scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A chromosome-level genome assembly of the *Bullacta exarata* (Cephalaspidea: Haminoeidae)

Yongchao Zhao[1], Yong Chi[2], Ailing Wu[2,3], Yiwo Wang[1], Duo Li[1], Yusheng Chen[1], Shuyin Chen[1], Hongjiu Ji[1] ✉ & Yangzhen Li[2] ✉

The mudsnail, *Bullacta exarata*, is an important marine food mollusk for both economic and aquaculture purposes. However, the scarcity of genomic information has hindered genetic research and breeding efforts. To address this gap, we developed a chromosome-level genome assembly utilizing short reads, HiFi long reads, and Hi-C sequencing data. The final assembly measures 836.77 Mb, with scaffold N50 and contig N50 values of 48.61 Mb and 1.00 Mb, respectively. Using Hi-C technology, 815.25 Mb (97.43%) of the contigs were anchored and arranged into 18 pseudochromosomes. A total of 358.66 Mb was identified as repeat elements, representing 42.86% of the genome. Furthermore, we predicted 85,914 non-coding RNAs and 17,996 protein-coding genes, with 93.48% of the protein-coding genes annotated. The BUSCO analysis indicated that the completeness of the genome assembly and annotation is 94.55% and 96.65%, respectively. Phylogenetic analysis clarified the evolutionary relationships between *B. exarata* and representative species in gastropoda. This high-quality reference genome for *B. exarata* serves as a valuable resource for aquaculture, fisheries, and ecological research.

Background & Summary

Bullacta exarata (Philippi, 1849), commonly known as mudsnail, is a marine herbivorous gastropod belonging to the order Cephalaspidea and family Haminoeidae[1]. This species primarily inhabits intertidal mudflats along the eastern coast of China, where it feeds predominantly on microalgae and detritus[2,3]. As an herbivorous gastropod, *B. exarata* plays a significant ecological role in coastal ecosystems. It exhibits gonochoristic reproductive traits characterized by external fertilization, and its life cycle includes planktonic larval stages, which are critical for dispersal and maintaining population connectivity across coastal regions[4].

Beyond its ecological importance, *B. exarata* has substantial economic value as a fishery resource in eastern China, with its aquaculture expanding steadily since artificial cultivation practices began in the 1980s[5]. This growth has prompted extensive research into diverse aspects of its biology, including reproductive biology[6], ecotoxicology[7,8], population dynamics[9,10], and spatial distribution patterns[11]. Furthermore, due to its sensitivity to environmental fluctuations, *B. exarata* is recognized as a valuable bioindicator, capable of reflecting coastal ecological health and pollution levels[12].

Conserving germplasm resources and preventing potential decline is crucial for safeguarding economically important species. Recently, the aquaculture of *B. exarata* has faced challenges, including overfishing, inadequate management, decreasing yields, and smaller individuals, which are linked to reduced genetic variation in cultured populations[13]. The scarcity of genetic data for this species hampers research on genetic variation and the development of breeding programs. Genomic data is vital for studying germplasm resources and evaluating the genetic structure and diversity of populations[14]. However, there is currently no research available on the genomic information of *B. exarata*. Advances in sequencing technologies, including long-read and Hi-C sequencing, along with the growing availability of extensive sequencing data, have accelerated the production of high-quality genome assemblies across numerous species[15,16]. Additionally, the advancement of various genome assembly algorithms has facilitated this process[17].

[1]Jiangsu Marine Fisheries Research Institute, Nantong, 226007, China. [2]State Key Laboratory of Mariculture Biobreeding and Sustainable Goods, Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao, 266071, China. [3]College of Fisheries and Life Science, Shanghai Ocean University, Shanghai, China. ✉e-mail: jijiu99@163.com; liyz@ysfri.ac.cn

Fig. 1 The adult individual of *B. exarata* used in this study.

Library type	Tissue	Raw data (Gb)	Clean data (Gb)	Coverage (×)
Illumina sequencing	Head	45.19	41.34	49.40
PacBio sequencing	Head	—	21.23	25.37
Hi-C	Head	97.10	90.77	108.48
RNA-seq sequencing	Head	5.49	5.39	—
	Foot	7.913	7.77	—

Table 1. Statistics of sequencing data used for *B. exarata* genome assembly in this study. The coverages of DNA-Seq data were calculated using clean data size divided by the genome size 836.77 Mb = 0.83677 Gb.

The availability of a high-quality genome for *B. exarata* will address key biological questions such as the genetic basis underlying adaptive responses to environmental stressors, mechanisms driving population declines in aquaculture, and the evolution of gastropod genomes through comparative genomic analyses. Furthermore, it will significantly advance genetic research and breeding programs by enabling the identification of markers associated with economically important traits, such as growth, stress resilience, and reproductive efficiency.

