scientific data

OPEN

DATA DESCRIPTOR

Chromosome-level genome assembly of the clam, Xishi tongue *Coelomactra antiquata*

Yawen Shen[1], Yanlin Wang[1] & Lingfeng Kong[1,2] ✉

Xishi tongue (*Coelomactra antiquata*), a commercially valuable marine bivalve, is distributed along the coastal waters of East Asia. In China, significant morphological and genetic differences have been observed between northern and southern populations. Overfishing and pollution have caused a severe decline in its natural populations, rendering the species endangered. In this study, we constructed the first chromosome-level genome of *C. antiquata* based on PacBio HiFi and Hi-C sequencing data. The assembled genome was 791.83 Mb in size, with the scaffold N50 of 44.05 Mb, and 99.79% of the sequences (790.13 Mb) were anchored to 19 chromosomes. A total of 24,592 protein-coding genes were predicted in the final assembly, of which 89.88% were functionally annotated. The BUSCO analysis revealed a genome completeness of 97.69%. The high-quality genome serves as a critical resource for advancing research on population genetics and germplasm conservation of this commercial shellfish, thereby facilitating sustainable management and conservation efforts.

Background & Summary

Coelomactra antiquata (Veneroida: Mactridae), a marine bivalve calm distributed in the coastal waters of East Asia, has high economic value[1]. This benthic clam inhabits the sandy or muddy substrate from near the low tide line to a depth of about 10 meter, and usually buries at depths of 6-7 centimeters[2,3]. It is widely distributed along the coast of China, from Liaoning Province in the north southward to Guangxi Province, with particularly high abundance in Fujian Province[4]. *C. antiquata*, as a delicious and nutritious seafood, is very popular in Chinese fish markets[5]. Reports have indicated that commercial-sized *C. antiquata* is being sold in southern China markets for $15–20 per kilogram[4]. In terms of size, the shell length of adult individuals can reach 15 cm[4,5]. And their delicate meat, enriched with flavor-enhancing amino acids due to its high protein content, offers a naturally sweet taste[5,6]. In addition, the soft part of *C. antiquata* contains a high proportion of unsaturated fatty acids and mineral elements such as calcium and iron, which contribute to its exceptional nutritional value[5,7].

In the 1980s and 1990s, the resources of *C. antiquata* were abundant in China, with estimated annual harvests reaching nearly 10,000 tons[4,8]. However, due to overfishing, the death of seedlings and the habitat destruction caused by the application of fishing techniques such as electric trawling, as well as pollution, the natural resources of *C. antiquata* have declined dramatically[1,4,9,10]. In the China Species Red List of Threatened Species published in 2004, *C. antiquata* was listed as 'endangered', highlighting the urgent need for conversation[1,11,12]. From 2005 to 2008, surveys of fishery resources in China showed that only three coastal provinces, Shandong, Jiangsu and Fujian, had found *C. antiquata* populations[4,13]. In order to protect and recover the natural resources of *C. antiquata*, the exploration on artificial breeding technology has been carried out[5,14]. Although the artificial breeding has achieved a successful breakthrough, large-scale culture has not been reported up to now[15]. We believe that genome information of *C. antiquata* will be useful in further research on conservation genomics as well as genomic selection in breeding to improve economically important traits in this commercially valuable shellfish.

It is worth mentioning that recent studies on population genetics of *C. antiquata* based on different kinds of genetic markers have revealed significant genetic differences between northern and southern populations in China, as follows. The analysis of polymorphic allozyme loci and morphological variables showed that there was significant morphological and genetic differentiation between the northern populations, including Rizhao

[1]Key Laboratory of Mariculture, Ministry of Education, Ocean University of China, Qingdao, 266003, China. [2]Laboratory for Marine Fisheries Science and Food Production Processes, Qingdao Marine Science and Technology Center, Qingdao, Shandong, 266237, China. ✉e-mail: klfaly@ouc.edu.cn

Libraries	Insert size (bp)	Clean data (Gb)	Reads number	Read length (bp)	Sequence coverage (X)
Illumina reads	350	65.30	441,061,468	150	82.47
PacBio reads	15,000	37.10	2,178,181	17,031 (mean)	46.85
Total	—	102.40	443,239,649	—	129.32

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

in Shandong Province, Jimo in Shandong Province, Qidong in Jiangsu Province, and the southern population represented by Changle in Fujian Province[4]. Studies based on AFLP (amplified fragment length polymorphism) showed that there was significant genetic differentiation between the northern population represented by Jiaonan in Shandong Province and the southern population represented by Changle and Xiamen in Fujian Province, while there was no significant genetic differentiation between the two southern populations[16]. Given the significant role of cytochrome c oxidase I (COI) as a marker in species identification[17], a COI-based systematic differentiation study reported an unusually high level of variation (14.9%) within the COI gene of *C. antiquata*[12]. Based on the publication of mitochondrial genome data of *C. antiquata*[18,19], phylogenetic analysis of 12 protein-coding genes in the mitochondrial genome showed that the genetic distance between the northern population (represented by Rizhao in Shandong Province) and the southern population (represented by Zhangzhou in Fujian Province) was greater than that between the two species from the same genus[20]. These findings, combined with significant morphological differences between the southern and northern populations[4], suggest a high degree of morphological and genetic differentiation, and raise the possibility that *C. antiquata* may differentiate into different cryptic species[4,21]. To date, there is no definite conclusion on whether the *C. antiquata* population has differentiated into subspecies, and the genetic mechanism of adaptive evolution among different populations is still unknown. The subsequent research on the differentiation and population division in *C. antiquata* needs the support of whole genome data containing more comprehensive genetic information.

In this study, we combined PacBio HiFi and Hi-C sequencing technology to construct the first chromosomal-level genome assembly of *C. antiquata*. The assembled genome was 791.83 Mb in size with the scaffold N50 of 44.05 Mb, and 99.79% (790.13 Mb) of the assembly successfully anchored on 19 chromosomes. A total of 24,592 protein-coding genes were predicted, with 89.88% (n = 22,104) of them being functionally annotated. The Benchmarking Universal Single-Copy Orthologs (BUSCO)[22] analysis of the genome demonstrated a completeness of 97.69% (n = 932) of chromosome-level genome. The acquisition of high-quality *C. antiquata* genome provides a comprehensive genetic resource for advancing research in molecular genetics, population genetics and germplasm conservation.

