.1% Missing BUSCOs.

The predicted gene sets were functionally annotated using diamond v2.0.7.145 with an E-value cutoff of 1e-5 against databases Swiss-Prot (http://www.expasy.ch/sprot) and NR databases through sequence homology searches. Functional annotations, including Gene Ontology (GO) terms and KEGG Orthology (KO) identifiers, were extracted from the best-hit records. These KO identifiers were subsequently used to map genes to KEGG pathways.[42–44]. Functional annotation was also searched against the TREMBL, InterPro (IPR), and KOG databases. A total of 21,670 genes - representing 98.76% of all predicted genes - were functionally annotated using public databases.

We re-calculated the annotation rate after filtering out genes annotated exclusively by NR. This yielded a revised rate of 97.63% for the filtered set of 21,419 genes (Table 6).

For collinearity analysis among seven Labridae fish genomes (*M. meleagris*, *N. celidotus*, *C. undulatus*, *S. ocellatus*, *T. adspersus*, *L. bergylta* and *C. schoenleinii*), an all-versus-all sequence search was conducted between the predicted gene sets of the two Labridae species using diamond v2.0.7.145 with an E-value cutoff of 1e-5. MCscanX v1.0.0 was used to detect syntenic blocks with default parameters[45]. We observed chromosome-scale synteny conservation among seven Labridae fish genomes (Fig. 5).

Data Records

The raw sequencing data for *M. meleagris* (short-read WGS Illumina data, long-read PacBio HiFi data and Hi-C data) generated in this study are available in the Sequence Read Archive (SRA) under SRP605163 and accession numbers SRR34792648, SRR34822252 and SRR34822253[46]. The assembled genome has been deposited in GenBank under accession number JBQPRG010000000[47]. The genome annotations can be accessed via Figshare (https://doi.org/10.6084/m9.figshare.30827831)[48].

Technical Validation

Four complementary approaches were used to evaluate the quality of the *M. meleagris* genome assembly. BUSCO v5.2.2 was used to evaluate the completeness of *M. meleagris* genome assembly with the actinopterygii_odb10 database (containing 3,640 orthologs). The result showed 98.7% completeness, including 98.0% complete single-copy, 0.7% duplicated, and 0.4% fragmented genes, with 0.9% missing (Fig. 2b). Furthermore, 97.46% of the Hi-C data were anchored to the 24 pseudo-chromosomes, confirming the accuracy of the chromosome-level assembly, and the Hi-C heatmap clearly visualized the interaction contacts both within and between chromosome regions (Fig. 1b). To evaluate the assembly's accuracy, WGS Illumina short reads, PacBio HiFi long reads and RNAseq reads were re-aligned to the genome using BWA v0.7.17-r1188, minimap2 v2.24, and hisat2 v2.2.1 respectively[37,49,50]. The high mapping rates − 98.05% for WGS Illumina reads, 99.86% for HiFi reads and 92.64% for RNAseq reads - demonstrate the reliability of the genome assembly. This high level of synteny among seven Labridae fish genomes underscores the extensive conservation of homologous genes across corresponding chromosomes and further validates the accuracy and integrity of the current genome assembly (Fig. 5).

Ethics approval. All animal experiments were approved by the Committee for Research Ethics and Academic Conduct, South China Sea Institute of Oceanology, Chinese Academy of Sciences (approval number: SCSIO-2022-0011).

Data availability

The raw sequencing data are available in the NCBI SRA repository SRP605163 under BioProject accession number PRJNA1299242[46]. The genome assembly has been deposited in GenBank under the accession number GCA_054073175.1[47]. The annotated files can be accessed via Figshare (https://doi.org/10.6084/m9.figshare.30827831)[48].

Code availability

All data analyses in this study were conducted strictly following the official manuals and protocols of the respective software tools, as detailed in the Methods section. No custom scripts or user-defined code were employed.

Received: 5 September 2025; Accepted: 3 February 2026;
Published online: 17 February 2026

References

1. Lucon-Xiccato, T. Inhibitory control in teleost fish: a methodological and conceptual review. *Animal Cognition*. **27**(1), https://doi.org/10.1007/s10071-024-01867-5 (2024).
2. Tariel-Adam, J. *et al.* Tool use by New World Halichoeres wrasses. *Coral Reefs*. **44**(3), 791–807, https://doi.org/10.1007/s00338-025-02633-w (2025).
3. Delrieu-Trottin, E., Williams, J. T. & Planes, S. Macropharyngodon pakoko, a new species of wrasse (Teleostei: Labridae) endemic to the Marquesas Islands, French Polynesia. *Zootaxa* **3857**(3), https://doi.org/10.11646/zootaxa.3857.3.6 (2014).