In the present study, we combined short reads, HiFi long reads, and Hi-C sequencing data to produce a chromosome-level genome assembly for for *B. exarata*, representing the first genomic resource for the family Haminoeidae. The length of the genome assembly was 836.77 Mb, with a scaffold N50 of 48.61 Mb and a contig N50 of 1.00 Mb. We identified 358.66 Mb of repetitive elements, representing 42.86% of the assembled genome, along with 17,996 protein-coding genes. This genome assembly and annotation will serve as a foundational resource for identifying genome-wide molecular markers, which are essential for future genetic improvement initiatives.

Methods

Sample collection. The *B. exarata* individual (total length 3.2 cm, Fig. 1) used for sample collection was collected from a natural population in the Nantong, Jiangsu, China, and identified based on morphological features. Fresh tissues, including foot and head trunks, were used for genome and transcriptome sequencing and analysis. All samples were flash-frozen in liquid nitrogen and stored at −80 °C for later extraction. Before sampling, this individual was anesthetized with Tricaine methanesulfonate (MS-222, 0.05 ppm). Genomic DNA was extracted from *B. exarata* tissues using the phenol-chloroform method.

Genomic DNA sequencing. An Illumina short read library was initially prepared and sequenced on the HiSeq X Ten platform, producing 150 bp paired-end short reads. A total of 45.19 Gb of raw Illumina short read data was generated, resulting in 41.34 Gb of clean data after removing adapters and low-quality reads using Fastp version 0.21.0 with default settings (Table 1). A PacBio HiFi-read library was generated and sequenced with a SMRT cell on the PacBio Sequel II platform, resulting in 21.23 Gb of clean data from the raw reads produced using the CCS model (Table 1). Using the standard protocol, the Hi-C library was prepared through *in situ* cross-linking of nuclear DNA, extraction, and digestion with the restriction enzyme DpnII. It was sequenced on the Illumina HiSeq X Ten platform by Genedenovo Biotechnology Co., Ltd. (Guangzhou, China) to generate paired-end 150 bp reads, yielding 97.10 Gb of raw data and 90.77 Gb of clean data after filtering (Table 1).

Transcriptome sequencing. Total RNA was extracted from two tissues (head and foot) using TRIzol reagent for transcriptome sequencing (on Illumina Novaseq 6000 platform by Genedenovo Biotechnology Co., Ltd., Guangzhou, China), resulting in 5.39 Gb and 7.77 Gb of clean RNA-seq data from the two tissues, respectively (Table 1).

Genome size and heterozygosity estimation. The frequency distribution of 21-mers from clean short reads was generated using Jellyfish v2.2.10[18]. This k-mer database was then analyzed with GenomeScope 2.0[19] to estimate various genome features, such as genome size, heterozygosity and repeat content. The estimated genome size is approximately 860.25 Mb, with a heterozygosity rate of about 1.98% (Fig. 2).

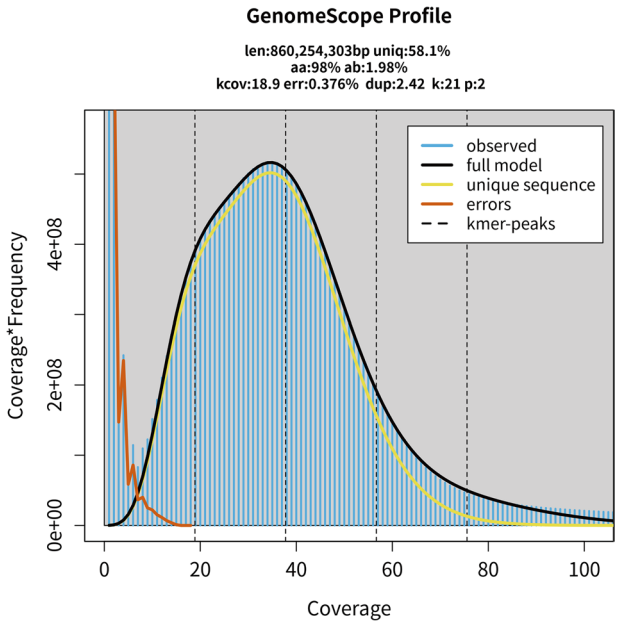

Fig. 2 Genome size estimation summary using k-mer (21-mer) analysis.