Methods

Sample preparation. The sample was collected in Qingdao, Shandong Province, China in September 2022. The collected fresh individual was dissected on the ice, and seven tissues (mantle, gills, foot, adductor muscle, siphons, viscera and gonad) were excised and then immediately frozen in liquid nitrogen for nucleic acid extraction. When using scalpels, forceps, modified scissors, etc. to extract different tissues, it was necessary to wash the tissues and sampling tools with phosphate buffered saline (PBS, 1X) and ethanol, respectively to prevent any possible impurities outside the sample and contamination between different parts of the tissue. The samples were kept at −80 °C for storage. With a modified SDS-based method[23], high molecular weight genomic DNA was extracted from adductor muscle for both Illumina and PacBio Hifi sequencing. Total RNA was isolated from a total of seven tissues using the TRIzol reagent (Vazyme, China) for transcriptome sequencing.

Illumina sequencing. The high-quality genomic DNA was randomly fragmented into segments by the Covaris ultrasonic crusher to prepare a paired-end library with the inserted fragments of 350 bp. The constructed library was sequenced on Illumina Hiseq platform. To ensure the quality of analysis, the raw reads were processed by fastp (v0.23.4)[24] to filter out adapters and low-quality reads, and 65.30 Gb of clean data was obtained (Table 1).

PacBio HiFi sequencing. The high-quality genomic DNA was randomly fragmented into 15–18 kb to build a PacBio HiFi library 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 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 37.10 Gb of HiFi clean data was generated, with an average read length of 17,031 bp (Table 1).

Hi-C sequencing. The muscle from the foot of *C. antiquata* was used to construct the Hi-C library following the standard protocol[25] with several modifications. Briefly, after grinding with liquid nitrogen, the tissue was cross-linked by 4% formaldehyde solution to preserve the 3D structure of DNA. After an overnight digestion with the 4-cutter restriction enzyme MboI, the DNA ends were marked with biotin-14-dCTP. The following steps involved blunt-end ligation of the cross-linked fragments, re-ligation of the proximal chromatin DNA, and reverse crosslinking of the nuclear complexes. Then, DNA was purified by the phenol-chloroform extraction. After removal of biotin from non-ligated fragment ends and repair of sheared fragments ends, biotin-labeled Hi-C samples were enriched, ligated by sequencing adapters and amplified by polymerase chain reaction (PCR, 12–14 cycles). The qualified library was sequenced on Illumina Hiseq platform. After removing adapters and low-quality reads by Trim Galore (v0.6.10, https://github.com/FelixKrueger/TrimGalore), we obtained 99.69 Gb clean data, with a Q20 of 96.80% (Table 2).

Type	Data
Raw paired reads	334,077,409
Raw Base(bp)	100,223,222,700
Clean Base(bp)	99,690,485,851
Effective Rate(%)	99.47%
Q20(%)	96.80
Q30(%)	91.22
GC Content(%)	35.06

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

Sample	Raw data (Gb)	Clean data (Gb)	Raw Reads	Clean Reads	Q20 (%)	Q30 (%)	GC (%)
Mantle	6.65	6.42	44,335,212	42,803,966	96.95	91.58	35.06
Gills	6.22	6.10	41,498,594	40,662,494	97.34	92.20	34.30
Foot	6.87	6.74	45,804,886	44,959,106	97.72	93.25	36.73
Adductor muscle	6.61	6.44	44,078,548	42,905,486	97.28	92.44	36.17
Siphons	6.82	6.58	45,450,476	43,895,912	97.38	92.66	35.60
Viscera	6.51	6.29	43,403,102	41,963,250	97.18	92.25	35.57
Gonad	6.32	6.32	42,146,312	40,323,936	97.40	92.75	37.09

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

RNA-seq sequencing. The high-quality total RNA from seven tissues were used for construction of cDNA libraries by NEBNext Ultra RNA Library Prep Kit for Illumina (NEB, USA) following manufacturer's recommendations. The qualified libraries were sequenced on Illumina Novaseq6000 platform with PE150 strategy. We utilized fastp (v0.23.4)[24] to conduct quality control for raw data, and a total of 44.89 Gb clean RNA-seq data was obtained (Table 3).

Genome survey. Prior to PacBio HiFi sequencing, we conducted genomic surveys to assess the genomic characteristics including genome size, heterozygosity, and replication rates, in order to determine the amount of subsequent sequencing data. We used jellyfish (v 2.2.10)[26] to count k-21mers with parameters '-m 21 -s 10 G -C'. The generated histogram was used as input file for GenomeScope (v2.0)[27] to estimate the genetic characteristics. The results showed that the genome size and heterozygosity of *C. antiquata* were approximately 771.14 Mb and 1.99%, respectively (Fig. 1).

Genome assembly. Based on PacBio HiFi reads and Hi-C reads, the genome was *de novo* assembled by Hifiasm (v0.19.9-r616)[28] in 'Hi-C Intergrated Assembly' mode with default parameters. Then, purge_dups (v1.2.5)[29] was used to remove haplotypic and heterozygous duplication from the *de novo* assembly, resulting in a total length of 791.66 Mb.

To obtain a chromosome-level genome assembly of *C. antiquata*, we utilized ALL-HIC (v0.9.13)[30] to carry out Hi-C scaffolding with four steps including correction, partition, optimization and building. The binary format file storing a HIC matrix was generated using 3D-DNA (v180114)[31]. Then, the JuiceboxGUI (v1.11.08)[32] was employed to correct assembly manually based on the Hi-C heatmap and the final genome-wide heatmap of Hi-C interactions (Fig. 2) was plotted by the HiCExplorer (v3.7)[33]. The final *C. antiquata* genome size was 791.83 Mb, and a total of 790.13 Mb (99.79%) sequences were anchored to 19 chromosomes of which the longest and shortest were 66.60 Mb and 15.76 Mb, respectively (Table 4, Fig. 3). The result of Hi-C scaffolding is consistent with the reported karyotype analysis of *C. antiquata*[34].

Annotation of repetitive elements. We combined Extensive *de novo* TE Annotator (EDTA, v2.0.1)[35] with RepeatModeler (v2.0.5)[36] to conduct a comprehensive repeat library for *C. antiquata* genome. Based on this *de novo* repeat library, repetitive elements were predicted and classified by RepeatMasker (v4.1.5)[37]. The result of repeat annotation demonstrated that 50.55% of the genome were identified as dispersed repeats (48.74%) or tandem repeats (1.81%). The most abundant repetitive element was DNA transposons (28.14%), followed by short interspersed nuclear elements (SINEs, 6.58%), while unclassified repeats accounted for 9.30% of the genome (Table 5, Fig. 3).

Noncoding RNA (ncRNA) annotation. Barrnap (v0.9, https://github.com/tseemann/barrnap) was applied to predict ribosomal RNAs (rRNAs) with default parameters. As for prediction of transfer RNAs (tRNAs), tRNAscan-SE[38] was employed. Based on aligning the genome to Rfam database (v14.8)[39], small nuclear RNAs (snRNAs) and microRNAs (miRNAs) were predicted by Infernal (v1.1.4)[40]. As a result, a total of 78,226 ncRNAs were annotated, including 353 rRNAs, 77,701 tRNAs, 120 snRNAs and 52 miRNAs (Table 6, Fig. 3).

