scientific data

Check for updates

OPEN

DATA DESCRIPTOR

The chromosome-level genome assembly and annotation of the silver-lipped pearl oyster, *Pinctada maxima*

Qianqian Mu[1,2,4], Zijian Li[1,2,4], Mingyang Liu[1,2,4], Baojun Zhao[1], Zhihui Yang[1], Zhenmin Bao[1,3], Jingjie Hu[1,3] & Pingping Liu[1,3] ✉

The silver-lipped pearl oyster (*Pinctada maxima*) is a valuable tropical aquaculture species, playing a crucial economic role in the global pearl industry. However, the lack of genomic reference limits our in-depth understanding of this species in genome-based breeding, conservation, evolution and adaptation. Here, annotated chromosome-level reference genome for *P. maxima* was generated by integrating PacBio long-read sequencing, Illumina short-read sequencing, and Hi-C sequencing data. The total genome size is 1,264.93 Mb, with contig N50 and scaffold N50 of 649 kb and 89.19 Mb, respectively. The majority (97.94%) of the assembled genome was anchored to the 14 chromosomes by Hi-C analysis. The relatively high genome completeness was observed, with 97.38% (metazoa_odb10 database) and 95.26% (mollusca_odb10 database) in BUSCO analysis. Genome annotation revealed approximately 65.46% of the repeat sequences and 26,315 protein-coding genes. Comparative genome analysis revealed 28 expanded and 48 contracted families ($p < 0.05$) in *P. maxima*, with 3.2% of genes (894) being species-specific. This chromosome-level genome serves as an essential resource for research in evolutionary genomics, phylogenetics, and biomineralization.

Background & Summary

Bivalves, including a diverse group of organisms such as clams, oysters, mussels, and scallops, serve dual ecological and economic roles across aquatic ecosystems. Ecologically, they act as natural biofilters to purify water through nutrient recycling and serve as early-warning indicators for aquatic ecosystem changes due to their environmental sensitivity[1–3]. Their population viability emerges as an integrated metric of ecosystem stressors, encompassing chemical contamination, climate change, and habitat alteration[4,5]. Beyond their ecological significance, numerous bivalve species, including oysters, mussels, and scallops, are of economic importance in aquaculture with 2,700 tons production in 2022, representing a commercial value of nearly 138.5 million dollars[6]. Additionally, pearls and shells produced by bivalves are highly valued in jewelry and decorative industries, further emphasizing their economic importance.

The *P. maxima*, an important tropical aquaculture species, is naturally distributed in the central Indo-Pacific region from Myanmar to the Solomon Islands like Australia, Southeast Asia, Philippines, and South China Sea[7]. *P. maxima* is a vital economic resource for mariculture, valued for their ability to produce high-quality pearls with high economic value[8,9]. It is known for generating the largest pearls in the world, and the size of pearls often exceed 10 mm. The larger size of the pearls generated by *P. maxima*, also called highly sought "South Sea" pearls[10], make them especially desirable in the luxury market. Regions such as Australia, Indonesia, the Philippines, and French Polynesia, which cultivate these oysters, have gained huge economic benefits from pearl farming[11].

[1]Key Laboratory of Tropical Aquatic Germplasm of Hainan Province, Sanya Oceanographic Institution, Ocean University of China, Sanya, 572000, China. [2]MOE Key Laboratory of Marine Genetics and Breeding, College of Marine Life Sciences, Ocean University of China, Qingdao, 266003, China. [3]State Key Laboratory of Mariculture Biobreeding and Sustainable Goods, Beidaihe Central Experiment Station, Chinese Academy of Fishery Sciences, Qinhuangdao, 066100, China. [4]These authors contributed equally: Qianqian Mu, Zijian Li, Mingyang Liu. ✉e-mail: liupingping@ouc.edu.cn

However, overfishing and environmental changes have led to a steep decline in populations of pearl oyster[12]. China has designated them as a national second-class protected species[13]. Although artificial breeding techniques allow the production of pearl oysters, the culture industry growth has been hindered by high larvae mortality rates in mariculture[14,15]. Genomic resources are crucial for the conservation of *P. maxima* and the development of aquaculture industry of *P. maxima*. In addition, the pearl oyster serves as a crucial model organism for investigating the genetic mechanisms of biomineralization[16], a field of considerable scientific importance. However, the limited genome resources available for this key bivalve species have hindered the identification of genes involved in regulating critical quality traits and the unique biological characteristics of pearls, such as biomineralization. Furthermore, genomic data is also of great value for the study of evolution, adaptation, longevity, gonad development, and sex determination in bivalves[17–19].

In this study, the first chromosome-level genome of the *P. maxima* was generated using PacBio long-read sequencing, Illumina short-read sequencing and Hi-C technology. The repeats and protein-coding genes were annotated, and comparative genome analysis was conducted, including molecular phylogenetic and genome synteny analysis. The high-quality reference genome resources for the *P. maxima* are of immense value for genome-based breeding programs, understanding complex biological processes and conserving germplasm resources, marking a significant advancement in the field of bivalve genomics.

Methods

Sample collection. The adult *P. maxima* was collected from Lingshui County, Hainan Province, China. After dissection, the adductor muscle, smooth muscle, gonad, mantle, gill, hepatopancreas, foot, and intestine tissues were immediately frozen in liquid nitrogen and stored at −80 °C until DNA and RNA extraction for subsequent sequencing.