Key	Hifiasm-derived contigs	Hi-C scaffolded assembly
Total length	836,469,034	836,765,634
Contig/scaffold number	1614	139
Contig/scaffold N50	1,000,240	48,614,088
Average contig/scaffold length (bp)	518,258.39	6,019,896.65
Largest contig/scaffold length (bp)	4,703,038	87,836,962

Table 2. Information of *B. exarata* genome assembly based on the Hifiasm-derived contigs and Hi-C scaffolded assembly.

Genome assembly. HiFi reads were assembled into contigs using hifiasm v0.18.4[20], and Gfatools was used to convert the sequence graphs from GFA to FASTA format. Consequently, the *B. exarata* genome assembly has a total length of approximately 836.47 Mb, comprising 1,614 contigs, with a contig N50 of 1.00 Mb (Table 2). A multi-step strategy was implemented to assemble the *B. exarata* to the chromosome level using Hi-C reads. Initially, raw Hi-C data were filtered with Fastp, followed by mapping the filtered data to the assembled genome using Bowtie2 v2.4.5[21]. This process retained 90.77 Gb of clean data (Table 1). Next, HiC-Pro v3.1.0[22] was employed to identify and retain valid interacting paired-end reads from the uniquely mapped pairs, filtering out invalid sequences. YaHS v1.2a[23] was then used to assemble scaffolds into a chromosome-level genome, with manual adjustments made using Juicebox to finalize the pseudochromosomes[24]. The Hi-C-assisted chromosome-length scaffolds achieved a final size of 836.77 Mb (Table 2, Fig. 3A), anchored on 18 pseudo-chromosomes that represent 97.43% of the genome, with lengths ranging from 25.27 Mb to 87.84 Mb (Table 3; Fig. 3B).

Repetitive sequence annotation. To identify and mask repeated elements, we employed both homology-based and de novo methods. A de novo repeat sequence library was constructed using tools such as LTR_Finder[25], LTRharvest[26], DeepTE[27], HelitronScanner[28], MITE_Hunter_2[29], Tirvish[30] and RepLoc[31]. The RepeatMasker v4.1.5[32] and RepeatProteinMask v4.1.5[33] were then used to annotate both novel and known repeat sequences based on the de novo library and Repbase databases. Tandem repeats were identified using TRF v4.0.9[34]. The analysis revealed that the *B. exarata* genome consists of approximately 42.86% (358.66 Mb) repeat elements, including long terminal repeats (LTR, 22.06%), DNA transposons (9.62%), long interspersed nuclear elements (LINE, 5.26%), tandem repeats (16.86%), and other elements (Table 4; Fig. 4).

Protein-coding gene prediction and annotation. To annotate protein-coding genes in the *B. exarata* genome, we utilized three methods: de novo prediction, homology search, and transcript-based assembly. For the transcriptome-based method, RNA-seq data were aligned to the *B. exarata* genome using Hisat2 v2.2.1[35], followed by assembly with StringTie v2.2.1[36]. The homologous method involved comparing the *B. exarata* genome with the protein sequences of *Aplysia californica* using Miniprot v0.7[37]. Additionally, we incorporated protein sequences from the BUSCO and SwissProt databases as indirect homologous evidence. The de novo prediction method was carried out using Augustus v3.5.0[38] and SNAP[39]. The results from all three prediction methods were