Prediction and functional annotation of protein-coding genes. The predictions of gene structure were performed with three combined strategies including *de novo*, homology-based, and transcriptome-based

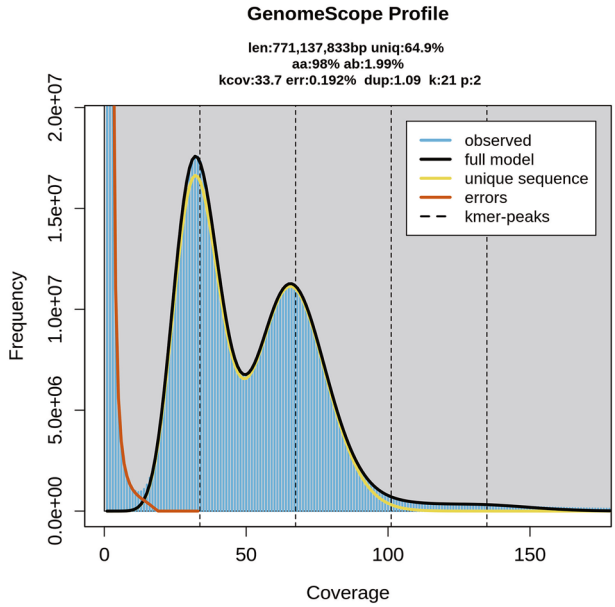

Fig. 1 Genome survey at 21-mer of *C. antiquata*.

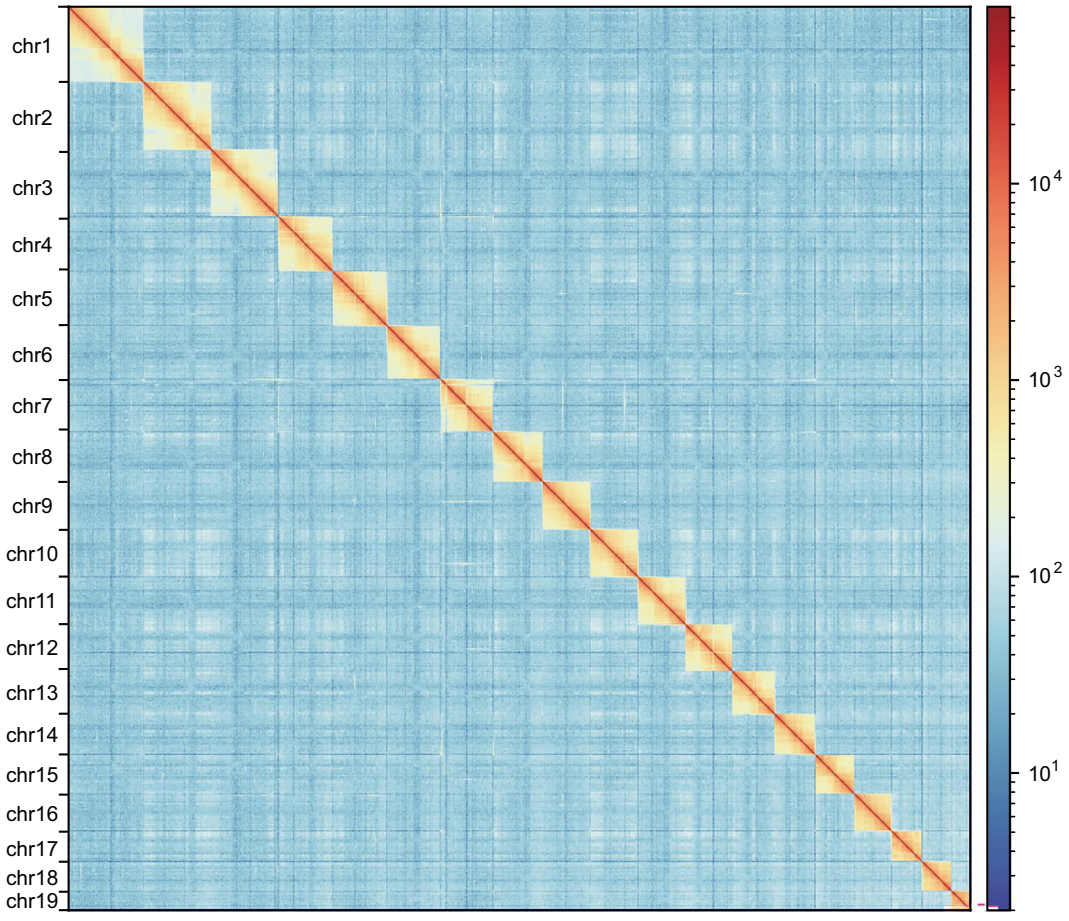

Fig. 2 Genome-wide heatmap of Hi-C interactions among 19 chromosomes in *C. antiquata*. Various colors represent different interaction frequencies of Hi-C links, and the aggregated color blocks represent the interaction frequencies between individual chromosomes.

Type	Contig (bp)	Scaffold (bp)
Total Number	476	111
Total Length	791,660,678	791,834,592
Average Length	1,663,153	7,133,645
Max Length	26,798,085	66,601,155
N50 Length	8,652,042	44,051,477
N50 Number	28	8
N90 Length	2,115,000	32,142,164
N90 Number	89	16

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

prediction, respectively. Augustus (v3.5.0)[41], GeneMark (v4.71_lic)[42], GlimmerHMM (v3.0.4)[43], SNAP (v2006-07-28)[44] and BRAKER2 (v2.1.6)[45] were performed for *de novo* prediction with default parameters. MetaEuk (v6.a5d39d9)[46] was utilized to perform homology-based prediction using protein sequences of *Crassostrea virginica*[47], *Cyclina sinensis*[48], *Mercenaria mercenaria*[49], *Mizuhopecten yessoensis*[50], *Modiolus philippinarum*[51], *Scapharca broughtonii*[52] and *Spisula solida*[53] with default parameters. As for transcriptome-based prediction, two approaches were used. Firstly, we combined clean RNA-seq data from seven tissues, and mapped it to the assembled genome using HISAT2 (v2.1.0)[54] with the parameter '-dta'. Subsequently, transcript structures were recovered by StringTie (v2.2.1)[55]. Secondly, we *de novo* assembled the transcriptome using Trinity (v2.15.1)[56] with default parameters and utilized Program to Assemble Spliced Alignments (PASA, v2.5.3) pipeline (https://github.com/PASApipeline/PASApipeline) to predict open reading frames of the transcripts. After completing prediction process of these three approaches, we integrated the results using EvidenceModeler (EVM, v1.1.1)[57] and Funannotate (v1.8.16) pipeline (https://github.com/nextgenusfs/funannotate) to obtain protein-coding gene models. After manually removing low-quality gene structures which were only supported by one approach, we predicted a total of 24,592 protein-coding genes, among which 17,783 genes (72.31%) were supported by evidence from all three prediction strategies of gene structures (Fig. 4). The average length of protein-coding genes from *C. antiquata* genome was 15,039.92 bp, with an average exon number of 7.63 and an average exon length of 251.04 bp (Table 7). To visualize gene distribution, we plotted the density of genes on 19 chromosomes with a window of 1 Mb in length using RIdeogram (v0.2.2)[58] in R (Fig. 5). On whole-genome level, we compared gene length, intron length, CDS length, exon number per gene and exon length of *C. antiquata* and other seven species used in homology-based predictions (Fig. 6).

