scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly and annotation of the spottedtail morwong (*Cheilodactylus zonatus*)

Mengyang Chang[1], Kunpeng Shi[1,2], Wensheng Li[3,4], Wenhui Ma[3,4], Songlin Chen[2,3,5] ✉ & Zhenxia Sha[1,2,3] ✉

The spottedtail morwong (*Cheilodactylus zonatus*), belonging to the family Cheilodactylidae and the genus *Cheilodactylus*, is an important economic fish known for its delicate flesh and rich nutritional value. It mainly feeds on crustaceans and mollusks and is widely distributed in the western Pacific Ocean. However, the lack of effective reference genomes has seriously hindered further research on its biology and genomic breeding. Here, we present a high-quality chromosome-level genome assembly of *C. zonatus* using PacBio sequencing and Hi-C technologies. The assembled genome has a total length of 612.58 Mb with a contig N50 of 25.86 Mb. Approximately 97.8% of the assembly length was anchored to 24 pseudochromosomes. A total of 26,083 protein-coding genes were predicted, among which 23,393 genes (89.67%) were functionally annotated. The completeness of the assembly was estimated to be 98.2% by BUSCO assessment. As we know, this is the first reference genome in the family Cheilodactylidae, which will provide pivotal genomic resources for further biological and evolutionary studies of *C. zonatus*, and facilitate the development of genomic hybridization breeding in the artificial *C. zonatus* industry.

Background & Summary

The family Cheilodactylidae, within the order Perciformes, comprises approximately five genera and around 26 species, with *Cheilodactylus zonatus* representing one of its most characteristic members[1]. However, no reference genome has yet been reported for any member of the family Cheilodactylidae. This species is widely distributed in the western Pacific, from Japan to the South China Sea. The *C. zonatus* has an oval, laterally compressed body, with the dorsal profile rising steeply behind the eyes and reaching its highest point near the origin of the dorsal fin before sloping downward[2]. It inhabits benthic zones around reef-sand interfaces, where it exhibits a stop-and-go swimming pattern. Individuals frequently rest above reef platforms, lie in wait to ambush prey, or use their elongated pectoral fin rays to probe sandy or muddy substrates in search of food, primarily benthic crustaceans[3]. Due to its distinctive taste and texture, *C. zonatus* possesses high economic value and holds promise as a potential candidate for aquaculture development.

In contrast, high-quality reference genomes have been successfully sequenced for several important fish species in recent years, including giant grouper (*Epinephelus lanceolatus*)[4], leopard coral grouper (*Plectropomus leopardus*)[5], and tomato hind (*Cephalopholis sonnerati*)[6]. These studies have provided a solid foundation for the conservation of aquaculture genetic resources, environmental adaptability research, and molecular breeding. With the widespread application of advanced genomic technologies, such as Single Molecule Real-Time

[1]Institute of Aquatic Biotechnology, College of Life Sciences, Qingdao University, Qingdao, 266071, China. [2]Laboratory for Marine Fisheries Science and Food Production Processes, Qingdao National Laboratory for Marine Science and Technology, Qingdao, 266071, China. [3]Shandong Center of Technology Innovation for Biological Breeding of Premium Fish (Preparatory), Qingdao, 266071, China. [4]Laizhou Mingbo Aquatic Co., Ltd, Yantai, 261400, China. [5]State Key Laboratory of Mariculture Biobreeding and Sustainable Goods, Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao, 266071, China. ✉e-mail: chensl@ysfri.ac.cn; shazhenxia@qdu.edu.cn

Libraries	Insert size (bp)	Clean data (Gb)	Reads number	Read length (bp)	Sequence coverage (×)
Illumina reads	350	38.27	127,575,305	150	64.86
PacBio reads	15,000	55.76	3,465,867	17,031 (mean)	94.51

Table 1. Statistical analysis of sequencing reads from Illumina and PacBio.

Type	Data
Raw paired reads	239,294,110
Raw Base (bp)	71,788,233,000
Clean Base (bp)	70,804,734,207
Effective Rate (%)	98.63%
Q20 (%)	98.52
Q30 (%)	94.88
GC Content (%)	45.16

Table 2. Statistical analysis of sequencing data from Hi-C.

sequencing (SMRT) and high-throughput chromatin conformation capture (Hi-C), chromosome-level genome assembly for non-model fish species has now become achievable.

In this study, we employed an integrated strategy of HiFi long reads, Hi-C, Iso-seq and RNA-seq sequencing technologies to assemble a high-quality genome of *C. zonatus*. This genome was 612.58 Mb with contig N50 of 25.86 Mb. Approximately 97.8% (598.6 Mb) of assembled sequences were placed into 24 pseudochromosomes with the support of Hi-C contact map. 26,083 protein-coding genes were predicted and 89.67% were functionally annotated. The Benchmarking Universal Single-Copy Orthologs (BUSCO)[7] assessment of the assembly showed 3,336 (98.2%) BUSCOs was complete. This high-quality *C. zonatus* reference genome will provide an important genomic resource for genetic breeding and molecular mechanism related studies.

Methods

Sample preparation. The sample was collected in September 2024 from Laizhou Mingbo Aquatic Co., Ltd. and derives from a wild population in the South China Sea. The individual was dissected on ice, and then seven tissues (muscle, gills, liver, intestine, kidney, brain, and blood) were excised. After that, these tissues were immediately frozen in liquid nitrogen for nucleic acid extraction. During tissue extraction using scalpels, forceps, and modified scissors, the tissues and sampling tools were washed separately with phosphate-buffered saline (PBS, 1X) and ethanol in order to prevent contamination from external impurities or cross-contamination between different tissue types. The samples were stored at −80 °C. High molecular weight genomic DNA was extracted for Illumina and PacBio Hifi sequencing by TIANamp Genomic DNA Kit (Qiagen, Germany). Total RNA was isolated from all seven tissues using TRIzol reagent (Vazyme, China) for transcriptome sequencing.

Illumina sequencing. The high-quality genomic DNA of muscle tissue was randomly fragmented into segments by the Covaris ultrasonic crusher to prepare a paired-end library with the inserted fragments of 350 bp using the Illumina DNA PCR-Free Prep kit (Illumina, USA). The constructed library was sequenced on Illumina Novaseq X platform (Illumina, USA). To ensure the quality of analysis, the raw reads were processed by fastp (v0.24.1)[8] to filter out adapters and low-quality reads, and 38.27 Gb of clean data was obtained (Table 1).