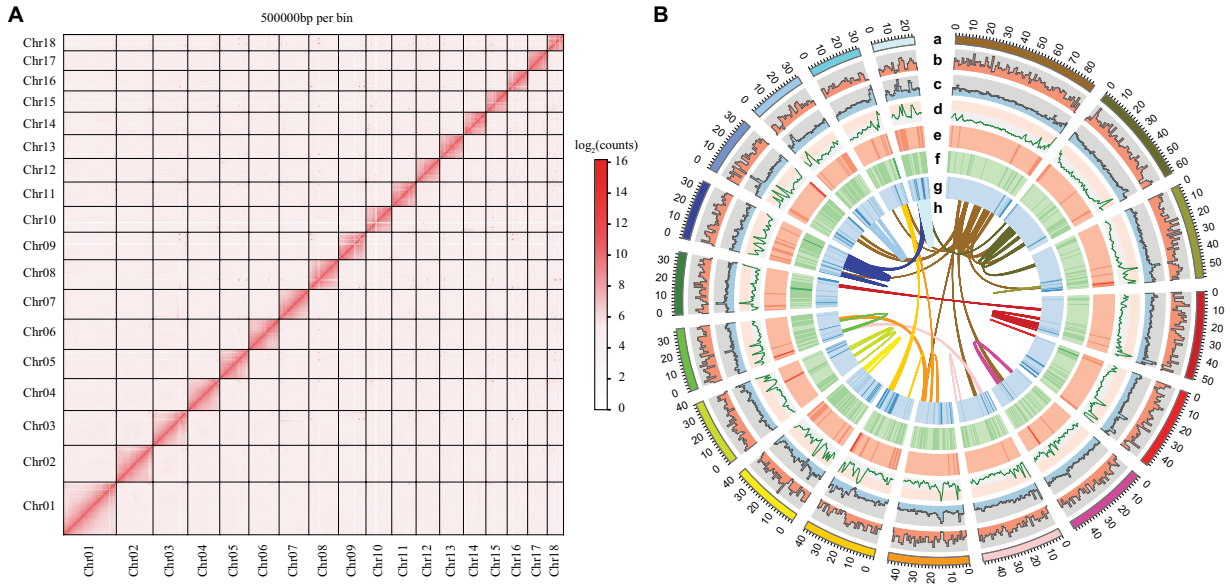

Fig. 3 The landscape of genome assembly and annotation of the *B. exarata*. (**A**) Hi-C interaction heat map of chromosomal level genome in *B. exarata*. (**B**) Chromosomal level genome feature distribution in *B. exarata*. Outer to inner tracks: (**a**) Length of chromosome, with each tick indicating a span of 1 Mb; (**b**) Coding gene density; (**c**) GC content distribution; (**d**) Repetitive sequence density; (**e**) DNA transposon density; (**f**) Density of Line elements; (**g**) LTR element density; (**h**) The interconnections within the circle represent syntenic gene pairs.

Chromosome	Size(bp)	Contig Number	GenBank AccNo
Chr01	8783692	118	CM106650.1
Chr02	60977552	92	CM106651.1
Chr03	56502489	104	CM106652.1
Chr04	52618390	87	CM106653.1
Chr05	49120446	59	CM106654.1
Chr06	48955948	108	CM106655.1
Chr07	48881058	78	CM106656.1
Chr08	48614088	82	CM106657.1
Chr09	44481240	81	CM106658.1
Chr10	41827734	80	CM106659.1
Chr11	40203854	70	CM106660.1
Chr12	37664760	79	CM106661.1
Chr13	37409015	86	CM106662.1
Chr14	35980102	71	CM106663.1
Chr15	34546647	73	CM106664.1
Chr16	32987537	73	CM106665.1
Chr17	31358973	68	CM106666.1
Chr18	25273163	81	CM106667.1
Total	815249958	1490	

Table 3. Statistics of 18 chromosomes of *B. exarata* genome.

combined using EvidenceModeler v.1.1.1[40]. A total of 17,996 protein-coding genes were identified by combining the results from these three approaches (Table 5).

For functional annotation, we utilized five databases: NCBI non-redundant proteins (NR), SwissProt, KOG, Gene Ontology (GO), and KEGG. The annotation was performed with DIAMOND v2.10[41] with a threshold for e-value $\leq$ 1e-5. In total, 16,822 genes (93.48% of the overall) were successfully annotated across the GO, KEGG, KOG, NR and Swissprot database (Fig. 4).

Noncoding RNAs annotation. Transfer RNA (tRNA) and ribosome RNA (rRNA) were predicted using tRNAscan-SE v2.0.9[42] and Barrnap v0.9. MicroRNAs (miRNA) and small nuclear RNA (snRNA) were identified

Type	RepeatMasker (%)	RepeatProteinMask (%)	Trf (%)	Combined (%)
DNA	79763009(9.5323)	885603(0.1058)	0	80468179(9.6166)
LTR	184206601(22.0141)	796202(0.0952)	0	184596070(22.0607)
LINE	41343978(4.9409)	5036349(0.6019)	0	44045743(5.2638)
SINE	0(0)	0(0)	0	0(0)
Simple_repeat	39245975(4.6902)	0(0)	0	39245975(4.6902)
Low_complexity	6951307(0.8307)	0(0)	0	6951307(0.8307)
Tandem Repeats	0(0)	0(0)	141106503(16.8633)	141106503(16.8633)
Unknown	16094704(1.9234)	0(0)	0	16094704(1.9234)
Total	319197160(38.1465)	6789794(0.8114)	141106503(16.8633)	358657165(42.8623)

Table 4. Statistical outcomes regarding repetitive sequences in the *B. exarata* genome.