Funannotate (v1.8.16) pipeline performed functional annotation of the protein-coding genes based on databases, including Clusters of Orthologous Groups of Proteins (COG)[59], eggNOG[60], Gene Ontology (GO)[61], InterPro[62] and Pfam[63]. Subsequently, these genes were functionally annotated using diamond (v2.1.8.162)[64] against National Center for Biotechnology Information (NCBI) non-redundant protein (Nr) and Swiss-Prot[65] databases with an E-value cutoff of $1e^{-5}$. We also utilized eggNOG-mapper (v2.1.12)[66] to obtain Kyoto Encyclopedia of Genes and Genomes (KEGG)[67] annotations. As a result, 22,104 protein-coding genes accounting for 89.88% were functionally annotated with at least one public database (Table 8, Fig. 7).

Data Records

The PacBio, Illumina, RNA-seq, and Hi-C sequencing data of *Coelomactra antiquata* have been deposited in the Sequence Read Archive (SRA) at NCBI database (https://www.ncbi.nlm.nih.gov/) and are publicly accessible under the accession numbers of SRR30920928[68], SRR30936336[69], and SRR30936337[70]. The RNA-seq data have been deposited in the SRA at NCBI SRR30941538[71], SRR30941539[72], SRR30941540[73], SRR30941541[74], SRR30941542[75], SRR30941543[76], SRR30941544[77]. The dataset is available at NCBI SRA under SRP537372[78]. The final chromosome-level genome has been deposited in GenBank with accession number GCA_047288015.1[79]. The final chromosome-level assembly and genome annotation results are available in the Figshare database[80].

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 46.85 × PacBio HiFi reads and 125.90 × Hi-C reads to ensure a high-quality genome assembly. Prior to this, we conducted a 21-mer distribution analysis to estimate the genome size based on 82.47 × Illumina reads. To assess the quality of *C. antiquata* genome assembly, we adopted three methods as follows. Firstly, we used Merqury (v1.3)[81] to estimate the base-level accuracy and completeness based on k-mer counts generated from Illumina reads, resulting in a QV of 54.89 (Table 9). Secondly, Clipping Reveals Assembly Quality (CRAQ, v1.09)[82] 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 94.59 and a S-AQI of 99.24 (Table 9). Thirdly, we mapped Illumina clean reads to *C. antiquata* genome using bwa (v0.7.17-r1188)[83]. The statistical result from samtools (v1.9)[84] showed the genome mapping rate and the coverage rate were 99.12% and 99.96% (Table 9), respectively. These results collectively revealed the high quality of *C. antiquata* genome assembly.

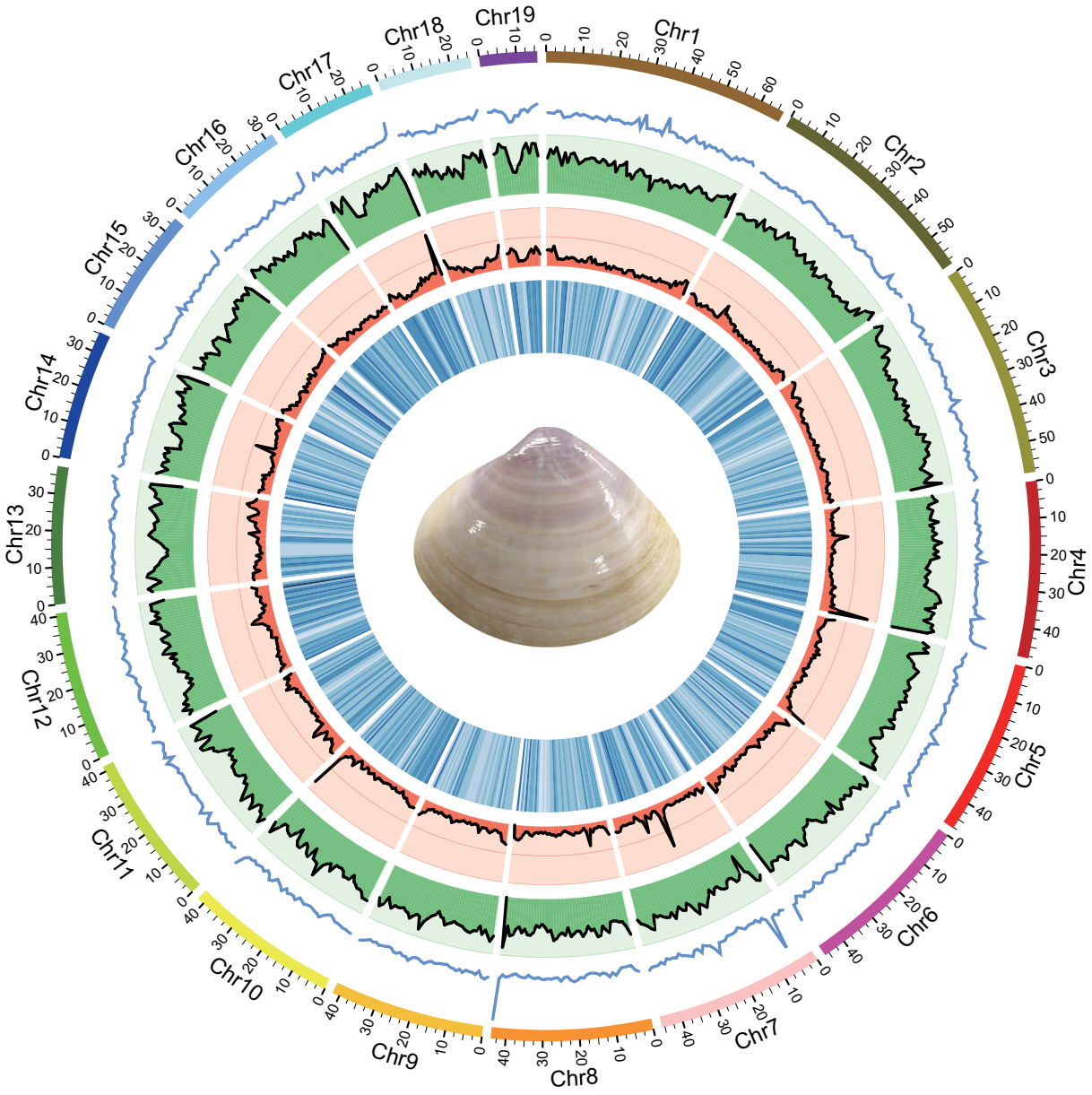

Fig. 3 Circos plot of genomic features in *C. antiquata*. Each track from outer to inner represents the chromosome length (Mb), and the distribution of GC content, repeat elements, ncRNAs, protein-coding genes with a sliding window size of 1 Mb.