PacBio HiFi sequencing. The high-quality genomic DNA of muscle tissue was randomly fragmented into 15–17 kb to build a PacBio HiFi library by using SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, USA) according to the manufacturer's instructions. The generated library was sequenced on the Pacbio Sequel II platform (Pacific Biosciences, USA) using SMRT (single molecule real-time) circular consensus sequencing (CCS) technology with SMRT Cell 8 M Tray. After removing adapters in polymerase reads, a total of 55.76 Gb of HiFi clean data was generated, with an average read length of 17,031 bp (Table 1).

Hi-C sequencing. Muscle tissue from the foot of *C. zonatus* was used to construct the Hi-C library according to a standard protocol[9] with several modifications. Briefly, the tissue was ground in liquid nitrogen and cross-linked with a 4% formaldehyde solution to preserve the 3D structure of the DNA. After overnight digestion with the 4-cutter restriction enzyme MboI, the DNA ends were labeled with biotin-14-dCTP. The subsequent steps involved blunt-end ligation of the cross-linked fragments, re-ligation of the proximal chromatin DNA, and reverse crosslinking of the nuclear complexes. DNA was then purified by using phenol-chloroform extraction. After removing biotin from non-ligated fragment ends and repairing of sheared fragment ends, biotin-labeled Hi-C samples were enriched, ligated with sequencing adapters, and amplified by polymerase chain reaction (PCR, 12–14 cycles). Hi-C library preparation and sequencing were performed with muscle tissue using a Dovetail Hi-C Core Kit (Dovetail Genomics, USA) following protocol instructions. The qualified library was sequenced on the Illumina Hiseq platform. After adapter removal and filtering of low-quality reads using Trim Galore (v0.6.10, https://github.com/FelixKrueger/TrimGalore), 70.8 Gb of clean data were obtained, with a Q20 score of 98.52% (Table 2).

Sample	Platform	Raw Reads	Clean Reads	Q20 (%)	Q30 (%)	GC (%)
Muscle	Illumina	44,335,212	42,803,966	96.95	91.58	35.06
Gills	Illumina	41,498,594	40,662,494	97.34	92.2	34.3
Liver	Illumina	45,804,886	44,959,106	97.72	93.25	36.73
Intestine	Illumina	44,078,548	42,905,486	97.28	92.44	36.17
Kidney	Illumina	45,450,476	43,895,912	97.38	92.66	35.6
Brain	Illumina	42,146,312	40,323,936	97.4	92.75	37.09
Mix	PacBio	10,171,245				

Table 3. Statistical analysis of RNA-seq from different tissues.

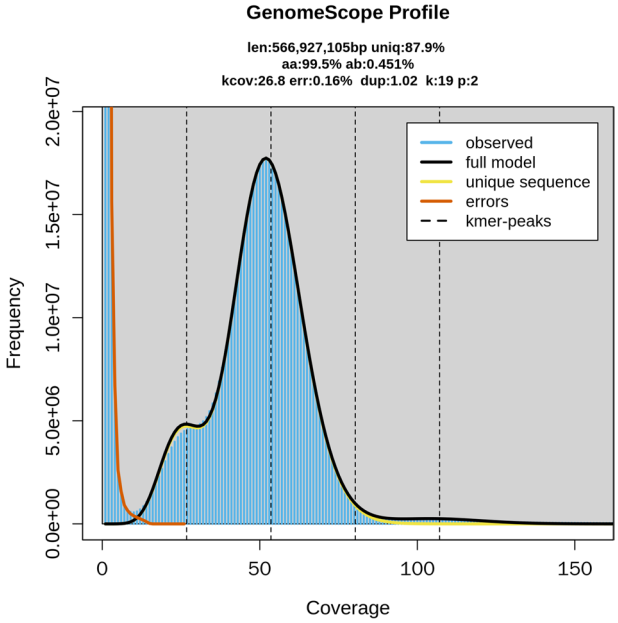

Fig. 1 Genome survey at 21-mer of *C. zonatus*.

RNA-seq sequencing. For transcriptome sequencing, muscle, gills, liver, intestine, heart, skin, and and whole blood of *C. zonatus* were collected to construct sequencing libraries. Next-generation RNA-seq libraries were constructed with the NEBNext® UltraTM RNA Library Prep Kit for Illumina (NEB, USA) according to the manufacturer's suggestions and sequenced on the Illumina Novaseq X platform (Illumina, USA) for PE150. The third-generation full-length RNA sequencing library was constructed by the Kinnex PCR cDNA Synthesis Kit (Pacific Biosciences, USA) and sequenced on PacBio Sequel II platform (Pacific Biosciences, USA). A total of 17.71 Gb of Isoform-Sequencing (Iso-Seq) reads was obtained (Table 3).

Genome survey. Genomic surveys were conducted to assess the genomic characteristics, which includes genome size, heterozygosity, and replication rates, so as to determine the amount of subsequent sequencing data. The software Jellyfish (v 2.2.10)[10] was utilized with parameters set as '-m 19 -s 20 G -C' to count k-19mers using the Illumina-generated short reads as input. The generated histogram was then employed as an input file for GenomeScope (v2.0)[11] to estimate the genome characteristics. It was found that the genome size and heterozygosity of *C. zonatus* were approximately 566.92 Mb and 0.45%, respectively (Fig. 1).

Genome assembly. Based on PacBio HiFi and Hi-C reads, the genome was *de novo* assembled by Hifiasm (v0.25.0)[12] in 'Hi-C Intergrated Assembly' mode with default parameters. Then, purge_dups (v1.2.6)[13] was used to remove haplotypic and heterozygous duplication from the de novo assembly, resulting in a total length of 612.58 Mb. To generate the chromosome-level genome of *C. zonatus*, the 121.66 × Hi-C data obtained from sequencing, the contigs sequences obtained from the assembly were anchored to the chromosome level using Haphic (v1.0.7)[14] (Fig. 2). The final *C. zonatus* genome size was 612.58 Mb, and a total of 598.6 Mb (97.8%) sequences were anchored to 24 chromosomes, of which the longest and shortest were 31.22 Mb and 13.63 Mb, respectively (Table 4, Fig. 3).