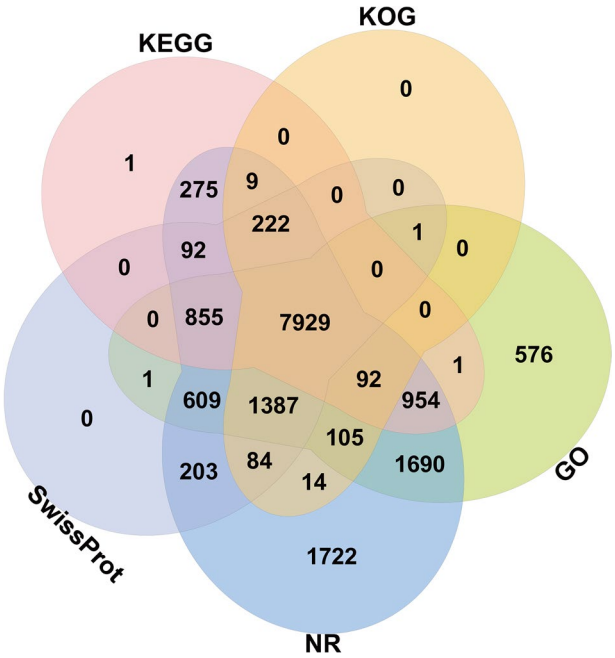

Fig. 4 Venn diagram illustrating the outcomes of functional annotation for *B. exarata* across five databases.

using Infernal Cmscan v 1.1.4[43] against the Rfam database[44]. In total, we predicted 84,428 tRNAs, 51 rRNAs, 241 snRNAs and 1194 miRNAs (Table 6).

Phylogenetic tree construction. Orthologous gene groups used for phylogenetic analysis were identified by performing pairwise comparisons of protein sequences from *B. exarata* and other 14 species, including *Candidula unifasciata*[45], *Biomphalaria glabrata*[46], *Elysia chlorotica*[47], *Pomacea canaliculata*[48], *Batillaria attramentaria*[49], *Patella vulgata*[50], *Lottia gigantea*[51], *Haliotis rufescens*[52], *Gigantopelta aegis*[53], *Pecten maximus*[54], *Mizuhopecten yessoensis*[55], *Crassostrea gigas*[56], *Octopus vulgaris*[57], and *Nautilus pompilius*[58], using DIAMOND BLASTP v2.1.11[41] (e-value $\leq 10^5$), followed by clustering into orthologous gene families using Proteinortho6[59]. Finally, a total of 1197 single-copy orthologous gene groups shared among these 15 species were selected for phylogenetic tree construction (maximum-likelihood based) using IQ-TREE v2.4.0[60]. The optimal substitution model for each gene was selected using ModelFinder[61]. Branch support was evaluated using UFBoot2[62] with 1,000 replicates. Divergence times were estimated using MCMCtree (PAML v4.9i)[63] under a Bayesian framework, with the tree rooted by *O. vulgaris* and *N. pompilius*, with root age at 603.7 million years ago (MYA).

Three calibration points were applied based on data from TimeTree (http://timetree.org/) and reference[48]: the divergence between *P. maximus* and *M. yessoensis* (220 MYA), *P. canaliculata* and *L. gigantea* (477.2 MYA), and *B. glabrata* and *P. canaliculata* (385.1 MYA). The phylogeny suggests that *B. exarata* (Cephalaspidea) diverged from *B. glabrata* (Stylommatophora) approximately 242.6 MYA (Fig. 5), highlighting their distinct evolutionary trajectories within Gastropoda, which originated 475.4 MYA.

Data Records

The genome assembly of *B. exarata* is available in GenBank under accession JBLOGN01[64]. The raw sequence data, as well as the assembly, reported in this paper has been deposited in the Sequence Read Archive under accession SRP553570[65]. Genome annotation files can be accessed on Figshare database[66].

Method	Software	Gene number	mRNA number	Average exon number
De novo	Augustus	24287	27045	8.61
	SNAP	30668	30668	7.67
Homology	Aplysia californica	10596	10596	7.60
	Swissprot	15174	15174	9.51
	Busco	572	572	8.83
RNA-seq	StringTie	17184	17184	10.51
Integration	EVidenceModeler	17429	17429	9.82
Final set	Mikado	17996	25060	11.26

Table 5. Statistical outcomes of gene structure annotation in the *B. exarata* genome, obtained through three different methods and subsequent integration.

Type	Number	Total length	Percentage (%)
miRNA	1194	192230	0.023
rRNA	51	30573	0.0037
snRNA	241	34877	0.0042
tRNA	84428	6255173	0.7475
Total	85914	6512853	0.7783

Table 6. Statistics for noncoding RNA genes in the genome of *B. exarata*.