4. Randall, J. E. A Revision of the Indo-Pacific Labrid Fish Genus Macropharyngodon, with Descriptions of Five New Species. *Bull. Mar. Sci.* **28**, 742–770 (1978).
5. Myers, R. F. *Micronesian Reef Fishes: A Practical Guide to the Identification of the Coral Reef Fishes of the Tropical Central and Western Pacific* (Coral Graphics, Barrigada, Guam, 1991).
6. Ewald Lieske, R. M. *Coral Reef Fishes: Indo-Pacific & Caribbean* (Haper Collins Publishers, 1993).
7. Myers, R. F. *Micronesian Reef Fishes: A Comprehensive Guide to the Coral Reef Fishes of Micronesia* (Coral Graphics 1999).
8. Kent E. & Carpenter, V. H. N. *FAO Species Identification Guide for Fishery Purposes. The Living Marine Resources of the Western Central Pacific* (Food and Agriculture Organization of the United Nations, Rome, 2001).
9. Marçais, 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).
10. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved assembly of diploid genomes without parental data. *Nature Biotechnology*. **40**, 1332–1335, https://doi.org/10.1038/s41587-022-01261-x (2022).
11. Hu, J. *et al*. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**(7), 2253–2255, https://doi.org/10.1093/bioinformatics/btz891 (2020).
12. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**(9), 2896–2898, https://doi.org/10.1093/bioinformatics/btaa025 (2020).
13. Servant, N. *et al*. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology* **16**(1), https://doi.org/10.1186/s13059-015-0831-x (2015).
14. 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).
15. Dudchenko, O. *et al*. *De novo* assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**(6333), 92–95, https://doi.org/10.1126/science.aal3327 (2017).
16. 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).
17. Heaton, B. *et al*. The genome sequence of the cuckoo wrasse, Labrus mixtus Linnaeus 1758. *Wellcome Open Research* **9**, https://doi.org/10.12688/wellcomeopenres.23063.1 (2024).
18. Liu, D. *et al*. Chromosome-level genome assembly of the endangered humphead wrasse Cheilinus undulatus: Insight into the expansion of opsin genes in fishes. *Molecular Ecology Resources* **21**(7), 2388–2406, https://doi.org/10.1111/1755-0998.13429 (2021).
19. Dong, Z. *et al*. Near telomere-to-telomere genome assembly of the blackspot tuskfish (Choerodon schoenleinii). *Scientific Data*. **12**(1), https://doi.org/10.1038/s41597-025-04893-1 (2025).
20. Nugent, C. M. *et al*. Whole-genome sequencing reveals fine-scale environment-associated divergence near the range limits of a temperate reef fish. *Molecular Ecology*. **32**(17), 4742–4762, https://doi.org/10.1111/mec.17063 (2023).
21. Waterhouse, R. M. *et al*. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Molecular Biology and Evolution*. **35**(3), 543–548, https://doi.org/10.1093/molbev/msx319 (2018).
22. Tea, Y. K. *et al*. The spotted parrotfish genome provides insights into the evolution of a coral reef dietary specialist (Teleostei: Labridae: Scarini: Cetoscarus ocellatus). *Ecology and Evolution* **14**(3), https://doi.org/10.1002/ece3.11148 (2024).
23. Kang, J. *et al*. Gene losses, parallel evolution and heightened expression confer adaptations to dedicated cleaning behaviour. *BMC Biology*. **21**(1), https://doi.org/10.1186/s12915-023-01682-3 (2023).
24. Lie, K. K. *et al*. Loss of stomach, loss of appetite? Sequencing of the ballan wrasse (Labrus bergylta) genome and intestinal transcriptomic profiling illuminate the evolution of loss of stomach function in fish. *BMC Genomics* **19**(1), https://doi.org/10.1186/s12864-018-4570-8 (2018).
25. Liang, Y. *et al*. *De Novo* Genome Assembly of the Whitespot Parrotfish (Scarus forsteni): A Valuable Scaridae Genomic Resource. *Genes*. **15**(2), https://doi.org/10.3390/genes15020249 (2024).
26. Bernardi, G. *et al*. Reference Genome of the California Sheephead,Semicossyphus pulcher(Labridae, Perciformes), A Keystone Fish Predator in Kelp Forest Ecosystems. *Journal of Heredity* **113**(6), 649–656, https://doi.org/10.1093/jhered/esac032 (2022).
27. Mattingsdal, M. *et al*. A continuous genome assembly of the corkwing wrasse (Symphodus melops). *Genomics* **110**(6), 399–403, https://doi.org/10.1016/j.ygeno.2018.04.009 (2018).
28. Cruz, F. *et al*. Chromosome-level assembly and annotation of the Xyrichtys novacula (Linnaeus, 1758) genome. *DNA Research* **30**(5), https://doi.org/10.1093/dnares/dsad021 (2023).
29. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences*. **117**(17), 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
30. Ou, S. *et al*. Differences in activity and stability drive transposable element variation in tropical and temperate maize. *Genome Research* **34**(8), 1140–1153, https://doi.org/10.1101/gr.278131.123 (2024).
31. Bao, W., K. K. Kojima & O. Kohany. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* **6**(1), https://doi.org/10.1186/s13100-015-0041-9 (2015).
32. Osmanski, A. B. *et al*. Insights into mammalian TE diversity through the curation of 248 genome assemblies. *Science* **380**(6643), https://doi.org/10.1126/science.abn1430 (2023).
33. Brown, M. R., Gonzalez de La Rosa, P. M. & Blaxter, M. tidk: a toolkit to rapidly identify telomeric repeats from genomic datasets. *Bioinformatics*. **41**(2), https://doi.org/10.1093/bioinformatics/btaf049 (2025).
34. Stanke, M. *et al*. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics* **24**(5), 637–644, https://doi.org/10.1093/bioinformatics/btn013 (2008).
35. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**(16), 2878–2879, https://doi.org/10.1093/bioinformatics/bth315 (2004).
36. Slater, G. S. C. and E. Birney. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**(1), https://doi.org/10.1186/1471-2105-6-31 (2005).
37. Pertea, M. *et al*. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols*. **11**(9), 1650–1667, https://doi.org/10.1038/nprot.2016.095 (2016).
38. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16), 2078–2079, https://doi.org/10.1093/bioinformatics/btp352 (2009).
39. Haas, B. J. *et al*. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* **8**(8), 1494–1512, https://doi.org/10.1038/nprot.2013.084 (2013).
40. Campbell, M. A. *et al*. Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. *BMC Genomics* **7**(1), https://doi.org/10.1186/1471-2164-7-327 (2006).
41. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* **9**(1), https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
42. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods* **18**(4), 366–368, https://doi.org/10.1038/s41592-021-01101-x (2021).
43. Ashburner, M. *et al*. Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**(1), 25–29, https://doi.org/10.1038/75556 (2000).
44. Kanehisa, M. S.G. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**(1), 27–30, https://doi.org/10.1093/nar/28.1.27 (2000).

