scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A telomere-to-telomere gap-free genome assembly of the protandrous maroon clownfish (*Premnas biaculeatus*)

Xinhui Zhang[1,2,7], Cancan Hu[3,7], Jieming Chen[2], Man Yuan[4], Yukai Yang[5], Tao Li[6] ✉ & Qiong Shi[1,2] ✉

As a protandrous hermaphrodite with natural male-to-female sex change, maroon clownfish (*Premnas biaculeatus*) serves as an ideal model organism for investigating sequential hermaphroditism. However, genomic resources for this interesting species remain scarce, thereby limiting in-depth research on its unique biological traits. In this study, we generated the first telomere-to-telomere (T2T) gap-free genome assembly of maroon clownfish by integrating multi-platform sequencing data, including MGI short reads, PacBio HiFi long reads, ONT ultra-long reads, and Hi-C sequencing data. The final haplotypic genome spans 884.39 Mb, with all sequences successfully anchored onto 24 chromosomes. This assembly is highly contiguous, with a contig N50 of 37.98 Mb. Comprehensive genomic characterization revealed the precise localization of telomeric repeats and centromeric region within each chromosome. Independent quality assessments, such as QV of 71.01, CRAQ score of 98.98%, and BUSCO completeness of 99.98%, confirmed good assembly accuracy. Additionally, alignment of ONT ultra-long and PacBio HiFi reads to the assembly yielded a high mapping rate exceeding 99%, further validating a good assembly integrity. Repetitive elements constituted 33.51% (296.37 Mb) of the assembled genome, and a total of 24,556 protein-coding genes were annotated. This high-quality T2T genome assembly will not only provide a valuable genetic resource to advance related research in comparative genomics, population genetics, molecular breeding, and functional genomics of maroon clownfish, but also lay a solid foundation for resolving molecular mechanisms underlying its protandrous reproductive strategy.

Background & Summary

Maroon clownfish (*Premnas biaculeatus*), also named as spinecheek anemonefish, is a marine teleost under the order Perciformes and the family Pomacentridae. However, it is the only clownfish species that does not fall within the genus Amphiprion[1]. As a representative of coral reef ecosystems, maroon clownfish is distinguished by its vivid body coloration, species-specific behaviors, and its obligate symbiotic relationship with sea anemones[2]. These traits make it not only a globally popular ornamental fish but also a classic model organism for investigating ecological and molecular mechanisms of marine symbiosis[3,4].

Evolution of sex change strategies in fish is tightly associated with their mating systems, as these strategies directly impact reproductive fitness[5,6]. Previous field observations and experimental studies have consistently demonstrated that, to maximize reproductive output under the constraints of social structure (e.g., dominance

[1]Laboratory of Aquatic Genomics, College of Life Sciences and Oceanography, Shenzhen University, Shenzhen, 518057, China. [2]Center for Aquatic Genomics, BGI Academy of Marine Sciences, Shenzhen, 518081, China. [3]Shenzhen Research Institute, Ocean University of China, Shenzhen, 518100, China. [4]Shenzhen BGI Marine Sci &Tech Co., Ltd, Shenzhen, 518119, China. [5]Key Laboratory of South China Sea Fishery Resources Exploitation & Utilization, Ministry of Agriculture and Rural Affairs, South China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Guangzhou, 510300, China. [6]Shenzhen Base of South China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Shenzhen, 518121, China. [7]These authors contributed equally: Xinhui Zhang, Cancan Hu. ✉e-mail: lit151415@126.com; shiqiong@szu.edu.cn; shiqiong@genomics.cn

hierarchy), some fish species have evolved a genetic feature of sex change[7]. Sex change and hermaphroditic strategies in bony fishes are taxonomically and functionally diverse, and they can be categorized into four primary types based on the direction and timing of sexual differentiation: protogyny (female-first sex change), protandry (male-first sex change), bidirectional sex change (flexible transition between sexes), and synchronous hermaphroditism (simultaneous possession of functional male and female gonads)[8]. Representative species include Asian seabass (*Lates calcarifer*)[9], black seabream (*Acanthopagrus schlegelii*)[10], and orange-red pygmygoby (*Trimma okinawae*)[11]. Within the family Pomacentridae, fishes of the genera *Amphiprion* and *Premnas* are typically protandrous hermaphrodites, which initially develop as functional males and can later undergo a transition to females. In contrast, some species of the genus *Dascyllus* display protogynous sex change, wherein individuals first appear as females and subsequently shift to males[6]. Despite sharing similarities in reproductive behaviors (e.g., egg guarding) and parental care strategies with other Pomacentridae species, these genera possess highly specialized mating systems, rendering them ideal models for investigating the evolution of sex change and the structure of social behaviors.

Maroon clownfish is the only anemonefish species within the genus *Premnas*. It exhibits strict host specificity to the bubble-tip anemone (*Entacmaea quadricolor*)[12]. Certain studies have demonstrated that maroon clownfish holds a competitive advantage over other organisms inhabiting *Entacmaea quadricolor*[13]. Compared with other anemonefish species, maroon clownfish possesses a unique social structure. That is to say, it typically exists solely in monogamous pairs, rather than in social groups that include immature subadults. Breeding females are generally twice the size of breeding males[14]. Like other anemonefishes, maroon clownfish is a protandrous hermaphrodite, and its social structure is characterized by female dominance[15].