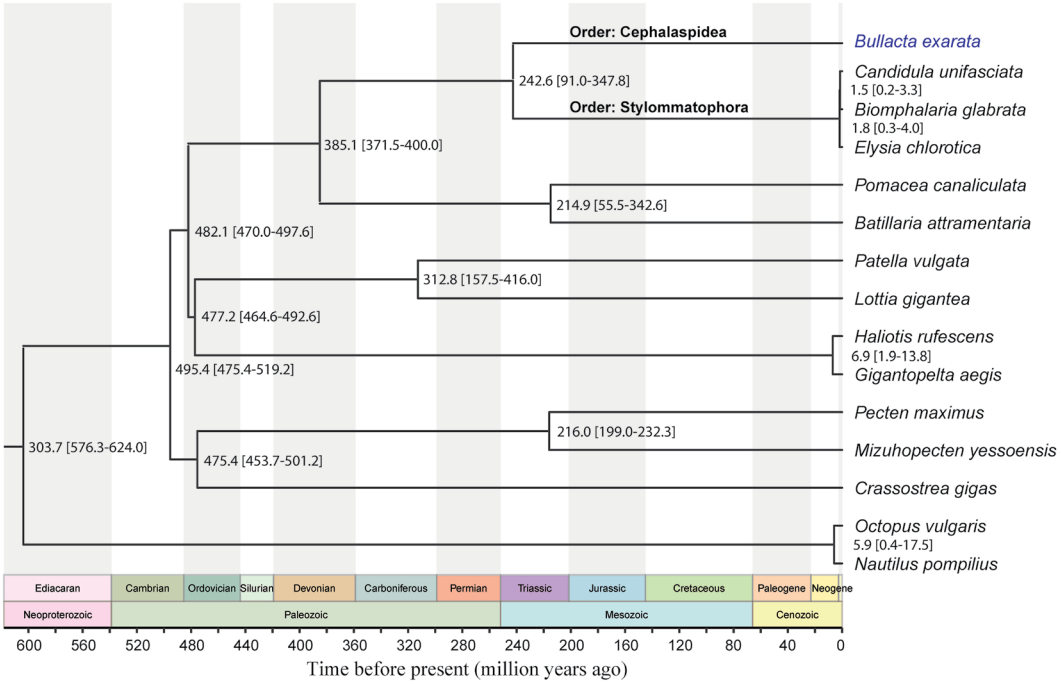

Fig. 5 Phylogenetic tree with estimated divergence times for *B. exarata* and other 14 species.

Technical Validation

To evaluate the accuracy and completeness of the *B. exarata* genome assembly, we performed a BUSCO assessment, identifying 902 complete BUSCO genes out of 954, which indicates a completeness of 94.55% (Table 7). For assembly accuracy, we calculated the mapping rates by aligning Illumina and PacBio reads to the final assembly using Samtools[67] and Minimap2[68], resulting in mapping rates of 98.71% and 98.86%, respectively. These assessments demonstrate the superior quality of genome assembly. Additionally, the BUSCO assessment for genome structure annotation revealed 922 complete genes out of 954 BUSCO groups, representing a completeness of 96.65% (Table 7).

Type	Assembly		Annotation	
	Number	Percent (%)	Number	Percent (%)
Complete BUSCOs (C)	902	94.55	922	96.65
Complete and single-copy BUSCOs (S)	877	91.93	896	93.92
Complete and duplicated BUSCOs (D)	25	2.62	26	2.73
Fragmented BUSCOs (F)	21	2.20	4	0.42
Missing BUSCOs (M)	31	3.25	28	2.94
Total BUSCO groups searched	954	100.00	954	100.00

Table 7. Statistics from the BUSCO assessment of the *B. exarata* genome assembly and annotation.

Code availability

All data processing commands and pipelines were executed following the manuals and protocols of the respective bioinformatics software.

Received: 27 February 2025; Accepted: 9 June 2025;
Published online: 16 June 2025