45. Wang, Y. *et al*. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* **40**(7), e49–e49, https://doi.org/10.1093/nar/gkr1293 (2012).
46. *NCBI Sequence Read Archive*. https://identifiers.org/ncbi/insdc.sra:SRP605163).
47. Yu, H., Chao Li, M. Q., Lin, Q. & Wang, D. *GenBank*. https://identifiers.org/ncbi/insdc.gca:GCA_054073175.1 (2025).
48. Yu, H., Chao Li, M. Q., Lin, Q. & Wang, D. Chromosome-level genome assembly of the Leopard Wrasse Macropharyngodon meleagris. *Figshare* https://doi.org/10.6084/m9.figshare.30827831 (2025).
49. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* **26**(5), 589–595, https://doi.org/10.1093/bioinformatics/btp698 (2010).
50. Li, H. & Alkan, C. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**(23), 4572–4574, https://doi.org/10.1093/bioinformatics/btab705 (2021).

Acknowledgements

This research was supported by National Natural Science Foundation of China 42425004 (Q.L.), 42230409 (Q.L.); Guangdong Basic and Applied Basic Research Foundation 2025B1515020030 (M.Q.); Special Fund of South China Sea Institute of Oceanology of the Chinese Academy of Sciences (SCSIO2024QY01).

Author contributions

D.Z. WANG, Q.L. and M.Q. designed the project. C.L. collected and prepared all fish specimens used in this project. H.Y.Y. conducted the bioinformatic analyses. H.Y.Y. drafted the initial manuscript. D.Z. WANG, Q.L. and M.Q. contributed to manuscript revision. D.Z. WANG, Q.L. and M.Q. reviewed the results and provided further revisions. 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 M.Q. or D.W.

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.