In our present study, we combined MGI short-read, PacBio HiFi long-read, ONT (Oxford Nanopore Technologies) ultra-long, and Hi-C sequencing data to construct a high-fidelity telomere-to-telomere (T2T) genome assembly of maroon clownfish. This assembly was rigorously assessed for quality, and its key genomic features were characterized. This gap-free reference genome assembly will not only facilitate population genetic research and evolutionary study, but also provide an important genetic resource for molecular breeding and investigating molecular mechanisms of sex change in this economically important fish.

Methods

Sample collection. The maroon clownfish used in this study (Fig. 1A) was obtained from a local base of the South China Sea Fisheries Research Institute (SCSFRI), Chinese Academy of Fishery Sciences (CAFS), in Shenzhen city, Guangdong province, China. It had a body length of 4.6 cm and body weight of 6.5 g. For whole-genome sequencing, muscle was dissected from the individual and subjected to multi-platform sequencing, including MGI short-read, PacBio HiFi long-read, ONT ultra-long-read, and Hi-C sequencing technologies. Additionally, transcriptome sequencing (RNA-seq) was performed using four distinct tissue samples (gill, eye, muscle and skin) (Table 1). All sampling procedures and experimental workflows were conducted in compliance with the guidelines established by the Animal Ethics Committee of SCRFRI, CAFS (No. nhdf2025-30).

DNA extraction and genome sequencing. Genomic DNA (gDNA) was extracted from the muscle tissue using a QIAamp DNA Mini kit (Qiagen, Valencia, CA, USA) in accordance with the manufacturer's protocols[16]. Quality of the extracted gDNA was evaluated using three complementary approaches. Fragment size was determined via 0.75% agarose gel electrophoresis, molecular integrity was assessed by an Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA), and absolute quantification was performed using a Qubit Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA).

For the MGI sequencing, a paired-end library with an insert size of 350 bp was constructed using a MGIEasy Universal DNA Library Preparation kit (MGI, Shenzhen, China). Sequencing was then conducted on a DNBSEQ-T7 platform (MGI) to generate 67.56 Gb of paired-end reads. These raw reads were filtered using fastp v0.12.6[17] with the following parameters: -n 0 -f 5 -F 5 -t 5 -T 5 -q 20. This filtering step removed adapter sequences, reads with excessive Ns, and low-quality bases (Phred quality score < 20), resulting in 59.92 Gb of clean reads (Table 1) that were used for subsequent data error correction and genome size estimation.

For the PacBio HiFi sequencing, approximately 10 μg of high-quality gDNA was used to construct a SMRTbell library following the standard protocol provided with the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, Menlo Park, CA, USA). The library was sequenced on a PacBio Sequel II System using the circular consensus sequencing (CCS) technology. HiFi reads were generated using CCS v6.0.0[18] with the optimized parameter "-min-passes 3". This process yielded a total of 101.76 Gb of HiFi reads with a N50 of 17,331 bp (Table 1).

Two ultra-long read libraries were prepared following Oxford Nanopore Technologies (ONT) standard protocols and sequenced on a PromethION platform (Oxford Nanopore Technologies Co., Littlemore, Oxford, UK). Raw ONT reads were quality-filtered using NanoFilt v2.8.0[19] to remove those reads with a quality value (QV) < 7. After filtering, 72.04 Gb of clean ultra-long reads were retained, which had an average length of 40,256 bp and a N50 of 62,144 bp (Table 1).

For the high-throughput chromosome conformation capture (Hi-C) sequencing, a Hi-C library was constructed using a GrandOmics Hi-C kit (GrandOmics, Wuhan, Hubei, China) according to the manufacturer's protocol. This library was sequenced on a DNBSEQ-T7 platform (MGI) using the 150-bp paired-end sequencing mode to generate 106.47 Gb of raw data. These raw reads were filtered using fastp v0.12.6[17] to remove adapters and low-quality sequences, resulting in 104.07 Gb of clean data (Table 1) for subsequent chromosome-level genome assembly.

RNA extraction and transcriptome sequencing. Total RNA was extracted from four tissues (Table 1) separately according to a standard Trizol protocol (Invitrogen, Frederick, MD, USA), followed by purification