DNA extraction and genome sequencing. High molecular weight genomic DNA was extracted from adductor muscle using the traditional Phenol-Chloroform protocol[20]. DNA purity and concentration were measured using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific), with acceptable thresholds of A260/A280 > 1.8 and A260/A230 > 2.0. DNA integrity and size were verified by electrophoresis on a 1% agarose gel, confirming fragments >30 kb. For survey sequencing, libraries with an approximate insert size of 300 bp were constructed by using the VAHTS Universal Plus DNA Library Prep Kit for Illumina, followed by paired-end 150 sequencing on the Illumina NovaSeq. 6000 platform. A total of 67.3 Gb data were generated with 51.56 × genome coverage.

Long read sequencing was performed by using the PacBio Sequel II system (Pacific Biosciences, California, USA). PacBio libraries were prepared using the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, California, USA) according to the manufacturer's guidelines. Briefly, the library was constructed through several steps, including magnetic bead enrichment, DNA repair and A-tailing reaction, DNA purification, adapters ligation, purification to remove small DNA fragments and excess reagents. A total of 52.1 Gb of data generated, with an N_{50} read length of 17 kb.

Chromosome-level assembly was achieved by using the Hi-C technique. The flash-frozen adductor muscle was processed to construct Hi-C library using Arima Genomics Hi-C Kit (San Diego, California, USA) by following the manufacturer's instructions. The samples underwent formaldehyde cross-linking, enzyme digestion, biotin marking of DNA ends and blunt end ligation, and DNA purification. Hi-C library was subjected to pair-end (2 × 150) sequencing on the Illumina NovaSeq. 6000 platform, yielding a total of 121.5 Gb of sequencing data.

RNA extraction and transcriptome sequencing. The gonad, mantle, gill, hepatopancreas, foot, adductor muscle, smooth muscle, and intestine tissues were used to extract RNA by using TRIzol reagent. NanoDrop2000 Spectrophotometers (Thermo Fisher Scientific, Waltham, MA) were used to determine the concentration of RNA, and Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA) was adopted to assess RNA integrity. mRNA from total RNA was captured by using poly-T oligo-attached magnetic beads. Library preparation was used NEBNext Ultra RNA library preparation kit (NEB) and the prepared libraries were sequenced on Illumina NovaSeq. 6000 platform (Novogene, Sacramento, CA). Finally, a total of 50.8 Gb high quality reads was generated. To acquire more comprehensive information on full-length transcripts, a third-generation full-length transcriptome (PacBio isoform sequencing Iso-seq) library was prepared by utilizing PacBio SMRT sequencing technology. Equal amounts of RNA from the eight sampled tissues were mixed together to prepare the Iso-seq library. The library was prepared by using the Clontech SMARTer PCR cDNA Synthesis Kit (Clontech, Mountain View, CA, USA) and the BluePippin Size Selection System (Sage Science, Beverly, MA, USA), as described in the Pacific Biosciences protocol (PN 100-092-800-03). The constructed PacBio library was sequenced on the PacBio Sequel II platform.

***De novo* genome assembly.** K-mer analysis was conducted with Jellyfish[21] and Genomescope2[22] to estimate genome size, repeat sequence content and heterozygosity, based on 17-mer frequency profiles derived from 67.3 Gb of Illumina raw data. A total of 60,102,996,443 k-mers were identified, exhibiting a depth of 49×. The haploid genome size was estimated at 1.21 Gb, with a heterozygosity rate of 1.01% and repetitive sequences accounting for 61.75% of the genome. A draft contig assembly was generated using PacBio HiFi sequencing data. Subreads obtained from the PacBio Sequel II platform were processed through SMRT Link v10.2 to generate Circular Consensus Sequences (CCS) via multi-pass subread integration. CCS reads were refined using the CCS algorithm (minimum passes = 3, minimum read quality = 0.99) to eliminate adapter sequences and low-quality reads. *De novo* genome assembly was performed using Hifiasm v0.20.0[23], leveraging its capacity for high accuracy and well-connected continuity to assemble PacBio HiFi reads.