Annotation of repetitive elements. A comprehensive repeat library for the *C. zonatus* genome was constructed by combining Extensive *de novo* TE Annotator (EDTA, v2.2.2)[15] with RepeatModeler (v2.0.6)[16]. Based on this *de novo* repeat library, repetitive elements were predicted and classified by RepeatMasker (v4.1.8)[17]. The repeat annotation results showed that repetitive elements accounted for 33.72% of the assembled genome. Among

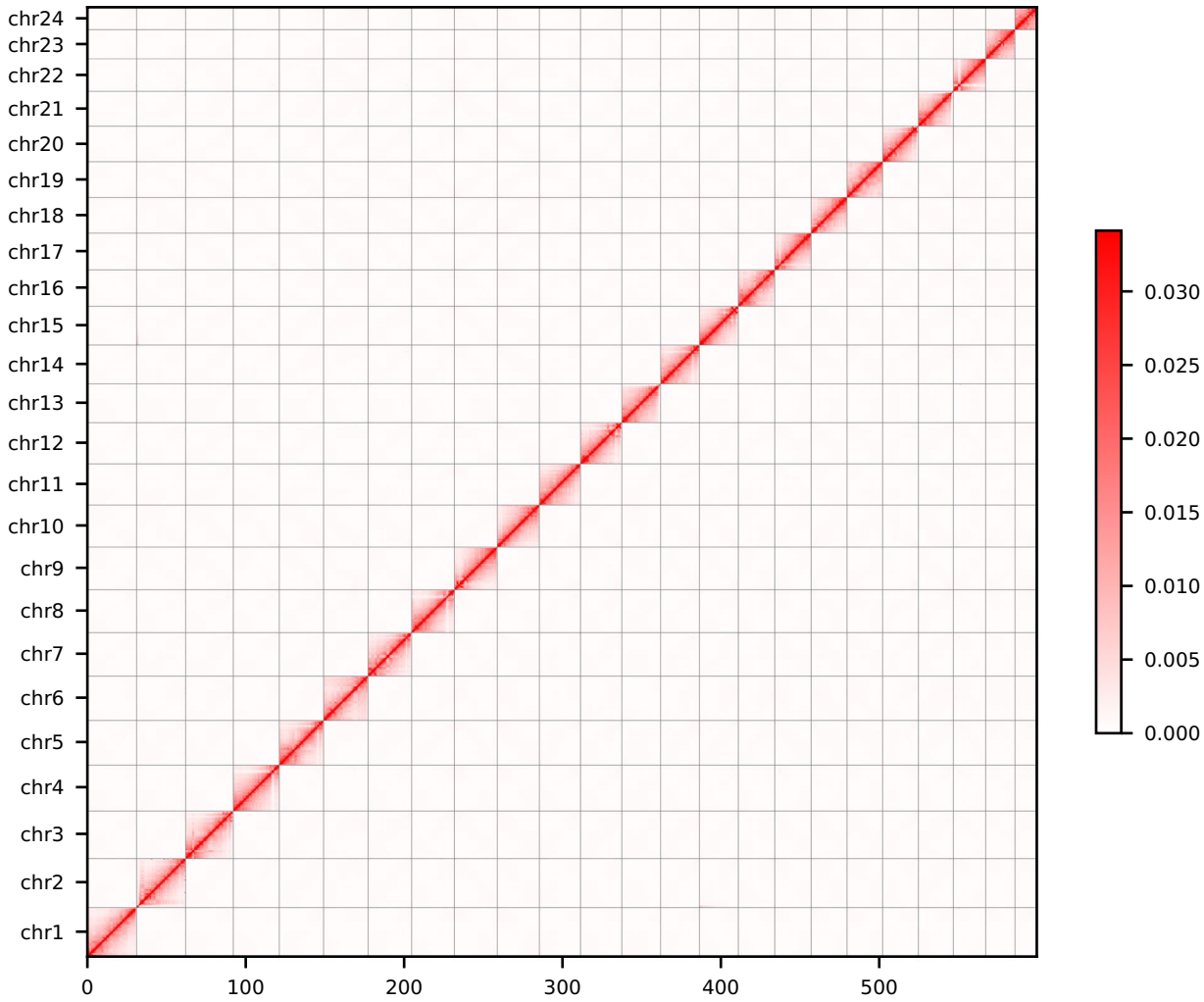

Fig. 2 Genome-wide heatmap of Hi-C interactions among 24 chromosomes in *C. zonatus*. Various colors represent different interaction frequencies of Hi-C links, and the aggregated color blocks represent the interaction frequencies between individual chromosomes.

Statistic	Contig (bp)	Scaffold (bp)
Total Number	111	86
Total Length	612,588,814	612,588,014
Average Length	5,518,810	5,947,464
Max Length	31,225,458	31,225,458
N50 Length	25,860,851	25,860,851
N50 Number	11	11
N90 Length	20,406,750	20,406,750
N90 Number	22	22

Table 4. Assembly statistics of *C. zonatus* genome.

these, DNA transposons were the most abundant (13.13%), followed by long interspersed nuclear elements (LINEs, 6.17%) and unclassified repeats (10.71%). Other types included long terminal repeats (LTRs, 1.63%), Penelope elements (0.17%), short interspersed nuclear elements (SINEs, 0.61%), and DIRS elements (0.62%). Additionally, tandem repeats such as satellites (0.43%), simple repeats (0.14%), and low-complexity regions (0.11%) were also identified (Table 5, Fig. 3).