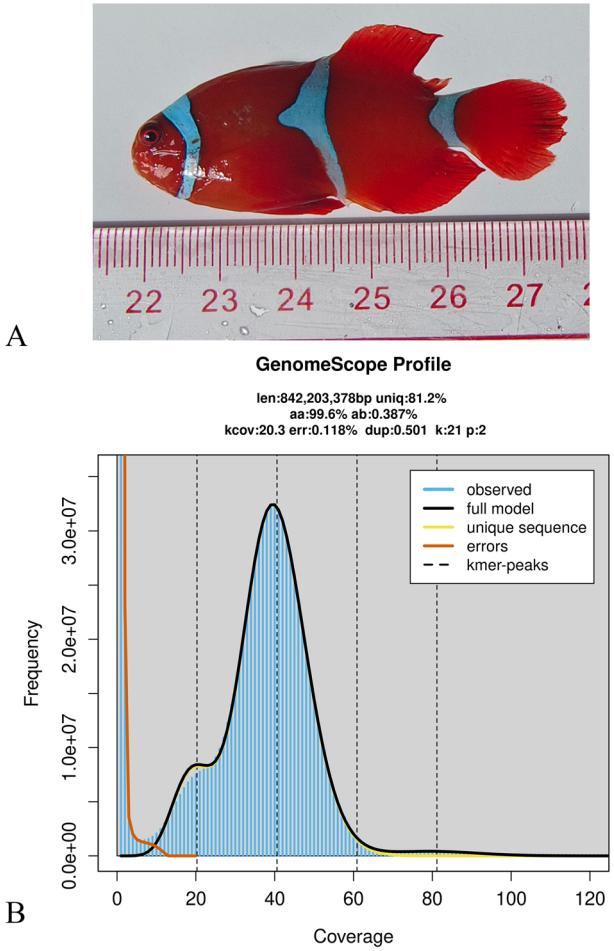

Fig. 1 Maroon clownfish and its whole-genome sequence distribution. (**A**) A morphological image of the sequenced fish. (**B**) A k-mer (21-mer) distribution curve for estimation of the genome size.

Type	Library type	Raw data (Gb)	Clean data (Gb)	Read N50/ length (bp)	Coverage of the genome (×)
Genome sequencing	MGI	67.56	59.92	150	71.16
	PacBio HiFi	/	101.76	17,331*	120.85
	ONT Ultra-long	/	72.04	62,114*	85.55
	Hi-C	106.47	104.07	150	123.59
Transcriptome sequencing	Eye	6.237	5.525	150	/
	Muscle	6.355	5.721	150	/
	Skin	6.177	5.543	150	/
	Gill	8.021	7.295	150	/

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

with a Qiagen RNeasy Mini Kit (Qiagen, Germantown, MD, USA). RNA concentration and integrity were measured using a NanoDrop 8000 Spectrophotometer (Thermo Fisher Scientific) and an Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA), respectively. Only those RNA samples with OD260/280 $\geq$ 1.8 and RNA integrity $\geq$ 7.0 were selected for transcriptome sequencing. RNA was used for construction of a cDNA library followed the manufacture's guideline, which was then sequenced on a HiSeq X Ten platform (Illumina, San Diego, CA, USA). A total of 48.07 Gb of transcriptome raw data were generated (Table 1), which aided in annotation of protein-coding genes and prediction of gene structures.

Genome-size estimation and construction of a T2T genome assembly. To estimate the genome size of maroon clownfish, k-mer counting was performed using Jellyfish v2.2.10[20] with a k-mer length of 21. The analysis parameters were configured as follows: -m 21 -s 10 G -C. A k-mer frequency histogram generated by Jellyfish was used as the input for GenomeScope v2.0[21], a tool designed to infer genome-wide genetic

Category	Data
Genome survey (Mb)	842.2
Genome length (bp)	884,393,502
Longest scaffold (bp)	45,429,913
Number of scaffolds	24
Contig N50 (bp)	37,983,415
Scaffold N50 (bp)	37,983,415
GC content	39.4%
BUSCO	99.97% (S:99.78%; D:0.19%)
Number of chromosomes	24
Chromosome length (bp)	884,393,502
Repetitive sequence	33.51%

Table 2. Summary of the genome assembly for maroon clownfish. Abbreviations: S, single copy complete genes; D, duplicated complete genes.

characteristics from k-mer distribution patterns. This analytical workflow allowed for sequence-based estimation of genome features prior to *de novo* assembly, eliminating potential biases introduced by assembly artifacts. As shown in Fig. 1B, the genome size of maroon clownfish was calculated to be approximately 842.2 Mb. Additional key genomic characteristics derived from this analysis include an estimated heterozygosity of 0.387%.

Primary contigs were initially constructed via *de novo* assembly of PacBio HiFi reads and ONT ultra-long reads using Hifiasm v0.19.8[22] with default parameters. Subsequent to contig generation, the purge_dups v1.2.5[23] pipeline (--low 65 --mid 115 --high 150 --min-length 1000) was applied to filter out haplotypic duplications and heterozygous redundant sequences from the initial *de novo* assembly. This quality control step resulted in a genome assembly with a total length of 884.09 Mb.

Using this preliminary genome assembly as a reference, Hi-C clean reads were utilized to perform chromosome-level scaffolding. The workflow was implemented as follows: (1) Hi-C clean reads were mapped to the assembled contigs using Bowtie2 v2.2.5[24] with the following parameter settings: --very-sensitive -L 20 --score-min L,-0.6,-0.2 --end-to-end. (2) The HiC-Pro v2.8.1 pipeline[25] was applied to detect Hi-C ligation products, with only valid paired-end reads retained for downstream analysis to ensure high reliability of interaction signals. (3) Based on the valid Hi-C reads, the preliminary contig assembly was clustered into chromosome groups, ordered, and oriented to construct chromosome-level scaffolds using Juicer v1.5[26] and 3D-DNA v3.0[27]. Juicebox v1.11.08[28] was employed to visualize the chromosome-level assembly. Manual adjustments were further conducted to correct misclusters, misorders, or misorientations in the candidate chromosome assemblies, ensuring good consistency with Hi-C interaction patterns.