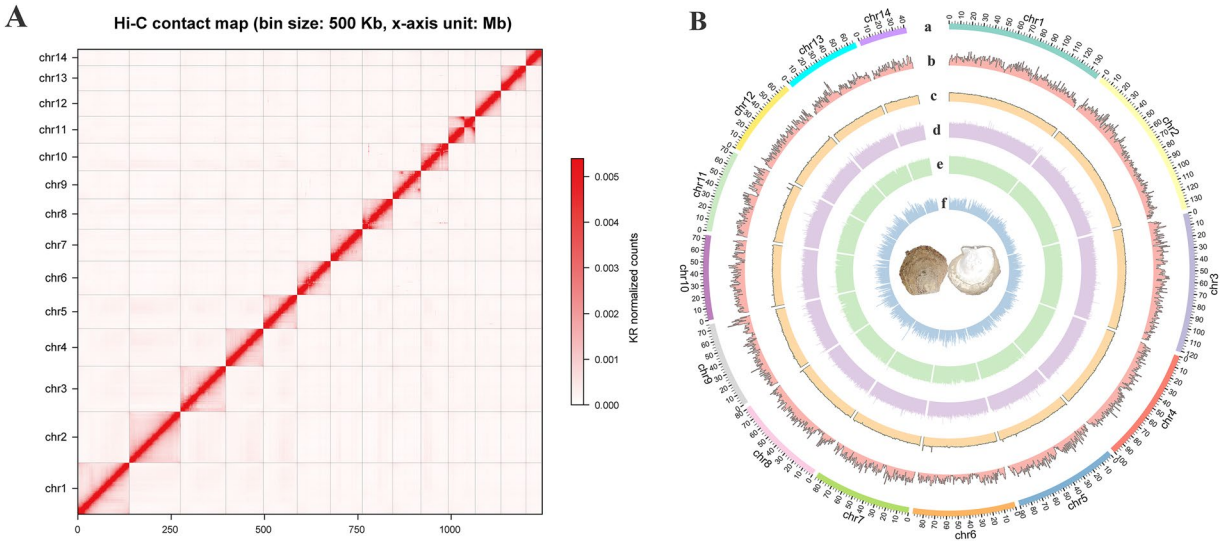

Fig. 1 Characteristics of the *Pinctada maxima* genome. (**A**) Hi-C interaction heat map of *Pinctada maxima*. (**B**) Circos plot of the *P. maxima* genome assembly. (**a**) The length of chromosomes in the size of Mb. (**b**) Density of genes with 500 kbp windows; (**c**) GC content with 500 kbp windows; (**d**) depth of coverage of PacBio HiFi reads with 100 kbp windows; (**e**) depth of coverage of Illumina short reads with 100 kbp windows; (**f**) distribution of heterozygous SNPs with 500 kbp windows.

Genome assembly	*Pinctada maxima*	*Pinctada fucata*	*Pteria penguin*	*Pinctada imbricata*	*Pinctada radiata*	*Crassostrea gigas*	*Crassostrea virginica*	*Crassostrea hongkongensis*	*Ostrea edulis*
Assembled genome size (bp)	1,264,926,820	1,039,545,996	838,713,081	990,984,031	931,129,676	563,985,803	684,741,128	608,622,584	894,786,893
Largest scaffold (bp)	137,924,361	113,122,503	93,594,500	104,615,532	98,733,858	76,070,991	104,168,038	64,470,731	112,480,954
GC rate (% of genome)	36.18	35.48	35.14	35.31	35.42	33.55	34.83	33.50	35.49
Scaffold N50 size (bp)	89,192,998	72,757,956	60,661,133	59,032,463	63,836,525	57,274,926	75,944,018	55,855,599	94,306,699
Contig N50 size (bp)	649,306	1,021,642	5,032,149	21,517	8,065,000	7,289,204	1,971,208	2,576,225	2,427,391
QV	55.64	\	\	\	\	\	\	\	\
BUSCO % (Mollusca_odb10/ Metazoa_odb10)									
Complete (C)	95.26/97.38	95.96/97.48	95.68/97.27	83.08/87.63	95.26/97.48	97.19/96.33	98.85/98.64	99.30/99.37	99.36/100.00
Complete and single-copy (S)	93.75/96.02	94.43/96.23	95.13/96.86	72.03/76.00	94.24/96.75	96.62/95.28	80.28/71.85	98.73/98.32	98.38/98.43
Complete and duplicated (D)	1.51/1.36	1.53/1.26	0.55/0.42	11.05/11.64	1.02/0.73	0.57/1.05	19.24/25.79	0.57/1.05	0.98/0.63
Fragmented (F)	0.62/0.21	0.60/0.84	0.64/0.31	3.25/4.93	0.53/0.63	0.09/0.31	0.17/0.31	0.19/0.21	0.32/0.42
Missing (M)	4.12/2.41	3.44/1.68	3.68/2.41	13.67/7.44	4.21/1.89	2.72/3.35	0.98/1.05	0.51/0.42	0.32/0.52
Total groups searched	5,295/954	5,295/954	5,295/954	5,295/954	5,295/954	5,295/954	5,295/954	5,295/954	5,295/954

Table 1. Comparison of genome assembly metrics between *Pinctada maxima* and other bivalve genomes.

Hi-C analysis and chromosome assembly. To assemble a high-quality chromosome-level genome, preliminarily assembled genome was anchored using Juicer[24] and 3D-DNA[25], with subsequent manual refinement implemented via Juicebox[24]. Hi-C chromatin interaction patterns resolved 14 chromosomal scaffolds (Fig. 1A), yielding a final assembly of 1,264.93 Mb with 36.18% GC content. Approximately 97.94% of the genome was anchored into these 14 chromosomes, and a contig N50 of 649 kb and a scaffold N50 of 89.19 Mb were observed. Genome architecture was visualized in the circos plot (Fig. 1B).