Type			Count	Length (bp)	% of Genome
Dispersed repeats	DNA transposons		1,906,025	229,070,289	28.14
	Retroelements	DIRS	193	87,735	0.01
		LTR	164,511	37,986,590	4.80
		LINE	90,854	28,507,495	3.60
		Penelope	200	30,226	4.82
		SINE	300,840	52,065,496	6.58
	Unclassified		167,954	38,149,966	9.30
Tandem repeats	Simple repeats		100,675	6,522,754	0.82
	Low complexity		11,407	601,427	0.08
	Satellite		19,336	7,240,311	0.91
Total			2,761,995	400,262,289	50.55

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

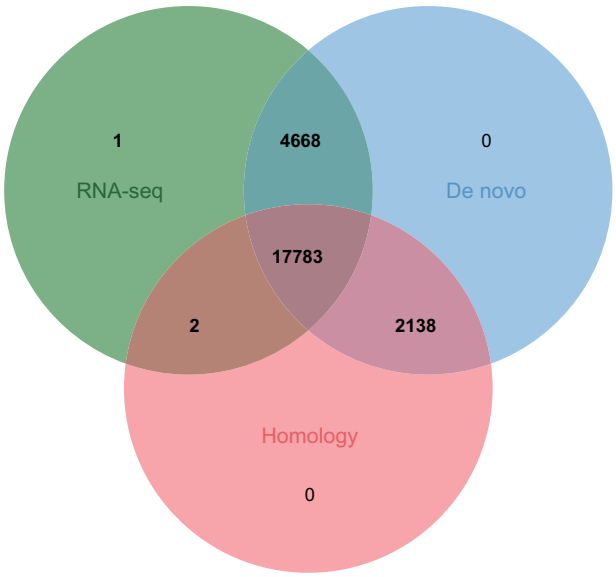

Fig. 4 Venn diagram of *C. antiquata* gene structure prediction through three strategies.

Type		Copy number	Average length(bp)	Total length(bp)	% of genome
tRNA		77,701	73.08	5,678,313	0.71710848
miRNA		52	80.71	4,197	0.00053003
snRNA	CD-box	33	94.73	3,126	0.00039478
	HACA-box	14	233.00	3,262	0.00041195
	scaRNA	1	128.00	128	0.00001616
	splicing	72	145.51	10,477	0.00132313
rRNA	5S	338	113.66	38,416	0.00485152
	5.8S	5	150.00	750	0.00009472
	18S	5	1,728.40	8,642	0.00109139
	28S	5	4,268.60	21,343	0.00269539

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

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	—	29,270	10,002.08	1,501.18	1,498.87	225.01	6.67
	BRAKER2	—	32,231	10,236.79	1,492.52	2,279.24	212.79	7.01
	GlimmerHMM	—	85,017	7,863.11	732.61	2,948.77	204.58	3.58
	GeneMark	—	64,614	4,309.28	836.21	1,083.12	198.79	4.21
	SNAP	—	77,935	3,008.78	527.51	1,907.87	229.30	2.30
RNA-seq	HISAT2 & StringTie	—	20,194	20,687.80	2,726.98	2,626.65	454.91	9.61
	PASA	—	29,928	21,043.10	1,202.82	2,869.62	417.63	7.27
Homology	MetaEuk	*C. virginica*	14,161	6,705.35	1,046.70	1,614.45	274.45	4.40
		C. sinensis	20,639	7,730.76	1,174.61	1,576.56	232.39	5.15
		M. mercenaria	22,309	6,519.86	981.48	1,600.26	224.53	4.45
		M. yessoensis	12,207	6,971.27	1,086.48	1,628.24	239.70	4.60
		M. philippinarum	21,892	3,771.01	809.98	1,619.86	289.93	2.82
		S. broughtonii	15,737	6,232.98	1,142.74	1,606.99	278.34	4.16
		S. solida	17,471	7,108.10	992.99	1,632.19	213.88	4.73
Final		—	24,592	15,039.92	1,520.94	1,976.08	251.04	7.63