To eliminate residual gaps in the chromosome-level assembly, corrected ultra-long ONT reads were used for gap filling. Two complementary tools were implemented, including TGS-GapCloser v1.2.1[29] with optimized parameters: --min_match 1000 --min_nread 3, and LR_Gapcloser v1.0[30] with the parameter settings: -t 35 -m 1000000 -v 500. The final gap-free genome assembly of maroon clownfish has a total length of 884.39 Mb (Table 2), with all sequences anchored onto 24 chromosomes (Chr; Fig. 2). The longest Chr is 45.43 Mb, while the shortest Chr is 26.18 Mb (Table 3). The number of assembled chromosomes (haplotypic n = 24) is consistent with the previously established diploid karyotype (2n = 48) of this species[31].

Identification of centromere and telomere sequences. Telomeres were annotated by searching for the conserved telomeric repeat motif (CCCTAA/TTAGGG) at both terminal regions of each chromosome using the Telomere-to-Telomere (T2T) Toolkit quarTeT v1.1.1[32]. To annotate centromeres in maroon clownfish, we first identified repetitive sequences using TRF v4.0.4[33] and RepeatMasker v4.0.6[34], generating a comprehensive transposable element (TE) annotation file. This TE annotation was then used as an input for quarTeT v1.1.1[32] to predict candidate centromeric intervals based on the enrichment of centromere-associated repetitive elements. Collectively, our annotations confirmed that the assembled genome contains a complete set of 24 centromeres and 48 telomeres (Fig. 3), consistent with the previously established diploid karyotype for this species[31].

Annotation of repeat elements. To annotate repetitive elements (REs) in the maroon clownfish genome, tandem repeats were identified using Tandem Repeats Finder (TRF) v4.0.4[33] and Genome-wide Microsatellite Analyzing Tool Package (GMATA) v2.2[35]. Specifically, TRF was employed to detect simple sequence repeats (SSRs), while GMATA was applied to characterize all tandem repeat families across the entire assembled genome, ensuring comprehensive coverage of tandemly repeated sequences.

For transposable element (TE) annotation, a combined strategy of homology-based and *de novo* prediction was adopted to minimize false negatives and improve annotation accuracy. In the former approach, RepeatMasker v4.0.6 and RepeatProteinMask v4.0.6[34] were utilized: RepeatMasker aligned the assembled genome against the RepBase database (v26.04) to identify known TE families, while RepeatProteinMask detected TE-related sequences via homology to conserved TE protein domains. In the latter approach, RepeatModeler v1.0.8[36] and LTR_FINDER v1.0.6[37] were used to construct a *de novo* repeat library with candidate TE consensus sequences. This custom library was then used as a reference for RepeatMasker to annotate *de novo*-identified REs. Finally, the integrated annotation revealed a total of 296.37 Mb of repetitive sequences in

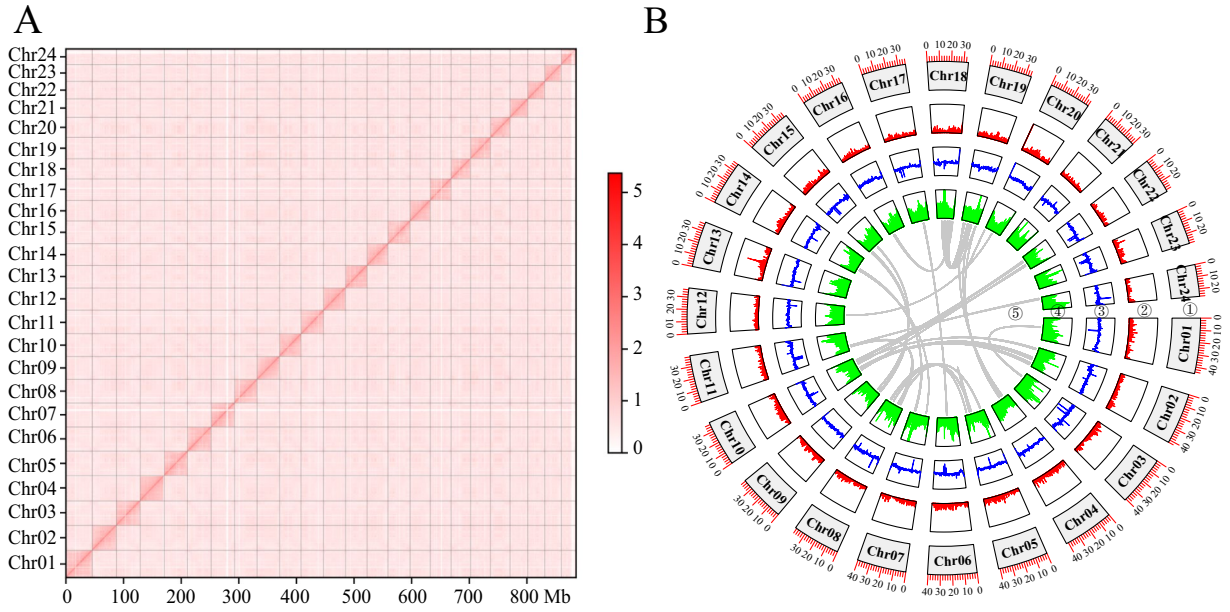

Fig. 2 T2T genome assembly of maroon clownfish. (**A**) Genome-wide chromatin interactions at a 500-kb resolution. Color blocks represent corresponding interactions, with various strengths from white (low) to red (high). (**B**) A Circos plot of the main genomic features. From outside to inside the details include (1) the 24 chromosomes, (2) gene density, (3) GC content, (4) repetitive sequences density, and (5) a colinear relationship among chromosomes of the maroon clownfish genome assembly. Note that the density calculation window is set as 100 kb.