Noncoding RNA (ncRNA) annotation. Ribosomal RNAs (rRNAs) were predicted by using Barrnap (v0.9, https://github.com/tseemann/barrnap) with default parameters. For the prediction of transfer RNAs (tRNAs), tRNAscan-SE[18] was utilized. Small nuclear RNAs (snRNAs) and microRNAs (miRNAs) were predicted by

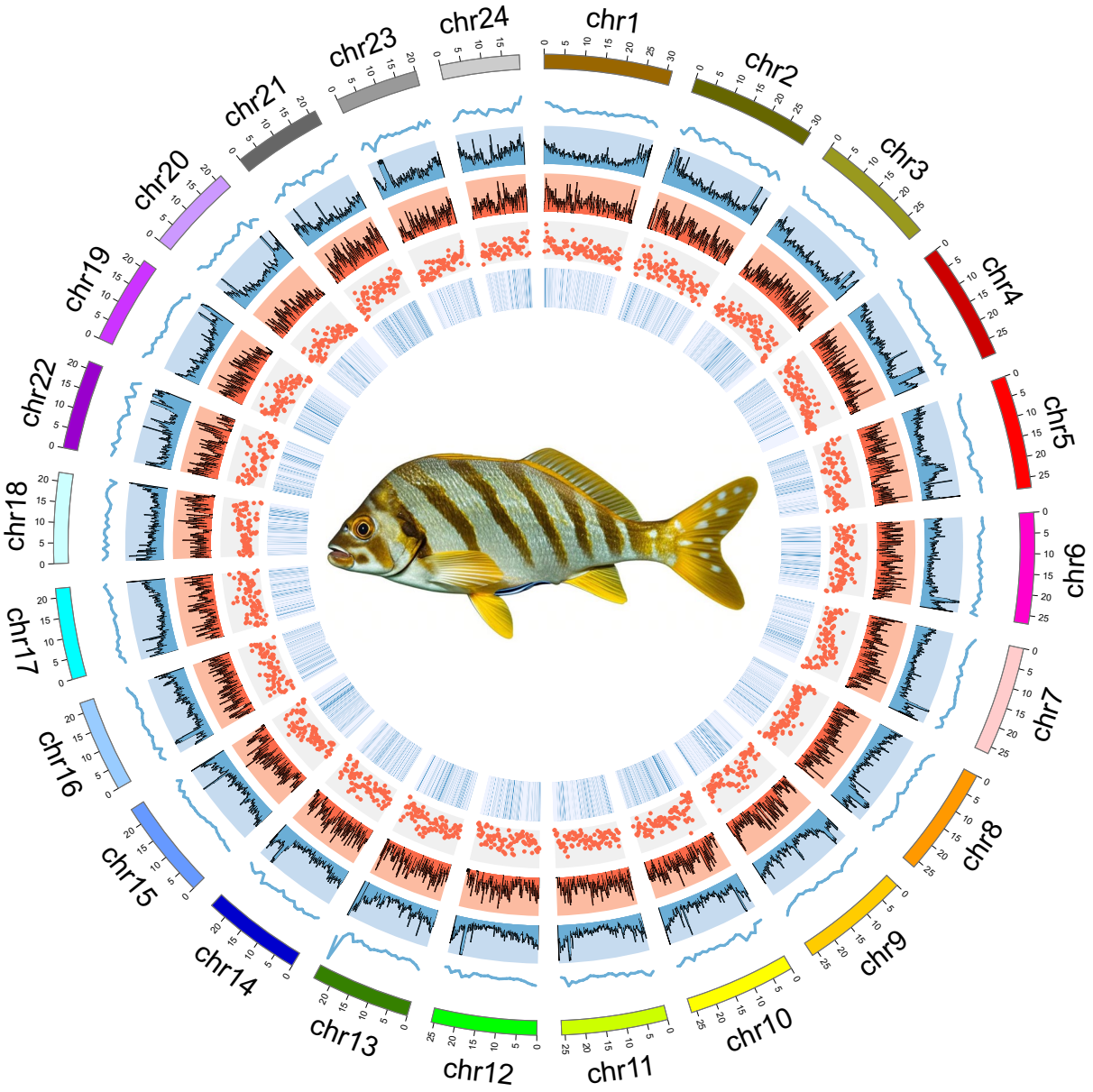

Fig. 3 Circos plot of genomic features in *C. zonatus*. Each track from outer to inner represents the chromosome length (Mb), and the distribution of GC content, DNA element, short interspersed nuclear element (SINE) density, long interspersed nuclear element (LINE) density, and protein-coding genes with a sliding window size of 1 Mb.

Infernal (v1.1.4)[19] based on genome alignment with the Rfam database (v14.8)[20]. In total, 2,041 non-coding RNAs (ncRNAs) were annotated, which included 342 rRNAs, 1,013 tRNAs, 128 snRNAs, 9 long non-coding RNAs (lncRNAs), and 549 miRNAs (Table 6).

Prediction and functional annotation of protein-coding genes. Protein-coding genes structure prediction was performed by integrating de novo prediction, homology-based prediction, transcriptome-based prediction, and prediction supported by other evidence in repeat-masked genome (Fig. 4). *De novo* prediction was used to predict gene structure (e.g. codon frequency, exon-intron distribution) using software Augustus (v3.3.3)[21], GlimmerHMM (v3.04)[22], GeneMark (v4.71_lic)[23], SNAP (v2006-07-28)[24] and BRAKER2 (v2.1.8)[25]. For homology prediction, reference protein-coding gene sets of six closely related species, including *Cynoglossus semilaevis* (GCA_000523025.1), *Epinephelus lanceolatus* (GCA_005281545.1), *Larimichthys crocea* (GCA_000972845.2), *Paralichthys olivaceus* (GCA_024713975.2), *Plectropomus leopardus* (GCA_008729295.2), and *Cephalopholis sonnerati* (GCA_043388385.1), were downloaded from NCBI databases to generate a homology-based protein-coding gene set. The homology-based coding protein sequence sets were aligned with the genomes and predicted with gene structure information using TblastN (v2.7.1)[26], GeneWise (v2.4.1)[27], and GenomeThreader

Type			Count	Length (bp)	% of Genome
Dispersed repeats	DNA transposons		760860	78619798	13.13
	Retroelements	DIRS	49652	3696398	0.62
		LTR	110997	9738616	1.63
		LINE	339366	36929206	6.17
		Penelope	12388	1042904	0.17
		SINE	34362	3627680	0.61
	Unclassified		619635	64114833	10.71
Tandem repeats	Simple repeats		11519	816219	0.14
	Low complexity		11407	601427	0.11
	Satellite		17133	2590956	0.43
Total			1967319	201778037	33.72

Table 5. Classification of repetitive sequences in *C. zonatus* genome.