Repeat annotation. Repeat element annotation was performed through a hybrid approach combining *de novo* prediction and structural features by using RepeatModeler v2.0.3[26] (https://github.com/Dfam-consortium/RepeatModeler), EDTA v2.0.0[27] and RepeatMasker v4.1.2[28] (https://www.repeatmasker.org/). Candidate LTR-RTs repeat sequence library was identified using LTR_finder[29] with parameters '-size 5000000 -time 1500 -w 2 -C -D 15000 -d 1000 -L 7000 -l 100 -p 20 -M 0.85' and LTRharvest v1.6.2[30] with parameters '-similar 90 -vic 10 -seed 20 -seqids yes -minlenltr 100 -maxlenltr 7000 -mintsd 4 -maxtsd 6 -motif TGCA -motifmis 1'. The identified LTR-RT candidates were filtered with LTR_retriever v3.0.1[31] by using default parameters. EDTA v2.0.0, LTR_retriever v3.0.1 and RepeatModeler v2.0.3 were used to build *de novo* repeat libraries. Finally, the perl script make_panTElib.pl in the EDTA v2.0.0 program was used to integrate and obtain combined repeat library. The combined repeat library was used as the final library to identify repeat sequences using RepeatMasker v4.1.2. The

Genome annotation	*Pinctada maxima*
Number of protein-coding genes	26,315
Average gene length (bp)	20,069
Average exon length (bp)	267
Average exon number per gene	7
Average CDS length (bp)	203.41
Average intron length (bp)	2,956
Percentage of repeat sequence (%)	65.46
LTR (%)	10.25
LINE (%)	7.52
SINE (%)	0.16
DNA transposons (%)	41

Table 2. Statistics of the genome-wide annotations in *Pinctada maxima*.

Type	*Pinctada maxima*	*Pinctada fucata*	*Pteria penguin*	*Crassostrea gigas*	*Crassostrea virginica*	*Ostrea edulis*
Total	26,315(−)	36,588(−)	36,733(−)	25,901(−)	34,596(−)	27,763(−)
SwissProt	13,819(52.51%)	16,947(46.32%)	13,877(37.78%)	14,845(57.31%)	18,525(53.55%)	15,209(54.78%)
KEGG	19,313(73.39%)	27,046(73.92%)	23,160(63.05%)	21,726(83.88%)	25,392(73.40%)	22,933(82.60%)
NR	21,937(83.36%)	29,532(80.71%)	24,922(67.85%)	25,717(99.30%)	34,557(99.89%)	27,422(98.77%)
GO	13,513(51.35%)	16,589(45.34%)	13,546(36.88%)	14,426(55.70%)	18,095(52.30%)	14,876(53.58%)
Pfam	16,992(64.57%)	22,277(60.89%)	17,907(48.75%)	20,138(77.75%)	25,695(74.27%)	20,767(74.80%)
Annotated	22,905(87.04%)	31,714(86.68%)	28,418(77.36%)	25,780(99.53%)	34,575(99.94%)	27,556(99.25%)

Table 3. The statistics of functional annotation for *Pinctada maxima* and five other bivalve species.

proportion of repeat sequences annotated was 65.46%, with DNA transposons accounting for the highest proportion (41.00%), followed by LTR (10.25%), LINE (7.52%) and SINE (0.16%) (Table 2).

Gene prediction and annotation. Protein-coding gene annotation was performed using the BRAKER v3.0.8[32] pipeline, which synergistically integrates multi-evidence approaches, including *de novo* prediction, homology-based searches and transcriptome-assisted methods. RNA-seq data generated by our study (SRP546131[33]) and published RNA-seq data (PRJNA362291,PRJNA636870 and PRJNA761869) were both used for *de novo* gene prediction. All the RNA-seq data were mapped to the soft masked genome using HISAT2 v2.2.1[34] with the alignment ranging from 93.93% to 98.17%. Then, BRAKER v3.0.8 and StringTie v2.2.1[35] were used to build transcript models of all mapping results. The transcript models were fed into AUGUSTUS v3.5.0[36] for gene model development and prediction. Homology-based annotation was conducted by using the amino acid sequences of *Pinctada fucata*[37], *Crassostrea gigas*[38], *Patinopecten yessoensis*[39], *Argopectens irradias*[40] and *Chlamys farreri*[41]. These amino acid sequences were aligned to the genome of *P. maxima* using TBLASTN with e-value threshold of 1e-10, and the aligned sequences were selected and provided to BRAKER v3.0.8. For the transcriptome-assisted annotation, Iso-seq data generated by our study was used to obtain full-length transcripts. The HiFi reads aligned were collapsed utilizing Isoseq. 3 v3.8.2 collapse pipeline (https://isoseq.how/classification/workflow.html) to remove the redundant transcripts resulting from 5′ RNA degradation. Then, the script gmst.pl from GeneMarkS v5.1[42] was used to predict the coding regions of the transcripts, and the prediction results were integrated using the script gmst2globalCoords.py from BRAKER v3.0.8. Finally, all evidences were merged to form a consensus gene set using TSEBRA v1.1.2.5[43] with parameters '--ignore_tx_phase -kl -f'. The weights of each part of the evidence are as follows: RNA-seq hints: 0.15; manual hints: 0.5; long reads hints: 0.5; protein hints: 3. In total, 26,315 protein-coding genes were identified (Table 3). NR, Pfam and SwissProt annotation of predicted protein-coding genes in *P. maxima* were performed by using BLASTP with e-value threshold of 1e-2. KEGG annotation was performed using KofamScan v1.3[44]. The GO annotation was obtained by mapping the annotation results from SwissProt. Finally, more than 87.04% (22,905) of protein-coding genes were annotated (Table 3). The results of functional annotations were displayed on the online platform Figshare (https://doi.org/10.6084/m9.figshare.28053659).