Chr	Contig	Length (bp)	Gap	Telomere (Te)				Centromere (Ce)	
				Upstream Start	Upstream End	Downstream Start	Downstream End	Start	End
Chr01	1	45,429,913	0	18	7,777	45,424,111	45,429,849	43,439,296	43,595,693
Chr02	1	42,086,916	0	5	1,170	42,086,515	42,086,895	5,885,772	6,130,821
Chr03	1	41,148,849	0	39	7,913	41,148,195	41,148,782	27,282,660	27,598,539
Chr04	1	41,717,530	0	24	8,042	41,713,346	41,717,529	41,260,044	41,713,346
Chr05	1	41,245,558	0	12	4,949	41,240,084	41,245,509	5,024	458,698
Chr06	1	40,440,592	0	14	718	40,439,730	40,440,590	1,700,066	1,991,945
Chr07	1	40,413,786	0	50	5,475	40,407,608	40,413,783	26,946,047	29,073,168
Chr08	1	39,140,442	0	1	498	39,138,728	39,140,427	328,499	559,843
Chr09	1	38,552,306	0	55	3,956	38,548,122	38,552,305	38,253,117	38,548,122
Chr10	1	37,764,731	0	5	4,938	37,763,375	37,764,729	32,153,794	32,247,715
Chr11	1	39,008,671	0	14	5,813	39,007,100	39,008,621	37,391,374	37,670,170
Chr12	1	37,502,806	0	5	4,515	37,502,134	37,502,749	2,515,076	2,811,636
Chr13	1	37,542,707	0	70	756	37,541,451	37,542,663	2,629,628	2,988,637
Chr14	1	36,292,040	0	3	7,508	36,286,943	36,292,036	35,943,122	36,159,383
Chr15	1	37,983,415	0	31	1,268	37,981,857	37,983,396	32,751,039	32,883,769
Chr16	1	34,972,747	0	42	5,846	34,970,898	34,972,743	33,897,919	34,268,553
Chr17	1	36,157,774	0	28	5,268	36,151,211	36,157,755	33,637,756	34,246,320
Chr18	1	33,640,073	0	21	2,177	33,639,010	33,640,068	14,002,114	14,498,088
Chr19	1	34,718,311	0	3	8,655	34,717,278	34,718,272	24,235,468	24,679,529
Chr20	1	34,203,512	0	16	917	34,193,813	34,203,512	30,593,062	31,055,797
Chr21	1	30,403,700	0	48	2,624	30,399,253	30,403,696	1,719,076	2,030,200
Chr22	1	29,553,411	0	57	2,295	29,546,498	29,553,357	28,475,090	29,103,835
Chr23	1	28,298,408	0	3	7,687	28,294,224	28,298,407	27,757,305	28,294,224
Chr24	1	26,175,304	0	32	740	26,169,830	26,175,255	25,740,005	25,880,411

Table 3. Telomere and centromere positions in the assembled genome.

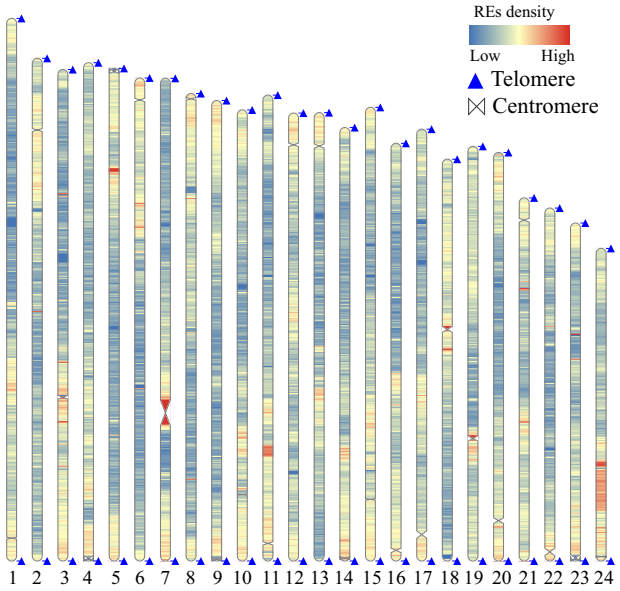

Fig. 3 Genome-wide localization of repetitive elements (REs), telomeres, and centromeres. Triangles at both ends of each chromosome represent the telomere regions, and the gully area within each chromosome stands for a centromere region.

Type			Length (bp)	Count	% of Genome
Dispersed repeats	DNA transposons		154,669,455	878,200	17.49
	Retroelements	LINE	60,483,307	268,517	6.84
		LTR	29,085,520	180,620	3.29
		SINE	8,666,949	61,382	0.98
Tandem Repeats	Simple repeats		11,635,726	161,597	1.32
	Satellites		5,201,134	23,229	0.59
Other			26,635,722	187,062	3.01
Total			296,377,811	1,883,021	33.51

Table 4. Classification of repetitive sequences in maroon clownfish genome.

the maroon clownfish genome, accounting for 33.51% of the assembled genome (Table 4, Fig. 3). Among these REs, DNA transposons were the most abundant component, representing 17.49% of the genome (154.69 Mb), followed by long interspersed nuclear elements (LINEs; 6.84%, 60.48 Mb) and long terminal repeats (LTRs; 3.29%, 29.08 Mb) (Table 4).