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

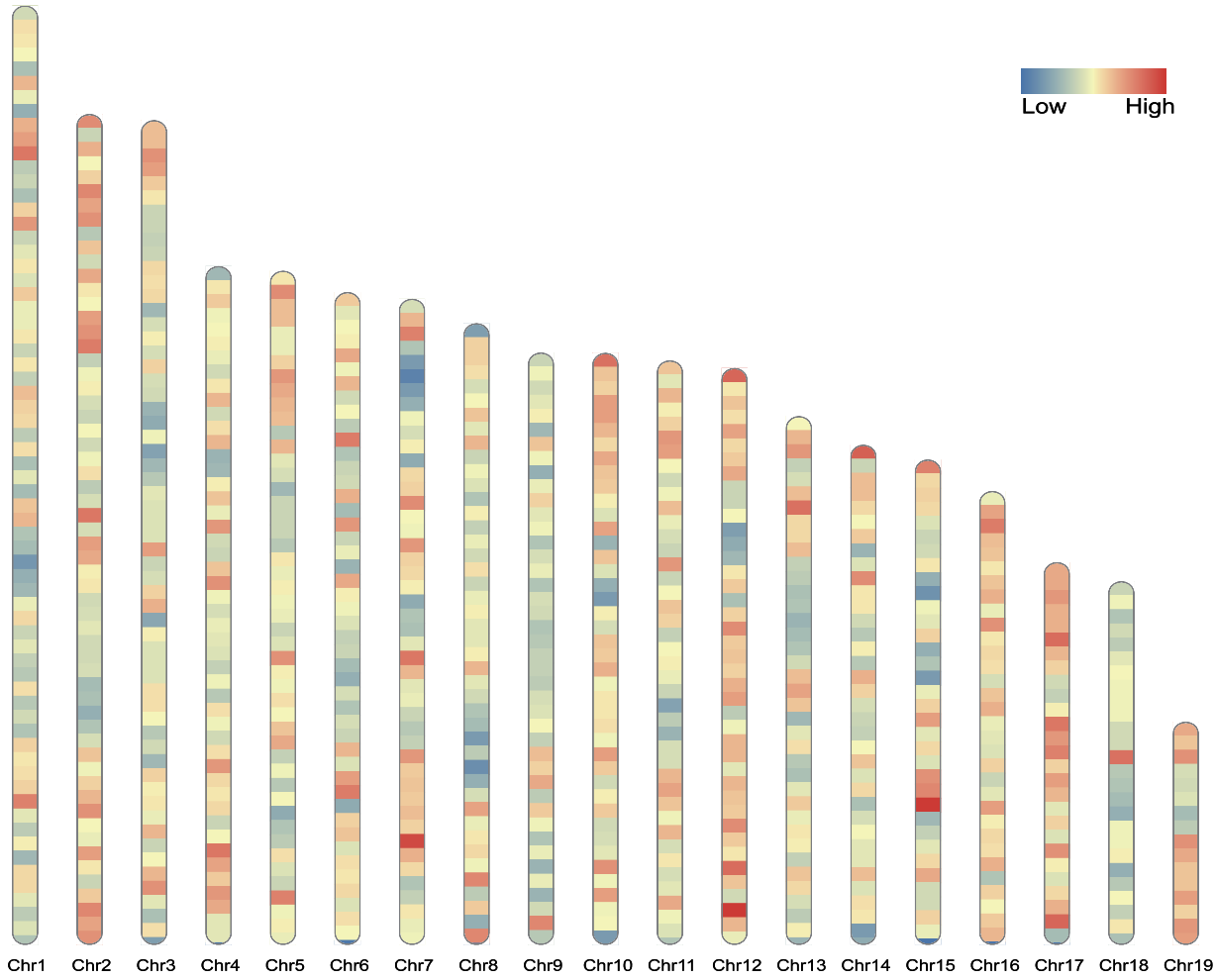

Fig. 5 Gene density of *C. antiquata* each chromosome within 1 Mb sliding windows.

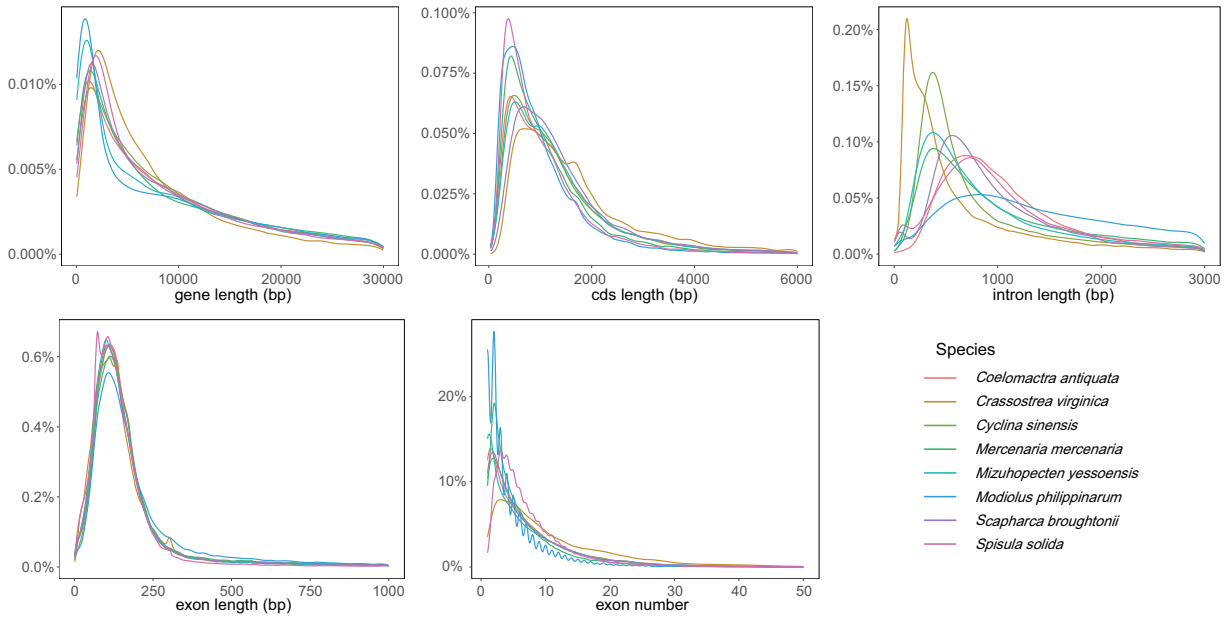

Fig. 6 *C. antiquata* Gene length, CDS length, intron length, exon length, exon number per gene compared with seven other species in Bivalva.

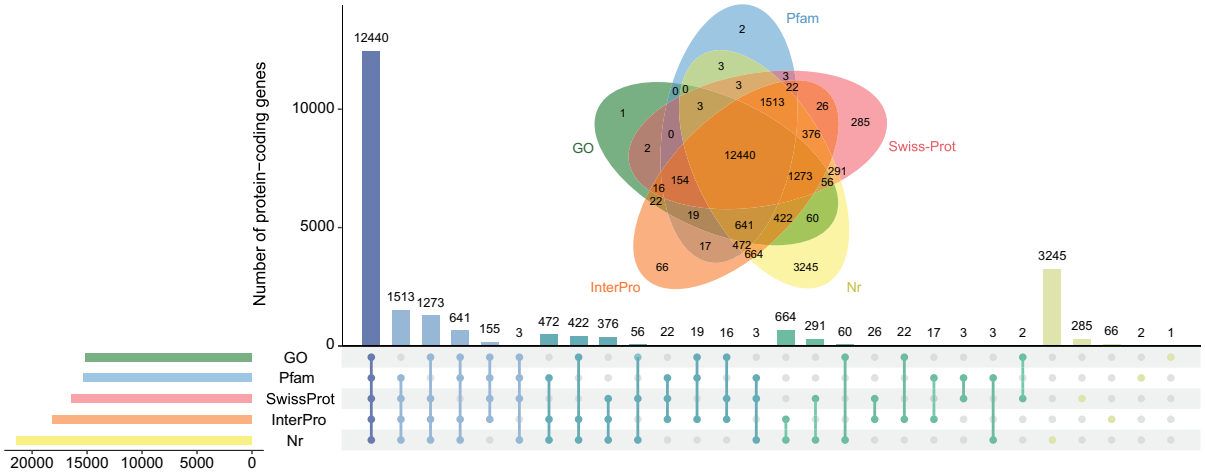

Fig. 7 Venn diagram and Upset plot of *C. antiquata* functional gene annotation based on five of the eight databases.