Type		Copy Number	Average Length (bp)	Total Length (bp)	% of Genome
tRNA		1013	73.08	74030	0.12371219
miRNA		549	80.71	44308	0.07395285
lncRNA		9	1200	10800	0.00001805
snRNA	CD-box	29	120	3480	0.00000581
	HACA-box	17	230	3910	0.00000653
	scaRNA	1	150	150	0.00000025
	splicing	81	160	12960	0.00002165
rRNA	5S	327	115	37605	0.06282348
	5.8S	5	160	800	0.00000134
	18S	5	1700	8500	0.0000142
	28S	5	4250	21250	0.00003547

Table 6. Classification of ncRNAs in *C. zonatus* genome.

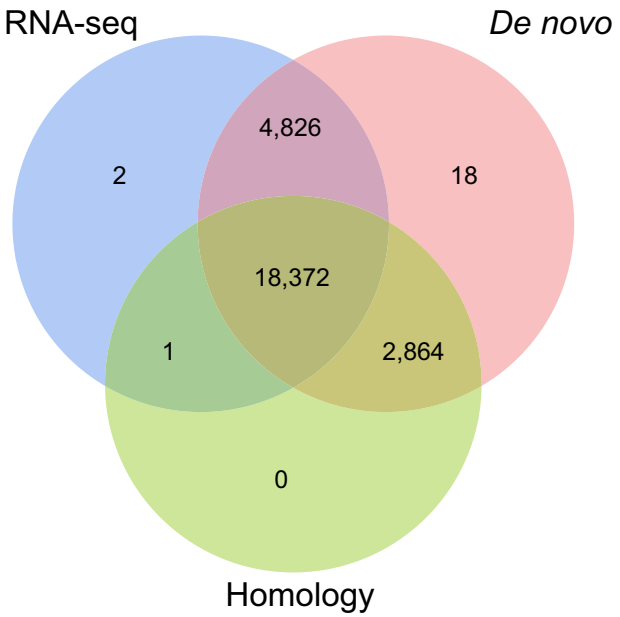

Fig. 4 Venn diagram of *C. zonatus* gene structure prediction through three strategies.

(v1.7.3)[28]. Other predictions supported by evidence are the prediction of gene structure by Blat alignment using EST or cDNA data from homologous species. In transcriptome-based prediction, RNA-seq data from different tissues were compared to the genome using HISAT2 (v2.2.1)[29] and Samtools (v1.19)[30], identifying exonic regions and splice sites using default parameters. Transcript assembly was performed using StringTie (v2.2.1)[31] against

Method	Software	Species	Gene Number	Gene length (bp)	CDS length (bp)	Average intron length (bp)	Average exon length (bp)	Exon per gene
De novo	Augustus		29,184	10,087.35	1,489.70	223.45	1,498.87	6.74
	BRAKER2		32,165	10,181.22	1,478.66	215.84	2,265.39	6.98
	GlimmerHMM		84,963	7,905.64	741.22	206.11	2,933.04	3.62
	GeneMark		64,508	4,358.92	829.46	200.25	1,099.33	4.28
	SNAP		77,842	3,057.49	538.19	231.74	1,918.21	3.35
RNA-seq	HISAT2 & StringTie		20,103	20,738.41	2,226.98	454.91	2,626.65	9.61
	PASA		29,996	21,091.46	1,217.94	413.28	2,884.11	7.33
Homology	MetaEuk	*C. semilaevis*	24,161	6,705.35	1,046.70	274.45	1,614.45	7.4
		E. lanceolatus	23,639	7,730.76	1,174.61	232.39	1,576.56	11.15
		L. crocea	24,309	6,519.86	981.48	224.53	1,600.26	9.45
		P. olivaceus	22,207	6,971.27	1,086.48	239.7	1,628.24	12.6
		P. leopardus	23,892	3,771.01	809.98	289.93	1,619.86	11.82
Final		—	26,083	15,094.75	1,506.73	252.77	1,989.42	7.58

Table 7. Statistical results of the gene structure annotation in *C. zonatus* genome.

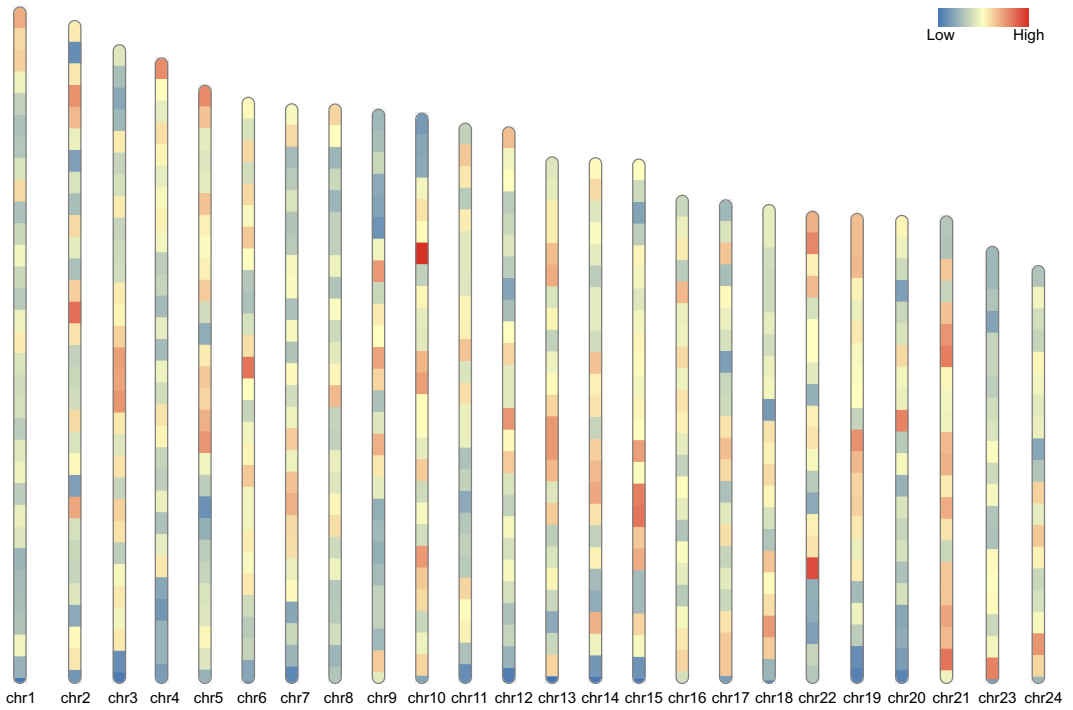

Fig. 5 Gene density of *C. zonatus* each chromosome within 1 Mb sliding windows.