Prediction and functional annotation of protein-coding genes. Prior to protein-coding gene prediction and structural annotation, repetitive regions in the assembled genome were masked to minimize interference from repetitive sequences on gene model construction. An integrated strategy of *de novo* prediction, homology-based annotation, and RNA-seq-supported annotation was applied for comprehensive protein-coding gene annotation. First, AUGUSTUS v3.2.1[38] and GlimmerHMM v3.0.4[39] were employed to perform the *ab inito* gene structure prediction. Second, GeMoMa v1.6.4[40] was employed to transfer gene annotations from five evolutionarily related fish species. Second, homologous protein sequences were downloaded from the NCBI database for the following species: *Amphiprion ocellaris* (clown anemonefish, GCA_022539595.1), *Amphiprion clarkii* (yellowtail clownfish, GCA_027123335.1), *Lates calcarifer* (Asian seabass, GCA_051027255.1), *Perca flavescens* (yellow Perch, GCA_004354835.1) and *Stegastes partitus* (Bicolor damselfish, GCA_000690725.1). These protein sequences were aligned to the masked genome so as to guide homology-based gene structure prediction. Third, RNA-seq data derived from the four tissues (Table 1) were assembled into transcript contigs with Trinity v2.5.1[41] using default parameters. These assembled transcripts were then used as evidence for gene structure refinement via PASA v2.3.3[42]. Finally, gene models generated from the three approaches were integrated using the Evidence Modeler (EVM) pipeline v1.0[43], with weights assigned to each evidence type (RNA-seq > homology > *de novo*) to prioritize reliable gene structures.

This integrated annotation yielded a total of 24,556 high-confidence protein-coding genes, with an average gene length of 19.72 kb and an average coding sequence (CDS) length of 1,734.11 bp (Table 5). Completeness of the predicted gene set was evaluated using Benchmarking Universal Single-Copy Orthologs (BUSCO) v5.4.7[44]

Method	Software/Species	Number	Average length (bp)				Average exon per gene
			gene	CDS	exon	intron	
De novo	Augustus	26,489	19,176.32	1,661.49	166.42	1,949.57	9.98
	Glimmer	53,694	15,371.11	981.61	141.69	2,427.35	6.93
Homolog	*A. ocellaris*	53,416	31,036.39	1,656.04	180.18	3,586.85	9.19
	A. clarkii	53,495	30,607.7	1,677.97	184.54	3,574.79	9.09
	L. calcarifer	53,774	32,665.24	1,643.63	178.07	3,769.15	9.23
	P. flavescens	54,166	32,262.33	1,634.2	180.44	3,801.43	9.06
	S. partitus	51,717	29,794.26	1,655.67	179.93	3,430.75	9.2
RNA-seq	PASA	22,940	24,147.33	3,877.47	331.28	1,893.56	11.7
Integrated	EVM	24,556	19724.07	1,734.11	170.76	1,964.99	10.16

Table 5. Summary of the predicted gene structures.

Database	Number	Percentage (%)
Total	24,556	100
NR	23,234	94.62
Swissprot	21,102	85.93
KEGG	16,581	67.52
GO	15,913	64.80
KOG	15,527	63.23
Overall	23,361	95.13

Table 6. Functional annotation of predicted protein-coding genes. Overall represents the total number of annotated genes with at least one hit from the five searched databases.

against the actinopterygii_odb10 database revealing that 99.12% of the complete BUSCO orthologs were successfully recovered, which confirms the high quality of gene annotation.

Functional annotation of these predicted protein-coding genes was conducted using Blastp v2.2.26[45], with deduced protein sequences aligned against five publicly available databases, including NCBI Non-Redundant Protein Sequence (NR) database, Swiss-Prot[46], Gene Ontology (GO)[47] (for functional classification), Kyoto Encyclopedia of Genes and Genomes (KEGG)[48] (for pathway annotation), and EuKaryotic Orthologous Groups (KOG)[49] (for orthologous gene family classification). All alignments were performed with an E-value threshold of 1e-5 to ensure significant sequence homology. In total, 23,361 protein-coding genes (95.13% of the total predicted gene set) were successfully annotated with at least one functional hit in at least one of the five searched databases (Table 6).

Data Records

Sequencing data (including MGI, PacBio, ONT, Hi-C, and transcriptome sequencing) and the assembled genome of maroon clownfish were deposited in the National Center for Biotechnology Information (NCBI) database under the BioProject accession number PRJNA1306080[50]. Raw sequencing reads are available in the Sequence Read Archive (SRA) with accession number SRP617804[51]. The genome assembly, predicted coding sequences, and functional annotation files of maroon clownfish were deposited in Figshare (No: https://doi.org/10.6084/m9.figshare.30104977)[52]. The complete genome assembly has also been deposited at the NCBI under accession number of GCA_053813585.1[53].