The non-coding RNA genes including rRNAs, tRNAs, snRNAs and miRNAs were screened using INFERNAL v1.1.2[45] and tRNAscan-SE v2.0.12[46]. Four types of noncoding RNAs, including 43 miRNAs, 4,042 tRNAs, 241 rRNAs and 609 snRNAs were identified in the *P.maxima* genome.

Comparative genome analysis. We conducted a systematic comparison with the four chromosomal-level Pteriidae genomes (*Pinctada fucata*, *Pinctada imbricata*, *Pteria penguin*, and *Pinctada radiata*) and four well-assembled oyster genomes (*Crassostrea gigas*, *Crassostrea virginica, Crassostrea hongkongensis* and *Ostrea edulis*) available on NCBI to outline the distinguishing features brought by our assembly. Our newly assembled *Pinctada maxima* genome exhibits superior scaffold-level contiguity compared to other species, with a scaffold

Database	Mollusca_odb10		Metazoa_odb10	
BUSCO	Data	Ratio	Data	Ratio
Complete BUSCOs (C)	90.59%	929	892	93.50%
Complete and single-copy BUSCOs (S)	89.07%	916	875	91.72%
Complete and duplicated BUSCOs (D)	1.53%	13	17	1.78%
Fragmented BUSCOs (F)	1.28%	2	12	1.26%
Missing BUSCOs (M)	8.12%	23	50	5.24%
Total BUSCO groups searched	5,295		954	

Table 4. BUSCO assessment the completeness of gene annotations.

N50 of 89.19 Mb comparable to Ostrea edulis (94.3 Mb) and surpassing the other seven genomes. Although contig N50 was lower than some species, possibly due to the large genome size, high proportion of repeat sequence combined with high heterozygosity, the high scaffold N50 demonstrates effective gap-closing during assembly. Our assembled genome shows excellent completeness, achieving 97.38% and 95.26% in Metazoa and Mollusca BUSCO assessments respectively, comparable to the available bivalve genomes (Table 1). The number of protein-coding genes in *P.maxima* (26,315) is less than the other two pearl oysters (36,588 and 36,733), while comparable to that of two oysters (25,901 and 27,763) (Table 3). Interestingly, the proportion of functionally annotated genes in oysters is higher than that in the pearl oysters (Table 3). The reason for the low annotation rate of genes in pearl oysters for functional annotation need further investigation.

The genome of *P.maxima* and 21 other species (*Acanthopleura granulate*, *Argopecten purpuratus*, *Bathymodiolus platifrons*, *Caenorhabditis elegans*, *Ciona intestinalis*, *Crassostrea gigas*, *Danio rerio*, *Drosophila melanogaster*, *Gallus gallus*, *Homo sapiens*, *Laticauda laticaudata*, *Lottia gigantea*, *Mus musculus*, *Octopus bimaculoides*, *Patinopecten yessoensis*, *Pictodentalium vernedei*, *Pinctada fucata*, *Pinctada imbricata*, *Scapharca broughtonii*, *Sinonovacula constricta*, *Xenopus tropicalis*) were used for gene family identification using OrthoFinder v2.5.5[47] with default parameters. Protein sequence alignment was executed using MUSCLE v3.8.31[48], following the alignment refinement conducted in GBLOCKS 0.91b[49] using stringent parameters (-b4 = 5 -b5 = h -t = p). The optimal amino acid substitution model (LG + I + G + F) was determined through ProtTest3 v3.4.2[50] prior to maximum likelihood tree construction in RAxML v8.2.12[51] with 1,000 bootstrap replicates. Divergence time estimation was performed using mcmctree in PAML[52]. Gene family contraction and expansion analysis was conducted in CAFE v5.0.0[53] using the result file generated by OrthoFinder. The constructed phylogenetic tree was visualized with the online interactive tool iTOL v7 (Interactive Tree Of Life) (https://itol.embl.de/). Syntenic genomic blocks between *P. maxima* and *P. fucata* were identified and visualized using MCScan implemented in jcvi v1.4.11[54] with the parameter--cscore = 0.99.

Data Records

The assembled genome has been deposited at GenBank under the accession JBLANZ000000000[55]. The raw Illumina PE150, PacBio, and Hi-C sequencing data have been deposited in Sequence Read Archive (SRA) with the accession number of SRP552859[56]. The raw RNA-seq sequencing and Iso-Seq sequencing data have been deposited in SRP546131[33], respectively. Assembled genome, functional annotation, and gene annotation files were uploaded to Figshare (https://doi.org/10.6084/m9.figshare.28053659)[57].