Database	Number	Percent(%)
Total	24,592	—
COG	17,019	69.21
eggNOG	17,917	72.86
GO	15,109	61.44
InterPro	18,143	73.78
KEGG	8,639	35.13
Pfam	15,292	62.18
Swiss-Prot	16,463	66.94
Nr	21,462	87.27
Annotated	22,104	89.88
Unannotated	2,961	10.12

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

	Type		Results
Genome accuracy and completeness	Mapping short-reads rate		99.12%
	Genome coverage rate with short-reads		99.96%
	Quality value scores (QVs)		54.89
	CRAQ	R-AQI	94.59%
		S-AQI	99.24%
Annotation quality	CEGMA	Complete Prots	90.32% (224)
		Partial Prots	98.38% (244)
	BUSCO	Complete BUSCOs (C)	97.69% (932)
		Complete and single-copy BUSCOs (S)	96.86% (924)
		Complete and duplicated BUSCOs (D)	0.84% (8)
		Fragmented BUSCOs (F)	1.05% (10)
		Missing BUSCOs (M)	1.26% (12)

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

To evaluate the quality of gene annotation, we utilized the Benchmarking Universal Single-Copy Orthologs (BUSCO, v5.5.0)[22] and the Core Eukaryotic Genes Mapping Approach (CEGMA, v2.5)[85]. In *C. antiquata* genome, 97.69% of complete BUSCOs in the metazoa_odb10 database were identified (Table 9). The results from CEGMA showed that 244 out of 248 (98.38%) core eukaryotic genes were successfully annotated (Table 9). These assessments validated the high quality of *C. antiquata* genome annotation.

Code availability

All of the software and pipelines utilized for data processing in this study were run according to the appropriate manuals and protocols. The versions and parameters of the tools used have been mentioned in the Method section, if not, the default parameters were applied. No custom code has been used.

Received: 22 October 2024; Accepted: 28 February 2025;
Published online: 11 March 2025

References

1. Kong, L. F. & Li, Q. Genetic evidence for the existence of cryptic species in an endangered clam *Coelomactra antiquata*. *Mar. Biol.* **156**, 1507–1515, https://doi.org/10.1007/s00227-009-1190-5 (2009).
2. Wen, W. Xishi's tongue—a promising precious species for aquaculture. *Guidance Profitable Fishery* **14**, 42 (2003).
3. Liu, D. J. Aquaculture biology and development prospect for cultivation of Xishi tongue. *Shandong Fisheries* **21**, 15–18 (2004).
4. Kong, L. F., Li, Q. & Qiu, Z. X. Genetic and morphological differentiation in the clam *Coelomactra antiquata* (Bivalvia: Veneroida) along the coast of China. *J. Exp. Mar. Biol. Ecol.* **343**, 110–117, https://doi.org/10.1016/j.jembe.2006.12.003 (2007).
5. Liu, H. *et al*. The clam, Xishi tongue *Coelomactra antiquata* (Spengler), a promising new candidate for aquaculture in China. *Aquaculture* **255**, 402–409, https://doi.org/10.1016/j.aquaculture.2005.12.027 (2006).
6. Meng, X. P., Gao, R. C., Dong, Z. G., Cheng, H. L. & Yan, B. L. Analysis and evaluation of nutritional composition in edible part of *Coelomactra antiquata*. *Marine Science* **31**, 17–22, https://doi.org/10.3969/j.issn.1000-3096.2007.01.004 (2007).
7. Li, Z. H. *et al*. Analysis and evaluation of nutritional composition of edible part of *Coelomactra antiquata* (Spengler) in nature sanctuary of *Coelomactra antiquata* of Fujian Province. *Food Science* **35**, 176–182, https://doi.org/10.7506/spkx1002-6630-201405035 (2014).
8. Meng, X. P., Cheng, H. L. & Dong, Z. G. The study status and prospect of *Coelomactra antiquata* in China. *Journal of Hebei Normal University of Science & Technology* **19**, 71–75, https://doi.org/10.3969/j.issn.1672-7983.2005.04.018 (2005).
9. Qi, Q. Z., Gao, R. C., Qiu, W. R. & Huang, X. Q. The life history of *Coelomactra antiquata*. *Journal of Fujian Normal University* **11**, 82–88 (1995).
10. Wu, J. F., Zhang, H. H., Liang, C. Y. & Chen, S. W. Resource and conservation strategy of *Coelomactra antiquata* in coast of Guangdong province, China. *Journal of Zhanjiang Ocean University* **22**, 68–69, https://doi.org/10.3969/j.issn.1673-9159.2002.03.014 (2002).
11. Wang, S. & Xie, Y., *China species red list* vol. III. Invertebrates. (Higher Education press, 2005)
12. Ni, L. H., Li, Q., Kong, L. F., Huang, S. Q. & Li, L. J. DNA barcoding and phylogeny in the family Mactridae (Bivalvia: Heterodonta): Evidence for cryptic species. *Biochem. Syst. Ecol.* **44**, 164–172, https://doi.org/10.1016/j.bse.2012.05.008 (2012).
13. Yuan, Y., Kong, L. F. & Li, Q. Mitogenome evidence for the existence of cryptic species in *Coelomactra antiquata*. *Genes Genomics* **35**, 693–701, https://doi.org/10.1007/s13258-013-0120-6 (2013).
14. Gao, R. C. Advances in studies on biology and artificial breeding of the bivalve *coelomatra antiquata*. *Journal of Xiamen University* **45**, 195–200, https://doi.org/10.3321/j.issn:0438-0479.2006.z2.028 (2006).
15. Lin, J. J. Current status and protection efficiency for *Mactra antiquata* stock in Changle, Fujian. *Journal of Applied Oceanography* **39**, 551–558, https://doi.org/10.3969/J.ISSN.2095-4972.2020.04.011 (2020).
16. Kong, L. F. & Li, Q. Genetic comparison of cultured and wild populations of the clam *Coelomactra antiquata* (Spengler) in China using AFLP markers. *Aquaculture* **271**, 152–161, https://doi.org/10.1016/j.aquaculture.2007.06.007 (2007).
17. Hebert, P. D. N., Cywinska, A., Ball, S. L. & DeWaard, J. R. Biological identifications through DNA barcodes. *Proc. R. Soc. Lond. B* **270**, 313–321, https://doi.org/10.1098/rspb.2002.2218 (2003).
18. Meng, X. P. *et al*. Complete mitochondrial genome of *Coelomactra antiquata* (Mollusca: Bivalvia): The first representative from the family Mactridae with novel gene order and unusual tandem repeats. *Comp. Biochem. Physiol. D: Genomics Proteomics* **7**, 175–179, https://doi.org/10.1016/j.cbd.2012.02.001 (2012).
19. Meng, X. P. *et al*. Mitogenomics reveals two subspecies in *Coelomactra antiquata* (Mollusca: Bivalvia). *Mitochondrial DNA* **24**, 102–104, https://doi.org/10.3109/19401736.2012.726620 (2012).
20. Shen, X. *et al*. Comparative mitogenomic analysis reveals cryptic species: A case study in Mactridae (Mollusca: Bivalvia). *Comp. Biochem. Physiol. D: Genomics Proteomics* **12**, 1–9, https://doi.org/10.1016/j.cbd.2014.08.002 (2014).
21. Yi, L. F. *et al*. Insights into cryptic diversity and adaptive evolution of the clam *Coelomactra antiquata* (Spengler, 1802) from comparative transcriptomics. *Mar. Biodivers.* **49**, 2311–2322, https://doi.org/10.1007/s12526-019-00964-w (2019).
22. Simao, 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).
23. Futoshi, A. PCR-RFLP analysis of nuclear nontranscribed spacer for mackerel species identification. *J. Agric. Food Chem.* **53**, 508–511, https://doi.org/10.1021/jf0484881 (2005).
24. Chen, S. F., Zhou, Y. Q., Chen, Y. R. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890, https://doi.org/10.1093/bioinformatics/bty560 (2018).
25. 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).
26. 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).
27. Vurture, G. W. *et al*. GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202–2204, https://doi.org/10.1093/bioinformatics/btx153 (2017).
28. 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).
29. 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).
30. Zhang, X. T., Zhang, S. C., Zhao, Q., Ming, R. & Tang, H. B. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. *Nat. Plants* **5**, 833–845, https://doi.org/10.1038/s41477-019-0487-8 (2019).
31. Dudchenko, O. *et al*. *de novo* assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Sci.* **356**, 92–95, https://doi.org/10.1126/science.aal3327 (2017).
32. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101, https://doi.org/10.1016/j.cels.2015.07.012 (2016).
33. Wolff, J. *et al*. Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. *Nucleic Acids Res.* **48**, W177–W184, https://doi.org/10.1093/nar/gkaa220 (2020).
34. Rao, X. Z., Xu, Y. Q., Chen, Y. S. & Gao, R. C. Karyotype Analysis of *Coelomactra antiquata*. *Chinese Journal of Zoology* **38**, 2–5, https://doi.org/10.13859/j.cjz.2003.02.001 (2003).