Technical Validation

To evaluate the quality of maroon clownfish genome assembly, four complementary approaches were employed (Table 7). First, BUSCO v5.4.7[44] was employed to evaluate assembly completeness against the actinopterygii_odb10 database. Our results showed that 99.98% of the complete BUSCOs were recovered, including 99.73% as single-copy complete genes (S) and 0.25% as duplicated complete genes (D), confirming high overall completeness of the assembly. Second, Merqury v1.328[54], a k-mer-based quality evaluation tool, was applied to assess base-level accuracy and assembly completeness. K-mers were generated from MGI short reads and PacBio HiFi long reads, respectively. This analysis yielded a Quality Value (QV) of 44.37 (short reads) and 71.01 (long reads), indicating high base-level accuracy of the assembly. Third, Clipping information for Revealing Assembly Quality (CRAQ) v1.09[55], a tool leveraging read clipping signals to quantify assembly accuracy, was used with PacBio HiFi and Illumina reads. The analysis determined a Reference-based Assembly Quality Indicator (R-AQI) of 97.93 and a Sample-based Assembly Quality Indicator (S-AQI) of 98.98, further verifying high accuracy of the assembled genome. Fourth, we mapped the sequencing data to the assembled genome using bwa v0.7.17[56] and minimap2 v2.26[57], which demonstrated mapping rates of 99.46% for the MGI data, 99.99% for the PacBio data, and 98.43% for the ONT data. Collectively, these quality metrics confirm that the maroon clownfish genome assembly is of high quality.

Type	Evaluation Methods		Results
Genome accuracy and completeness	Mapping short reads rate		99.21%
	Mapping HiFi reads rate		99.99%
	Mapping ONT reads rate		99.26%
	QV	Short reads	44.37
		HiFi reads	71.01
	CRAQ	R-AQI	97.93%
		S-AQI	98.98%
	BUSCO		99.98%
Annotation quality	Complete BUSCOs		99.12% (3,608)
	Complete and single-copy BUSCOs (S)		98.41% (3,582)
	Complete and duplicated BUSCOs (D)		0.71% (26)
	Fragmented BUSCOs (F)		0 (0)
	Missing BUSCOs (M)		0.88% (32)
	RNA-seq coverage ratio of the exonic regions		95.89%

Table 7. Assessment metrics of the genome assembly and annotation.

For the predicted protein-coding genes, additional quality evaluations were performed. In fact, BUSCO assessment against the actinopterygii_odb10 database showed a completeness value of 99.12% (Table 7), indicating comprehensive gene prediction. To validate annotation accuracy, transcriptome data were aligned to the assembled genome using STAR v2.7.11b[58], and exonic coverage was calculated using BEDTools v2.29.2[59]. Our results showed that 95.89% of exonic regions were covered by sequencing reads, further confirming high accuracy of the gene annotation (Table 7).

Data availability

In this study, sequencing data for *P. biaculeatus*, including MGI, PacBio, Nanopore, Hi-C, and transcriptome sequencing, have been deposited in the NCBI BioProject database under the accession number PRJNA1306080[50]. The raw sequencing reads are available in the NCBI Sequence Read Archive (SRA) under the accession numbers SRP617804[51]. The final genome assembly and predicted coding sequences files were stored in Figshare (No: m9.figshare.30104977)[52]. The genome assembly has also been deposited at the NCBI under accession number of GCA_053813585.1[53] for public availability.

Code availability

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

Received: 14 October 2025; Accepted: 24 December 2025;
Published online: 04 February 2026

References

1. Allen, G.R. Damselfishes of the South Seas. *New Jersey: Tropical Fish Hobbyist Publications* 1-273 (1975).
2. Dixson, D. L. *et al.* Influence of prior residents on settlement preferences in the anemonefish, *Premnas biaculeatus*. *Coral Reefs* **37**(2), 519–526 (2018).
3. Herrera, M. *et al.* Anemonefishes: A model system for evolutionary genomics. *F1000Research* **12**, 204 (2023).
4. Marcionetti, A. *et al.* Insights into the genomics of clownfish adaptive radiation: the genomic substrate of the diversification. *Genome Biology and Evolution* **15**(7), evad088 (2023).
5. Warner, R. R. Sex change and the size-advantage model. *Trends in Ecology & Evolution* **3**(6), 133–136 (1988).
6. Hattori, A. *et al.* Sex change strategies and group structure of damselfishes. *In Biology of damselfishes*. *CRC Press* 55-83 (2016).
7. Munday, P. L. *et al.* Diversity and flexibility of sex-change strategies in animals. *Trends in Ecology & Evolution* **21**(2), 89–95 (2006).
8. Mondal, R. C. *et al.* Beyond the Binary: Controlling Natural Sex-Change in Hermaphroditic Fishes. *Reviews in Aquaculture* **17**(4), e70066 (2025).
9. Zhang, X. *et al.* A telomere-to-telomere gap-free genome assembly of the protandrous hermaphrodite Asian seabass (*Lates calcarifer*). *Scientific Data* **12**, 1457 (2025).
10. Zhang, K. *et al.* A telomere-to-telomere genome assembly of the protandrous hermaphrodite blackhead seabream, *Acanthopagrus schlegelii*. *Scientific Data* **12**, 350 (2025).
11. Manabe, H. *et al.* Field evidence for bi-directional sex change in the polygynous gobiid fish *Trimma okinawae*. *Journal of Fish Biology* **70**(2), 600–609 (2007).
12. Srinivasan, M. *et al.* Experimental evaluation of the roles of habitat selection and interspecific competition in determining patterns of host use by two anemonefishes. *Marine Ecology Progress Series* **186**, 283–292 (1999).
13. Dunn, D. F. The clownfish sea anemones: Stichodactylidae (Coelenterata: Actiniaria) and other sea anemones symbiotic with pomacentrid fishes. *Transactions of the American Philosophical Society* **71**(1), 3–115 (1981).
14. Fishelson, L. Behaviour, socio-ecology and sexuality in damselfishes (Pomacentridae). *Italian Journal of Zoology* **65**(S1), 387–398 (1998).
15. Mariscal, R. N. *et al.* Field guide to anemonefishes and their Host Sea anemones. *Copeia* **1993**(3), 899 (1993).
16. Mei, L. *et al.* Evaluation of QIAamp® DNA Stool Mini Kit for ecological studies of gut microbiota. *Journal of Microbiological Methods* **54**(1), 13–20 (2003).
17. Chen S. *et al.* fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**(17), i884-i890.
18. Rhoads, A. *et al.* PacBio Sequencing and Its Applications. *Genomics Proteomics & Bioinformatics* **13**, 278–289 (2015).
19. De Coster, W. *et al.* NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**(15), 2666–2669 (2018).