References

1. Malaquias, M. A. E. Systematics, phylogeny, and natural history of *Bullacta exarata* (Philippi, 1849): an endemic cephalaspidean gastropod from the China Sea. *J. Nat. Hist.* **44**, 2015–2029 (2010).
2. Ge, B. M., Bao, Y. X. & Cheng, H. Y. Tempo-spatial variation of macrobenthic communities on a tidal flat of Wenzhou Bay, China. *Rev. Biol. Marina Y Oceanogr.* **46**, 281–286 (2011).
3. Liu, J. *et al*. Epizoic Ulva attached to intertidal animals in the Subei intertidal zone are not the additional source of the famed Yellow Sea green tides. *J. Sea Res.* **174**, 102065 (2021).
4. Ge, B. M., Bao, Y. X., Cheng, H. Y., Zhang, D. Z. & Tang, B. P. Temporal and spatial distribution pattern of *Bullacta exarata* in a tidal flat at south shore of Hangzhou Bay, China. *Iran. J. Fish. Sci.* **12**, 96–104.
5. Ying, X. P., Yang, W. X., Jiang, N. C. & Zhang, Y. P. Ultrastructure of spermatozoa of *Bullacta exarata* (Philippi) and its significance on reproductive evolution and physio-ecological adaptation. *J. Zhejiang Univ. SCIENCE.* **5**, 1211–1217 (2004).
6. Zheng, H. P. preliminary study on characteristics of behavior and reproductive biology for *bullacta exarata* (philippi). *Mar. Sci.* **27**, 69–71 (2003).
7. Li, L. N., Chen, Z. L., Xu, S. Y. & Bi, C. J. Acute toxicity test on benthic fauna, *Bullacta exarata*, polluted by lead in the tidal flat of Changjiang estuary. *Trans. Oceanol. Limnol.* **0**, 88–92 (2005).
8. Li, B. Q. *et al*. Effects of key species mud snail *Bullacta exarata* (Gastropoda) on oxygen and nutrient fluxes at the sediment-water interface in the Huanghe River Delta, China. *Acta Oceanol. Sin.* **38**, 48–55 (2019).
9. Wang, S. L., Niu, D. H., Jia, M. J. & Li, J. L. Polymorphic microsatellite loci for population studies of *Bullacta exarata*. *Conserv. Genet.* **11**, 1231–1233 (2010).
10. Li, X. Y. *et al*. Isolation of fourteen novel polymorphic microsatellite markers for alien or invasive species: monitoring and management of the mud snail *Bullacta exarata* Philippi, 1849. *Conserv. Genet. Resour.* **6**, 259–261 (2014).
11. Wu, W. G., Yu, L., Zhang, J. H., Zhang, S. B. & Du, M. R. Preliminary study on distribution characteristics and living environment of *Bullacta exarata* population in Yellow River Estuary, China. *Prog. Fish. Sci.* **34**, 38–45 (2013).
12. An, L. H. *et al*. Biomarker responses and genotoxicity in the mud snail (*Bullacta exarata*) as indicators of coastal contamination. *Mar. Pollut. Bull.* **64**, 303–309 (2012).
13. Zhu, S. Y. *et al*. Artificial hybridization and juvenile nursery of *Bullacta exarata*. *J. Zhejiang Ocean Univ.* **1**, 32–36 (2008).
14. Breed, M. F. *et al*. The potential of genomics for restoring ecosystems and biodiversity. *Nat. Rev. Genet.* **20**, 615–628 (2019).
15. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).
16. Huddleston, J. *et al*. Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res.* **27**, 677–685 (2017).
17. Li, H. & Durbin, R. Genome assembly in the telomere-to-telomere era. *Nat. Rev. Genet.* **25**, 658–670 (2024).
18. Marçais, G. & Kingsford, C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764–770 (2011).
19. 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 (2020).
20. 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 (2021).
21. Langmead, B. & Salzberg, S. Fast gapped-read alignment with Bowtie 2. *Nat. Methods.* **9**, 357–359 (2012).
22. Servant, N. *et al*. HiC-Pro: An optimized and flexible pipeline for Hi-C processing. *Genome Biol.* **16**, 259 (2015).
23. Zhou, C. X., McCarthy, S. A. & Durbin, R. YaHS: yet another Hi-C scaffolding tool. *Bioinformatics* **139**, btac808 (2023).
24. Durand, N. C. *et al*. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Syst.* **3**, 99–101 (2016).
25. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
26. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
27. Yan, H. D., Bombarely, A. & Li, S. DeepTE: a computational method for de novo classification of transposons with convolutional neural network. *Bioinformatics* **36**, 4269–4275 (2020).
28. Xiong, W. W., He, L. M., Lai, J. S., Dooner, H. K. & Du, C. G. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proc. Natl. Acad. Sci.* **111**, 10263–10268 (2014).
29. Han, Y. J. & Wessler, S. R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* **38**, e199 (2010).
30. Gremme, G., Steinbiss, S. & Kurtz, S. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* **10**, 645–56 (2013).
31. Feng, C., Dao, M., Liu, Y. J. & Chen, M. Sequence repetitiveness quantification and de novo repeat detection by weighted k-mer coverage. *Brief. Bioinform.* **22**, bbaa086 (2021).

32. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Curr. Protoc. Bioinformatics*. Chapter 4, unit 4.10 (2009).
33. Tempel, S. Using and Understanding RepeatMasker. *Methods Mol. Biol.* **859**, 29–51 (2012).
34. Gary, B. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
35. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
36. Kovaka, S. *et al*. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol.* **20**, 1–13 (2019).
37. Li, H. Protein-to-genome alignment with miniprot. *Bioinformatics* **39**, btad014 (2023).
38. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **1**, 34 (2006).
39. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 1–9 (2004).
40. Haas, B. J., Salzberg, S. L., Zhu, W. & Pertea, M. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
41. Buchfink, B., Reuter, K. & Drost, H. G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).
42. 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 (2021).
43. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
44. Griffiths-Jones, S. *et al*. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* **33**, D121–D124 (2005).
45. Chueca, L. J., Schell, T. & Pfenninger, M. De novo genome assembly of the land snail *Candidula unifasciata* (Mollusca: Gastropoda). *G3* **11**(8), jkab180 (2021).
46. Adema, C. M. *et al*. Whole genome analysis of a schistosomiasis-transmitting freshwater snail. *Nat. Commun.* **8**, 15451 (2017).
47. Cai, H. *et al*. A draft genome assembly of the solar-powered sea slug *Elysia chlorotica*. *Sci. Data* **6**(1), 1–13 (2019).
48. Liu, C. *et al*. The genome of the golden apple snail *Pomacea canaliculata* provides insight into stress tolerance and invasive adaptation. *Gigascience* **7**(9), giy101 (2018).
49. Patra, A. K. *et al*. Genome assembly of the Korean intertidal mud-creeper *Batillaria attramentaria*. *Sci. Data* **10**(1), 498 (2023).
50. Hawkins, S. J. *et al*. The genome sequence of the common limpet, *Patella vulgata* (Linnaeus, 1758). *Wellcome Open Res.* **8**, 418.
51. Simakov, O. *et al*. Insights into bilaterian evolution from three spiralian genomes. *Nature* **493**(7433), 526–531 (2013).
52. Griffiths, J. S. *et al*. A draft reference genome of the red abalone, *Haliotis rufescens*, for conservation genomics. *J. Hered.* **113**(6), 673–680 (2022).
53. Lan, Y. *et al*. Hologenome analysis reveals dual symbiosis in the deep-sea hydrothermal vent snail *Gigantopelta aegis*. *Nat. Commun.* **12**(1), 1165 (2021).
54. Kenny, N. J. *et al*. The gene-rich genome of the scallop *Pecten maximus*. *GigaScience* **9**(5), giaa037 (2020).
55. Wang, S. *et al*. Scallop genome provides insights into evolution of bilaterian karyotype and development. *Nat. Ecol. Evo.* **1**(5), 0120 (2017).
56. Penaloza, C. *et al*. A chromosome-level genome assembly for the Pacific oyster *Crassostrea gigas*. *GigaScience* **10**(3), giab020 (2021).
57. Destanović, D. *et al*. A chromosome-level reference genome for the common octopus, *Octopus vulgaris* (Cuvier, 1797). *G3* **13**(12), jkad220 (2023).
58. Zhang, Y. *et al*. The genome of *Nautilus pompilius* illuminates eye evolution and biomineralization. *Nat. Ecol. Evo.* **5**(7), 927–938 (2021).
59. Lechner, M. *et al*. Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC bioinformatics* **12**, 1–9 (2011).
60. Minh, B. Q. *et al*. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**(5), 1530–1534 (2020).
61. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K., Von Haeseler, A. & Jermiin, L. S. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat. methods* **14**(6), 587–589 (2017).
62. Hoang, D. T., Chernomor, O., Von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**(2), 518–522 (2018).
63. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**(8), 1586–1591 (2007).
64. Li, Y. *Bullacta exarata* isolate mudsnail3, whole genome shotgun sequencing project. GenBank. https://identifiers.org/ncbi/insdc:JBLOGN000000000.1 (2025).
65. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP553570 (2025).
66. Li, Y. Genome annotation of *Bullacta exarata*. *Figshare*. https://doi.org/10.6084/m9.figshare.28456985.v2 (2025).
67. Danecek, P. *et al*. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
68. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).

Acknowledgements

This study was supported by China Agriculture Research System (CARS-49), Jiangsu Natural Resources Development Special Fund (Marine Science and Technology Innovation) project (JSZRHYKJ202109), and Jiangsu Agricultural Science and technology innovation and promotion (2023-SJ-026).

Author contributions

Y.L. and H.J. conceived this study. Y.Z., Y.C. and A.W. performed data analyses. Y.Z., Y.W., D.L., Y.C. and S.C. collected the samples. Y.Z., Y.C., A.W., H.J. and Y.L. wrote and revised the manuscript. All authors have read and approved the final manuscript for publication.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to H.J. or Y.L.