Technical Validation

QUAST v5.3.0[58] was employed to assess the genome assembly quality, focusing on its size and genome continuity. The total genome size was generated to be 1,264.93 Mb, with a contig N50 of 649 kb and a scaffold N50 length of 89.19 Mb (Table 1). Subsequently, we evaluated the completeness of the genome assembly using Benchmarking Universal Single-Copy Orthologs (BUSCO v5.8.1) with the metazoa_odb10 and mollusca_odb10 database. For metazoa_odb10 database, a total of 97.38% complete core genes were found with 96.02% as single-copy and 1.36% as duplicated genes (Table 1). The mollusca_odb10 database contains a total of 5,295 conserved core genes for mollusca, and our assembled genome included 5,044 (95.26%) of the expected mollusca genes with 4,964 (93.75%) as single-copy and 80 (1.51%) as duplicated genes (Table 1). We also used BUSCO to evaluate the completeness of gene annotations, observing 93.50% and 90.59% of the expected metazoa and mollusca genes, respectively (Table 4). Merqury v1.3[59] was used to evaluate the genome quality with PacBio HiFi reads, ultimately obtaining a consensus quality value (QV) of 55.64. In addition, Illumina paired-end clean reads and PacBio HiFi reads were mapped to the final reference genome assembly by BWA v0.7.18[60] and Minimap2 v2.1[61] to evaluated the genome assembly, observing the extremely high mapping rate with 98.89% and 99.99% for Illumina and PacBio sequencing. The high quality of the genome assembly is also demonstrated by the successful mapping of 95.39% ± 1.73% of transcriptome reads.

Molecular phylogenetic analysis identified a total of 35,646 orthogroups, of which 119 were single-copy orthogroups. The ortholog analyses revealed that 24,684 genes in *P.maxima* were clustered into orthogroups, with 894 genes belonging to species-specific orthogroups. Among the three pearl oysters (*P. maxima*, *P. fucata*, and *P. imbricata*), 1,332, 1,860, and 1,281 genes were assigned to *Pinctada*-specific orthogroups, respectively. The resulting ML topology incorporated 1,000 bootstrap replications for robust branch support evaluation. Phylogenetic analysis indicated the closest evolutionary relationship between *P. maxima* and *P. fucata*, with an estimated divergence time of approximately 90 million years ago. Furthermore, 10,479 gene families were identified as undergoing expansion or contraction events. Specifically, 231 expanded and 633 contracted gene

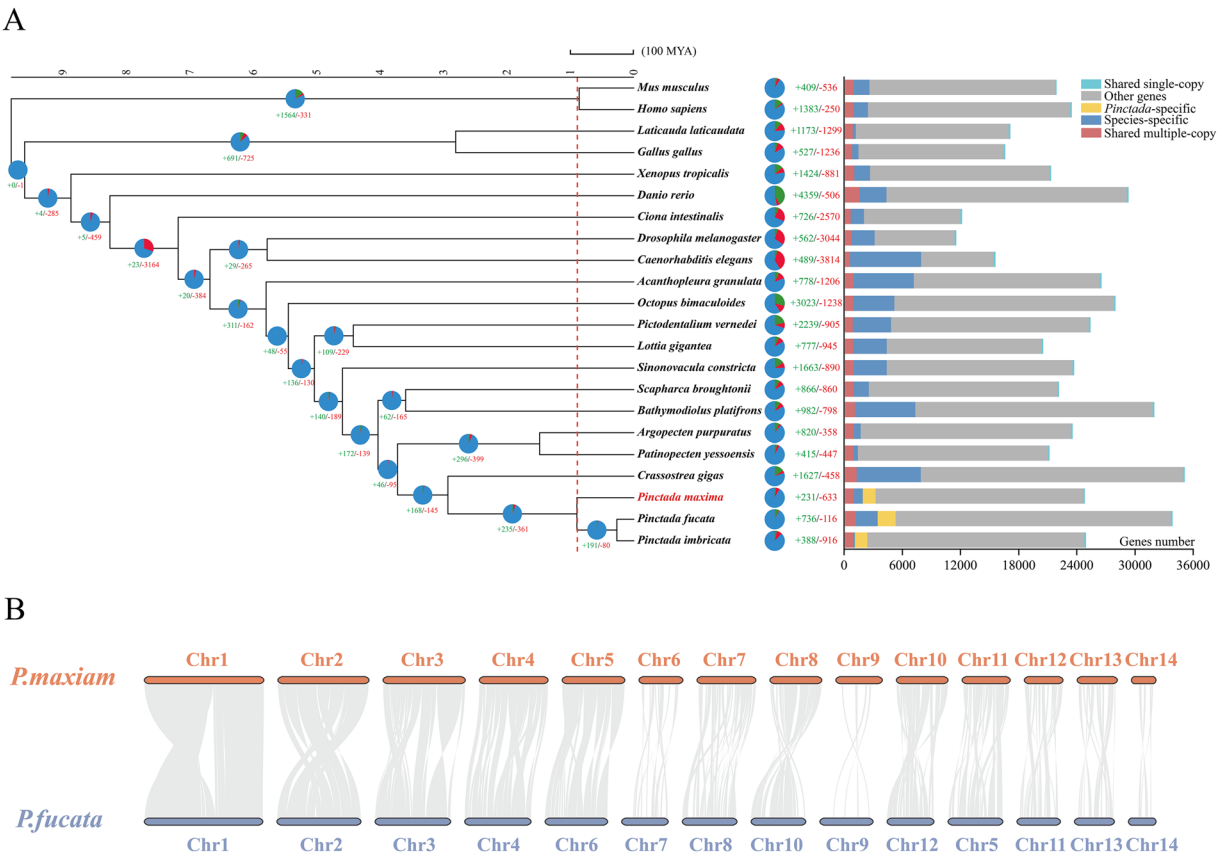

Fig. 2 Comparative genome analysis. (**A**) Maximum likelihood tree *P. maxima* and 21 other species with 1,000 bootstrap iterations. The scale bar represents 100 million years per unit. In the pie chart, red indicates contracted gene families, green represents expanded gene families, and blue denotes gene families that have neither contracted nor expanded. The stacked bar chart displays the number of genes in each species categorized as: Shared single-copy orthogroups, Shared multiple-copy orthogroups, Species-specific orthogroups, *Pinctada*-specific orthogroups, and Other genes. (**B**) Chord plot showing the chromosomal synteny between *P. maxima* and *P. fucata*.