20. Marçais, G. *et al*. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**(6), 764–770 (2011).
21. Vurture, G. W. *et al*. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**(14), 2202–2204 (2017).
22. Cheng, H. *et al*. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 170–175 (2021).
23. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**(9), 2896–2898 (2020).
24. Langmead, B. *et al*. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**(4), 357–359 (2012).
25. Dekker, J. *et al*. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology* **16**, 259 (2015).
26. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems* **3**(1), 95–98 (2016).
27. Dudchenko, O. *et al*. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
28. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Systems* **3**, 99–101 (2016).
29. Xu, M. *et al*. TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *GigaScience* **9**(9), giaa094 (2020).
30. Xu, G. C. *et al*. LR_Gapcloser: a tiling path-based gap closer that uses long reads to complete genome assembly. *Gigascience* **8**(1), giy157 (2019).
31. Takai, A. Karyotype Differentiation characterized by C-banded heterochromatin and nucleolus organizer regions in anemonefish (Pomacentridae). *Cytologia* **86**(2), 175–179 (2021).
32. Lin, Y. *et al*. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Horticulture Research* **10**(8), uhad127 (2023).
33. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**(2), 573–580 (1999).
34. Tarailo-Graovac, M. *et al*. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* Chapter 4, 4–10 (2009).
35. Wang, X. & Wang, L. GMATA: an integrated software package for genome-scale SSR mining, marker development and viewing. *Frontiers in Plant Science* **7**, 1350 (2016).
36. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Science of the United States of America* **117**, 9451–9457 (2020).
37. Xu, Z. *et al*. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* **35**, W265–268 (2007).
38. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **34**, W435–439 (2006).
39. Majoros, W. H. *et al*. TigrScan and GlimmerHMM: two open sources ab initio eukaryotic gene-finders. *Bioinformatics* **20**(16), 2878–2879 (2004).
40. Keilwagen, J. *et al*. GeMoMa: homology-based gene prediction utilizing intron position conservation and RNA-seq data. *Methods in Molecular Biology* **1962**, 161–177 (2019).
41. Haas, B. J. *et al*. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* **8**, 1494–1512 (2013).
42. Haas, B. J. *et al*. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31**(19), 5654–5666 (2003).
43. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology* **9**, R7 (2008).
44. Simao, F. A. *et al*. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
45. Altschul, S. F. *et al*. Basic local alignment search tool. *Journal of Molecular Biology* **215**(3), 403–410 (1990).
46. Bairoch, A. *et al*. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research* **28**(1), 45–48 (2000).
47. Ashburner, M. *et al*. Gene ontology: tool for the unification of biology. *Nature Genetics* **25**(1), 25–29 (2000).
48. Kanehisa, M. *et al*. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* **44**(D1), D457–D462 (2016).
49. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
50. *NCBI Bioproject* https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1306080 (2025).
51. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP617804 (2025).
52. Zhang, X. Genome assembly, predicted coding sequences and functional annotation files of *P. biaculeatus*. *Figshare* https://doi.org/10.6084/m9.figshare.30104977 (2025).
53. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_053813585.1 (2025).
54. Rhie, A. *et al*. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21**(1), 245 (2020).
55. Li, K. *et al*. Identification of errors in draft genome assemblies at single-nucleotide resolution for quality assessment and improvement. *Nature Communications* **14**(1), 6556 (2023).
56. Li, H. *et al*. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**(14), 1754–1760 (2009).
57. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**(18), 3094–3100 (2018).
58. Dobin, A. *et al*. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**(1), 15–21 (2013).
59. Quinlan, A. R. *et al*. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**(6), 841–842 (2010).

Acknowledgements

This work was supported by Shenzhen Natural Science Foundation (no. JCYJ20241202124511016) and Shenzhen Industrial Development Special Fund Project (no. 1303).

Author contributions

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

Competing interests

The authors declare no competing interests.

Additional information

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