the results. Combining the above prediction results, the gene sets obtained from the various methods of prediction were integrated into a non-redundant and more complete gene set using the EVidenceModeler (EVM, v2.1.0)[32] integration software. Lastly, the final non-redundant reference gene set was obtained by using PASA (v2.5.3)[33], combining the transcriptome assembly results, correcting the annotation results of EVM, and adding information such as UTR and variable shear to obtain the final non-redundant reference gene set. Finally, 26,083 protein-coding genes in the genome of *C. zonatus* were predicted, with an average gene length of 15,094.75 bp. The average lengths of CDSs, exons, and introns of each gene were 1,506.73, 1,989.42 bp, and 252.77 bp, respectively (Table 7). To visualize gene distribution, we plotted the density of genes on 24 chromosomes with a window of 1 Mb in length using RIdeogram (v0.2.2) in R (Fig. 5). Meanwhile, the genomic features of *C. zonatus*, such as CDS, gene, exon and intron length distributions, were similar to those of homologous species (Fig. 6).

Functional annotation of the protein-coding genes was carried out using the Funannotate pipeline (v1.8.16), based on multiple databases, including Clusters of Orthologous Groups of Proteins (COG)[34], eggNOG[35], Gene Ontology (GO)[36], InterPro[37], and Pfam[38]. Subsequently, functional annotation was performed by aligning the protein-coding genes to the National Center for Biotechnology Information (NCBI) non-redundant protein (Nr) and Swiss-Prot databases[39] using Diamond (v2.1.8.162)[40], with an E-value cutoff of $1e^{-5}$. Additionally, Kyoto Encyclopedia of Genes and Genomes (KEGG) annotations were obtained by eggNOG-mapper (v2.1.12)[41]. A

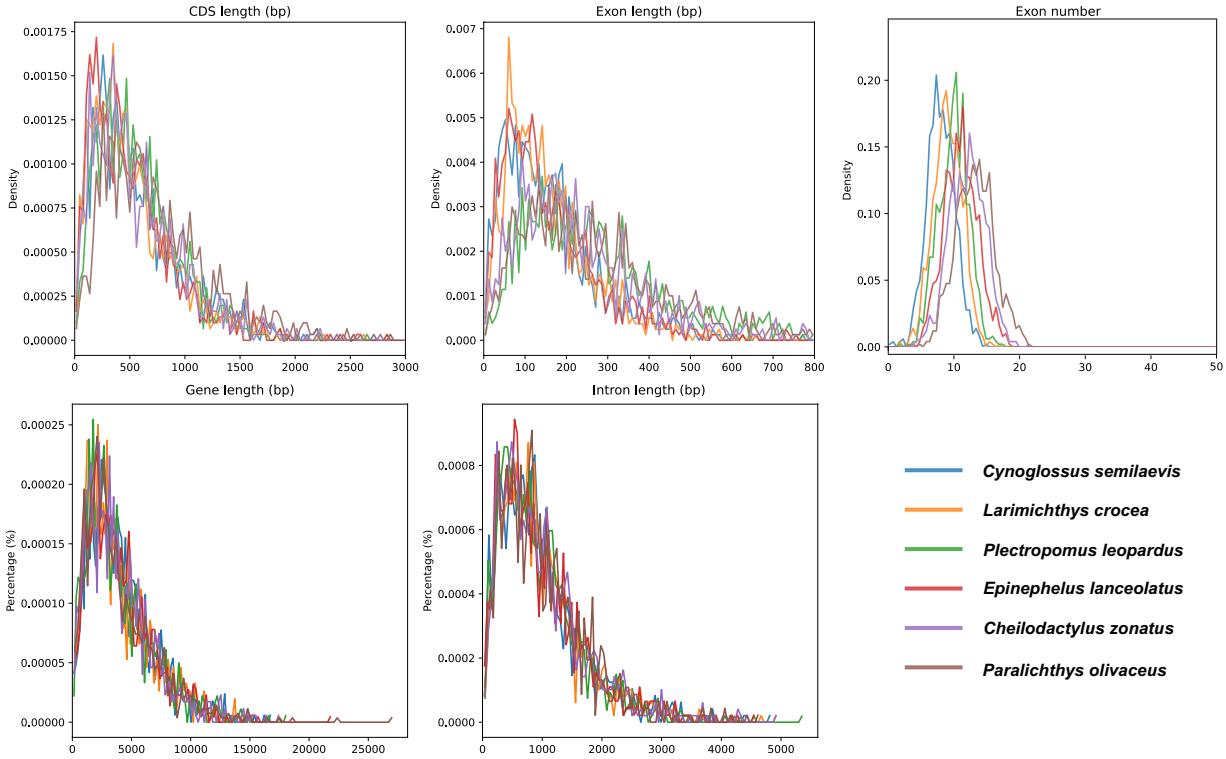

Fig. 6 The Gene length, CDS length, intron length, exon length, exon number per gene in *C. zonatus* compared with five other species.

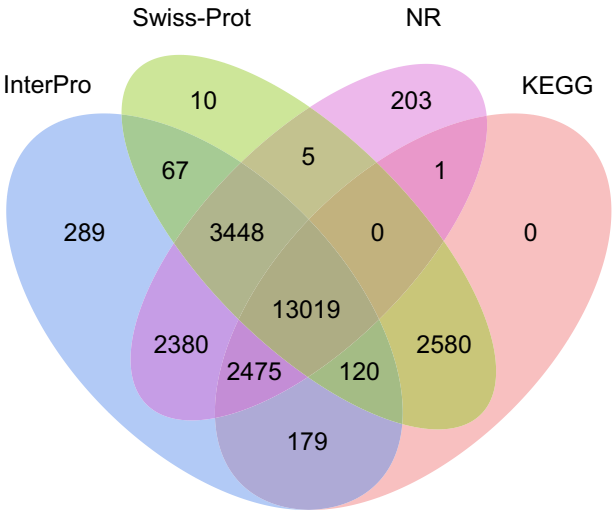

Fig. 7 Venn diagram of *C. zonatus* functional gene annotation based on four of the eight databases.

total of 23,393 genes (89.6%) of the predicted proteins were functionally annotated, of which 17,441 proteins were annotated in the Swiss-Prot database, 22,758 proteins in the NR database, 19,221 proteins in the InterPro database, and 16,221 proteins in the Pfam database (Fig. 7; Table 8).