families were observed in *P. maxima* (Fig. 2A). Subsequent statistical analysis ($p < 0.05$) identified 28 significantly expanded and 48 significantly contracted gene families in *P. maxima*. The collinearity analysis between *P. maxima* and *P. fucata* identified 17,191 highly matched genomic blocks pairs, with the genomes exhibiting complete one-to-one chromosomal synteny and no large-scale rearrangements (fission, fusion, or deletion) detected. The results suggested highly conserved genome synteny between *P. maxima* and *P. fucata* with generally one-to-one correspondence between their 14 chromosomes (Fig. 2B).

Code availability

This study did not utilize custom code for the curation and/or validation of the dataset. All commands and operational sequences executed during data processing were conducted in strict adherence to the guidelines and procedures delineated in the relevant bioinformatics software manuals and protocols. In cases where the software did not specify detailed parameters, the default parameters recommended by the developers were adopted.

Received: 22 February 2025; Accepted: 10 July 2025;
Published online: 26 July 2025

References

1. Vaughn, C. C. & Hoellein, T. J. Bivalve impacts in freshwater and marine ecosystems. *Annual review of ecology, evolution, and systematics* **49**(1), 183–208 (2018).
2. Adamkewicz, S. L., Harasewych, M. G., Blake, J., Saudek, D. & Bult, C. J. A molecular phylogeny of the bivalve mollusks. *Molecular biology and evolution* **14**(6), 619–629 (1997).
3. Yusof, A. M., Yanta, N. F. & Wood, A. K. H. The use of bivalves as bio-indicators in the assessment of marine pollution along a coastal area. *Journal of Radioanalytical and Nuclear Chemistry* **259**, 119–127 (2004).
4. Strehse, J. S. & Maser, E. Marine bivalves as bioindicators for environmental pollutants with focus on dumped munitions in the sea: A review. *Marine environmental research* **158**, 105006 (2020).
5. Aguirre-Rubi, J. R. *et al.* Prospective biomonitor and sentinel bivalve species for pollution monitoring and ecosystem health disturbance assessment in mangrove–lined Nicaraguan coasts. *Science of the Total Environment* **649**, 186–200 (2019).
6. FAO. *The State of World Fisheries and Aquaculture 2024 – Blue Transformation in action*. Rome (2024).
7. Southgate, P. C., Lucas, J. S. The pearl oyster. *Elsevier*, pp. 58—59 (2011).