35. 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).
36. 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).
37. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Curr. Protoc. Bioinf.* **Ch. 4**, 10.1–10.14, https://doi.org/10.1002/0471250953.bi0410s25 (2009).
38. 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).
39. 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).
40. 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).
41. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–439, https://doi.org/10.1093/nar/gkl200 (2006).
42. 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).
43. 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).
44. Korf, I. Gene finding in novel genomes. *BMC Bioinf.* **5**, 59, https://doi.org/10.1186/1471-2105-5-59 (2004).
45. 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).
46. Levy Karin, E., Mirdita, M. & Soding, J. MetaEuk—sensitive, high-throughput gene discovery, and annotation for large-scale eukaryotic metagenomics. *Microbiome* **8**, 48, https://doi.org/10.1186/s40168-020-00808-x (2020).
47. Gómez-Chiarri, M., Warren, W. C., Guo, X. M. & Proestou, D. Developing tools for the study of molluscan immunity: The sequencing of the genome of the eastern oyster, *Crassostrea virginica*. *Fish Shellfish Immunol.* **46**, 2–4, https://doi.org/10.1016/j.fsi.2015.05.004 (2015).
48. Wei, M. *et al.* Chromosome-Level clam genome helps elucidate the molecular basis of adaptation to a buried lifestyle. *iScience* **23**, 101–148, https://doi.org/10.1016/j.isci.2020.101148 (2020).
49. Song, H. *et al.* The hard clam genome reveals massive expansion and diversification of inhibitors of apoptosis in Bivalvia. *BMC Biol.* **19**, 15, https://doi.org/10.1186/s12915-020-00943-9 (2021).
50. Wang, S. Scallop genome provides insights into evolution of bilaterian karyotype and development. *Nat. Ecol. Evol.* **1**, 120, https://doi.org/10.1038/s41559-017-0120 (2017).
51. Sun, J. Adaptation to deep-sea chemosynthetic environments as revealed by mussel genomes. *Nat. Ecol. Evol.* **1**, 121, https://doi.org/10.1038/s41559-017-0121 (2017).
52. Bai, C.-M. *et al.* Chromosomal-level assembly of the blood clam, *Scapharca* (*Anadara*) *broughtonii*, using long sequence reads and Hi-C. *GigaScience* **8**, giz067, https://doi.org/10.1093/gigascience/giz067 (2019).
53. Holmes, A. The genome sequence of the surf clam, *Spisula solida* (Linnaeus, 1758). *Wellcome Open Res.* **8**, 227, https://doi.org/10.12688/wellcomeopenres.19486.1 (2023).
54. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360, https://doi.org/10.1038/nmeth.3317 (2015).
55. Kovaka, S. *et al.* Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 278, https://doi.org/10.1186/s13059-019-1910-1 (2019).
56. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652, https://doi.org/10.1038/nbt.1883 (2011).
57. 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).
58. Hao, Z. D. *et al.* RIdeogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms. *PeerJ Comput. Sci.* **6**, e251, https://doi.org/10.7717/peerj-cs.251 (2020).
59. 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).
60. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314, https://doi.org/10.1093/nar/gky1085 (2019).
61. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29, https://doi.org/10.1038/75556 (2000).
62. 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).
63. 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).
64. 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).
65. 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).
66. 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).
67. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–462, https://doi.org/10.1093/nar/gkv1070 (2016).
68. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30920928 (2024).
69. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30936336 (2024).
70. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30936337 (2024).
71. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30941538 (2024).
72. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30941539 (2024).
73. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30941540 (2024).
74. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30941541 (2024).
75. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30941542 (2024).
76. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30941543 (2024).
77. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR30941544 (2024).
78. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP537372 (2024).
79. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_047288015.1 (2024).
80. Shen, Y. W., Wang, Y. L. & Kong, L. F. Chromosome-level genome assembly of *Coelomactra antiquata*. *Figshare* https://doi.org/10.6084/m9.figshare.27201636.v2 (2024).

81. 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).
82. 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).
83. 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).
84. Li, H. *et al*. The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079, https://doi.org/10.1093/bioinformatics/btp352 (2009).
85. Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**, 1061–1067, https://doi.org/10.1093/bioinformatics/btm071 (2007).

Acknowledgements

This work was supported by the Key R&D Program of Shandong Province, China (2022TZXD002), the Shandong Provincial Natural Science Foundation (ZR2023MD008), the Qingdao Natural Science Foundation (23-2-1-166-zyyd-jch) and Laoshan Laboratory. We also acknowledge the support of the High-Performance Biological Supercomputing Center at the Ocean University of China for this research.

Author contributions

L.F.K. conceived this study. Y.W.S performed the sample preparation. Y.L.W. conducted the data analysis. Y.W.S. drafted the manuscript. L.F.K. and Y.L.W. provided revisions and suggestions for the manuscript. All authors reviewed and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to L.K.

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.