Data Records

The genomic Illumina, PacBio, Hi-C, and transcriptomic sequencing data were deposited in the Sequence Read Archive (SRA) at NCBI under the accession numbers SRP583805[42].

The final chromosome assembly was deposited in the GenBank under the accession GCA_050575295.2[43].

The annotation files have been uploaded to the figshare database, with the DOI number: https://doi.org/10.6084/m9.figshare.28953329[44].

Database	Number	Percent(%)
Total	26,083	100
COG	18,035	69.21
eggNOG	18,982	72.86
GO	16,004	61.44
InterPro	19,211	73.78
KEGG	14,157	54.32
Pfam	16,221	62.18
Swiss-Prot	17,441	66.94
Nr	22,758	87.27
Annotated	23,393	89.88
Unannotated	2,690	10.12

Table 8. Summary of the functional gene annotation in *C. zonatus* genome.

	Type		Result
Genome accuracy and completeness	Mapping short-reads rate		99.24%
	Genome coverage rate with short-reads		99.97%
	Quality value scores (QVs)		54.73
	BUSCOs	Complete BUSCOs (C)	98.2% (3,574)
		Complete and single-copy BUSCOs (S)	97.5% (3,549)
		Complete and duplicated BUSCOs (D)	1.4% (51)
		Fragmented BUSCOs (F)	0.4% (15)
		Missing BUSCOs (M)	0.7% (25)
	CRAQ	R-AQI	94.72%
		S-AQI	99.28%
Annotation quality	Complete BUSCOs (C)		95.9% (3,491)
	Complete and single-copy BUSCOs (S)		95.2% (3,465)
	Complete and duplicated BUSCOs (D)		1.2% (44)
	Fragmented BUSCOs (F)		0.8% (29)
	Missing BUSCOs (M)		3.4% (124)

Table 9. Assessment metrics of *C. zonatus* genome assembly and annotation.

Technical Validation

Quality evaluation of the extracted nucleic acid. The concentrations of DNA and RNA were measured by Nanodrop 2000 spectrophotometers (Thermo Fisher Scientific, USA) and 5400 Fragment Analyzer system (Agilent Technologies, USA). Agarose gel electrophoresis was applied to evaluate the integrity and purity of nucleic acid.

Evaluation of the genome assembly and annotation. The present study utilized a total of 94.51× PacBio HiFi reads and 121.66× Hi-C reads to ensure a high-quality genome assembly. Prior to this, we conducted a 19-mer distribution analysis to estimate the genome size based on 64.86× Illumina reads. To assess the quality of *C. zonatus* genome assembly, we adopted three methods as follows. Firstly, we used Merqury (v1.3)[45] to estimate the base-level accuracy and completeness based on k-mer counts generated from Illumina reads, resulting in a QV of 54.73 (Table 9). Secondly, Clipping Reveals Assembly Quality (CRAQ, v1.09)[46] was used to assess the accuracy of genome assembly based on PacBio HiFi reads and Illumina reads, resulting in a R-AQI (assembly quality indicator) of 94.72% and a S-AQI of 99.28% (Table 9). Thirdly, we mapped Illumina clean reads to *C. zonatus* genome using bwa (v0.7.17-r1188)[30]. The statistical result from samtools (v1.9)[47] showed the genome mapping rate and the coverage rate were 99.24% and 99.97%, respectively (Table 9).

The completeness of the *C. zonatus* genome assembly was evaluated using BUSCO (v5.2.2)[48] with the *actinopterygii_odb10* database comprising 3,640 single-copy orthologs. The assessment revealed that 98.2% (3,574) of BUSCOs were complete, including 97.5% (3,549) single-copy and 1.4% (51) duplicated BUSCOs. Additionally, 0.4% (15) were fragmented and only 0.7% (25) were missing, indicating that the assembled genome has high completeness and integrity. These results collectively revealed the high quality of *C. zonatus* genome assembly. The accuracy of gene annotation was evaluated using BUSCO (v5.2.2) on the basis of actinopterygii_odb10 database containing 3,640 BUSCOs. The results showed that 3,491 (95.9%) complete BUSCOs, containing 3,465 (95.2%) single-copy and 44 (1.2%) duplicated BUSCOs, were detected, 29 (0.8%) fragmented BUSCOs and 124 (3.4%) missing BUSCOs were identified.

Code availability

The detailed parameters were not specified in this study. All bioinformatics software and tools were employed strictly following their respective user manuals and standard protocols. No custom codes or scripts were developed or applied during the course of this study.

Received: 30 May 2025; Accepted: 23 September 2025;
Published online: 04 November 2025