8. Wang, Z., Liang, F., Huang, R., Deng, Y. & Li, J. Identification of the differentially expressed genes of *Pinctada maxima* individuals with different sizes through transcriptome analysis. *Regional Studies in Marine Science* **26**, 100512 (2019).
9. Hao, R. *et al*. Metabolomic responses of juvenile pearl oyster *Pinctada maxima* to different growth performances. *Aquaculture* **491**, 258–265 (2018).
10. Jones, D. B., Jerry, D. R., Khatkar, M. S., Raadsma, H. W. & Zenger, K. R. A high-density snp genetic linkage map for the silver-lipped pearl oyster, *Pinctada maxima*: a valuable resource for gene localisation and marker-assisted selection. *BMC Genomics* **14**(1), 1–18 (2013).
11. Wang, P. *et al*. Comparative proteomics reveal the humoral immune rejection of pearl oyster *Pinctada fucata* to xenograft from *Pinctada maxima*. *Aquaculture* **582**, 740515 (2024).
12. Deng, Y., Fu, S., Liang, F., Du, X. & Xie, S. Growth and survival of pearl oyster *Pinctada maxima* spat reared under different environmental conditions. *Journal of Shellfish Research* **32**(3), 675–679 (2013a).
13. Liang, M. *et al*. Transcriptome analysis provides novel insights into the factors influencing the settlement and metamorphosis of *Pinctada maxima*. *Aquaculture Reports* **39**, 102377 (2024).
14. Deng, Y., Fu, S., Liang, F. & Xie, S. Effects of stocking density, diet, and water exchange on growth and survival of pearl oyster *Pinctada maxima* larvae. *Aquaculture international* **21**, 1185–1194 (2013b).
15. Cheng, S. Y., Yu, D. H., Huang, G. J., Pan, L. L. & Wang, X. N. Intermediate culture of pearl oyster *Pinctada maxima* juveniles in deep waters. *Guangdong Agricultural Sciences* **38**(15), 102–104 (2011).
16. Gardner, L. D., Mills, D., Wiegand, A., Leavesley, D. & Elizur, A. Spatial analysis of biomineralization associated gene expression from the mantle organ of the pearl oyster *Pinctada maxima*. *BMC genomics* **12**, 1–16 (2011).
17. Wang, J., Zhang, L., Lian, S., Qin, Z., & Wang, S. Evolutionary transcriptomics of metazoan biphasic life cycle supports a single intercalation origin of metazoan larvae. *Nature Ecology & Evolution* (5), **4** (2020).
18. Moss, D. K. *et al*. Latitudinal life history gradients in two Pliocene species of *Glycymeris* (Bivalvia). *Historical Biology* 1−14 (2024).
19. Zhang, Q., Chen, J., Wang, W., Lin, J. & Guo, J. Genome-wide investigation of the TGF-β superfamily in scallops. *BMC genomics* **25**(1), 24 (2024).
20. Sambrook, J., Fritsch, E. F. & Maniatis, T. *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab Press, 1989).
21. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**(6), 764–770 (2011).
22. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nature communications* **11**(1), 1432 (2020).
23. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nature methods* **18**(2), 170–175 (2021).
24. 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).
25. Dudchenko, O. *et al*. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**(6333), 92–95 (2017).
26. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**(17), 9451–9457 (2020).
27. Ou, S. *et al*. Benchmarking Transposable Element Annotation Methods for Creation of a Streamlined, Comprehensive Pipeline. *Genome Biology* **20**(1), 275 (2019).
28. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current protocols in bioinformatics* **25**(1), 4–10 (2009).
29. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic acids research* **35**, W265–W268 (2007).
30. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC bioinformatics* **9**, 1–14 (2008).
31. Ou, S. & Jiang, N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant physiology* **176**(2), 1410–1422 (2018).
32. Gabriel, L. *et al*. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. *Genome Research* (2024).
33. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP546131 (2025).
34. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nature methods* **12**(4), 357–360 (2015).
35. Pertea, M. *et al*. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology* **33**(3), 290–295 (2015).
36. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**(5), 637–644 (2008).
37. Takeuchi, T. *et al*. A high-quality, haplotype-phased genome reconstruction reveals unexpected haplotype diversity in a pearl oyster. *DNA Research*. **29**(6) (2022).
38. Penaloza, C. *et al*. A chromosome-level genome assembly for the Pacific oyster *Crassostrea gigas*. *Gigascience* **10**(3) (2021).
39. Wang, S. *et al*. Scallop genome provides insights into evolution of bilaterian karyotype and development. *Nature Ecology & Evolution* **1**(5), 120 (2017).
40. Liu, X. *et al*. Draft genomes of two Atlantic bay scallop subspecies *Argopecten irradians irradians* and *A. i. concentricus*. *Scientific Data* **7**(1), 99 (2020).
41. Li, Y. *et al*. Scallop genome reveals molecular adaptations to semi-sessile life and neurotoxins. *Nature Communications* **8**(1), 1721 (2017).
42. Besemer, J., Lomsadze, A. & Borodovsky, M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic acids research* **29**(12), 2607–2618 (2001).
43. Gabriel, L., Hoff, K. J., Bruna, T., Borodovsky, M. & Stanke, M. TSEBRA: transcript selector for BRAKER. *Bmc Bioinformatics* **22**, 1–12 (2021).
44. Aramaki, T. *et al*. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**(7), 2251–2252 (2020).
45. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**(22), 2933–2935 (2013).
46. 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 research* **49**(16), 9077–9096 (2021).
47. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**(1), 238 (2019).
48. Edgar, R. C. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nature Communications* **13**, 6968 (2022).
49. Castresana, J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular biology and evolution* **17**(4), 540–552 (2000).

50. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics.* **27**(8), 1164–5 (2011).
51. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**(9), 1312–1313 (2014).
52. Yang, Z. PAML: a program package for phylogenetic analysis by maximum likelihood. *Bioinformatics* **13**, 555–556 (1997).
53. De Bie, T., Cristianini, N., Demuth, J. P. & Hahn, M. W. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* **22**(10), 1269–71 (2006).
54. Tang, H. *et al*. JCVI: A versatile toolkit for comparative genomics analysis. *Imeta* **3**(4), e211 (2024).
55. Mu, Q. Q. *Pinctada maxima* genome. *GenBank* https://identifiers.org/ncbi/insdc:JBLANZ000000000 (2025).
56. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP552859 (2025).
57. Mu, Q. Q. The chromosome-level genome assembly and annotation of the silver-lipped pearl oyster, *Pinctada maxima*. *figshare* https://doi.org/10.6084/m9.figshare.28053659 (2024).
58. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**(8), 1072–1075 (2013).
59. Rhie, A., Walenz, B. P., Koren, S. & Phillippy, A. M. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. *Genome Biology* **21**(1), 245 (2020).
60. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14), 1754–1760 (2009).
61. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**(18), 3094–3100 (2018).

Acknowledgements

This work was supported by Hainan Province Science and Technology Talent Innovation Project (KJRC2023A02), Project of Sanya Yazhouwan Science and Technology City Management Foundation (SKJC-KJ-2019KY01) and Sanya Science and Technology Special Fund (2022KJCX91).

Author contributions

P.L., Z.B. and J.H. conceived and designed the study. Z.B. and P.L. coordinated and supervised the whole study. M.L. performed sampling. Z.L., Q.M. carried out the experiments. Z.L., Q.M. and M.L. performed computational framework and analyzed the data. B.Z., Z.Y. participated in discussions and provided suggestions for manuscript improvement. Z.L., Q.M. and P.L. wrote the manuscript. P.L. and Z.B. revised the manuscript. The authors read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

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