References

1. Burridge, C. P. & Smolenski, A. J. Molecular phylogeny of the Cheilodactylidae and Latridae (Perciformes: Cirrhitoidea) with notes on taxonomy and biogeography. *Mol Phylogenet Evol* **30**, 118–127, https://doi.org/10.1016/s1055-7903(03)00157-x (2004).
2. Burridge, C. P. & Smolenski, A. J. Microsatellite loci from the marine fish Nemadactylus macropterus (Perciformes: Cheilodactylidae). *Mol Ecol* **9**, 1180–1181, https://doi.org/10.1046/j.1365-294x.2000.00954-6.x (2000).
3. Burridge, C. P. Molecular phylogeny of Nemadactylus and acantholatris (Perciformes: Cirrhitoidea: Cheilodactylidae), with implications for taxonomy and biogeography. *Mol Phylogenet Evol* **13**, 93–109, https://doi.org/10.1006/mpev.1999.0622 (1999).
4. Zhou, Q. *et al*. Telomere-to-telomere gapless genome assembly of the giant grouper (*Epinephelus lanceolatus*). *Sci Data* **11**, 1342, https://doi.org/10.1038/s41597-024-04219-7 (2024).
5. Zhou, Q. *et al*. De novo sequencing and chromosomal-scale genome assembly of leopard coral grouper, *Plectropomus leopardus*. *Mol Ecol Resour* **20**, 1403–1413, https://doi.org/10.1111/1755-0998.13207 (2020).
6. Xie, Z. *et al*. Chromosome-level genome assembly and transcriptome comparison analysis of *Cephalopholis sonnerati* and its related Ggrouper species. *Biology (Basel)* **11**, 1053, https://doi.org/10.3390/biology11071053 (2022).
7. 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, https://doi.org/10.1093/bioinformatics/btv351 (2015).
8. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890, https://doi.org/10.1093/bioinformatics/bty560 (2018).
9. Belton, J. M. *et al*. Hi-C: a comprehensive technique to capture the conformation of genomes. *Methods* **58**, 268–276, https://doi.org/10.1016/j.ymeth.2012.05.001 (2012).
10. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770, https://doi.org/10.1093/bioinformatics/btr011 (2011).
11. 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, https://doi.org/10.1038/s41467-020-14998-3 (2020).
12. 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, https://doi.org/10.1038/s41592-020-01056-5 (2021).
13. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898, https://doi.org/10.1093/bioinformatics/btaa025 (2020).
14. Zeng, X. *et al*. Chromosome-level scaffolding of haplotype-resolved assemblies using Hi-C data without reference genomes. *Nat Plants* **10**, 1184–1200, https://doi.org/10.1038/s41477-024-01755-3 (2024).
15. Ou, S. *et al*. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol* **20**, 275, https://doi.org/10.1186/s13059-019-1905-y (2019).
16. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA* **117**, 9451–9457, https://doi.org/10.1073/pnas.1921046117 (2020).
17. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr Protoc Bioinformatics* **4**, 4.10.1–4.10.14, https://doi.org/10.1002/0471250953.bi0410s25 (2009).
18. Chan, P. P., Lin, B. Y., Mak, A. J. & Lowe, T. M. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Res* **49**, 9077–9096, https://doi.org/10.1093/nar/gkab688 (2021).
19. Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335–1337, https://doi.org/10.1093/bioinformatics/btp157 (2009).
20. Kalvari, I. *et al*. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res* **49**, D192–D200, https://doi.org/10.1093/nar/gkaa1047 (2021).
21. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34**, W435–W439, https://doi.org/10.1093/nar/gkl200 (2006).
22. Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* **20**, 2878–2879, https://doi.org/10.1093/bioinformatics/bth315 (2004).
23. Besemer, J. & Borodovsky, M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. *Nucleic Acids Res* **33**, W451–454, https://doi.org/10.1093/nar/gki487 (2005).
24. Korf, I. Gene finding in novel genomes. *BMC Bioinf* **5**, 59, https://doi.org/10.1186/1471-2105-5-59 (2004).
25. Bruna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics Bioinf* **3**, lqaa108, https://doi.org/10.1093/nargab/lqaa108 (2021).
26. Gertz, E. M., Yu, Y. K., Agarwala, R., Schäffer, A. A. & Altschul, S. F. Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. *BMC Biol* **4**, 41, https://doi.org/10.1186/1741-7007-4-41 (2006).
27. Baranzini, S. E. *et al*. Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum Mol Genet* **18**, 2078–2090, https://doi.org/10.1093/hmg/ddp120 (2009).
28. Pina-Martins, F., Silva, D. N., Fino, J. & Paulo, O. S. Structure_threader: An improved method for automation and parallelization of programs structure, fastStructure and MavericK on multicore CPU systems. *Mol Ecol Resour* **17**, e268–e274, https://doi.org/10.1111/1755-0998.12702 (2017).
29. 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, https://doi.org/10.1038/s41587-019-0201-4 (2019).
30. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760, https://doi.org/10.1093/bioinformatics/btp324 (2009).
31. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**, 290–295, https://doi.org/10.1038/nbt.3122 (2015).
32. Haas, B. J. *et al*. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol* **9**, R7, https://doi.org/10.1186/gb-2008-9-1-r7 (2008).
33. Avram, O. *et al*. PASA: Proteomic analysis of serum antibodies web server. *PLoS Comput Biol* **17**, e1008607, https://doi.org/10.1371/journal.pcbi.1008607 (2021).
34. Galperin, M. Y. *et al*. COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res* **49**, D274–D281, https://doi.org/10.1093/nar/gkaa1018 (2021).

35. Huerta-Cepas, J. *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090organisms and 2502 viruses. *Nucleic Acids Res* **47**, D309–D314, https://doi.org/10.1093/nar/gky1085 (2019).
36. Ashburner, M. *et al*. Gene Ontology: tool for the unification of biology. *Nat. Genet* **25**, 25–29, https://doi.org/10.1038/75556 (2000).
37. Blum, M. *et al*. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res* **49**, D344–D354, https://doi.org/10.1093/nar/gkaa977 (2021).
38. Mistry, J. *et al*. Pfam: the protein families database in 2021. *Nucleic Acids Res* **49**, D412–D419, https://doi.org/10.1093/nar/gkaa913 (2021).
39. Bairoch, A. & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* **28**, 45–48, https://doi.org/10.1093/nar/28.1.45 (2000).
40. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59–60, https://doi.org/10.1038/nmeth.3176 (2014).
41. Cantalapiedra, C. P., Hernandez-Plaza, A., Letunic, I., Bork, P. & Huerta-Cepas, J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol. Biol. Evol* **38**, 5825–5829, https://doi.org/10.1093/molbev/msab293 (2021).
42. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP583805 (2025).
43. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_050575295.2 (2025).
44. Chang, M. The gene annotation file of *Cheilodactylus zonatus*. *figshare. Dataset.* https://doi.org/10.6084/m9.figshare.28953329 (2025).
45. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biol* **21**, 245, https://doi.org/10.1186/s13059-020-02134-9 (2020).
46. Li, K. P., Xu, P., Wang, J. P., Yi, X. & Jiao, Y. N. Identification of errors in draft genome assemblies at single-nucleotide resolution for quality assessment and improvement. *Nat. Commun* **14**, 6556, https://doi.org/10.1038/s41467-023-42336-w (2023).
47. Li, H. *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079, https://doi.org/10.1093/bioinformatics/btp352 (2009).
48. Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A. & Zdobnov, E. M. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. *Molecular Biology and Evolution* **38**, 4647–4654 (2021).

Acknowledgements

This study was supported by grants from the Academician Special Program of Shandong Province (2023ZLYS02).

Author contributions

Z.S. and M.C. conceived the research project. M.C. performed the bioinformatic analyses and wrote the manuscript. M.C. and K.S. accomplished the collection of samples.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.C. or Z.S.

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.