scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A chromosome-level genome assembly of *Macrobrachium hainanense*

Chao Bian[1,2], Rujingwen Huan[1], Yiting Jin[2], Qibin Yang[3], Weihong Zhao[4], Na Zhou[5,6] ✉ & Shengming Sun[2,6] ✉

***Macrobrachium hainanense*, an anadromous crustacean species within the genus *Macrobrachium*, is predominantly distributed along the southern coasts of China. This species is regarded as an important economic prawn. However, overfishing and breeding difficulties have caused a severe decline in the natural population of *M. hainanense*. Genomic data can provide valuable insights into the species genetic basis and evolutionary adaptations, thereby supporting its sustainable development and utilization. In this study, we generated a high-quality chromosome-level genome assembly of *M. hainanense* by combining PacBio HiFi sequencing and Hi-C technologies. This genome assembly, with assembled size of 4.3 Gb and a scaffold N50 of 87.0 Mb, was successfully anchored to 51 chromosomes. 51 centromeres and 97 telomeres were identified from 51 chromosome sequences. The BUSCO analysis revealed 93.8% completeness and 92.2% single-copy genes. A total of 46,945 protein-coding genes were annotated. This genome assembly provides a valuable resource for understanding the genetic makeup of *M. hainanense*, and it will aid in future breeding programs and hybridization studies.**

Introduction

Crustaceans play a crucial role in global fisheries and aquatic ecosystems, which not only have significant ecological importance, but also play a key role in economic activities[1]. In fact, the prawn genus *Macrobrachium* (Palaemonidae) was one of the most diversified and widely distributed decapods groups[2], although the majority of species require brackish or sea water to complete their larval development, the genus Macrobrachium are referred to as 'freshwater' prawns[3]. The differential growth of the larger second claw eventually leads to the development of the well known sexual dimorphism, which is a prominent feature of Macrobrachium with females of species with abbreviated larval development maturing at a smaller size[4].

Among prawn genus *Macrobrachium, Macrobrachium hainanense* (Parisi, 1919) is an anadromous migratory crustacean[5,6], which is primarily found along the coasts of southern China according to the Global Biodiversity Information Facility (GBIF) information (https://www.gbif.org/species/5863221)[7]. The larvae of this species develop and metamorphose in estuarine semi-saline environments before migrating to freshwater habitats for growth[8], and the size (6–10 cm) of mature individuals is larger than that of the *M. nipponense* (3–8 cm), however, which was smaller than that of *M. rosenbergii* (15–20 cm). Interestingly, *M. hainanense* is capable of inhabiting temperatures as low as 4 °C and as high as 40 °C[9]. *M. hainanense* has become an important economic prawn due to its excellent biological characteristics such as large individuals, fast growth, and low temperature tolerance[9]. However, problems such as overfishing wild populations and breeding difficulties have led to a significant decrease in its natural population size and put its germplasm resources at risk of decline[10]. Therefore, the absence

[1]Laboratory of Aquatic Genomics, College of Life Sciences and Oceanography, Shenzhen University, Shenzhen, 518057, China. [2]Key Laboratory of Exploration and Utilization of Aquatic Genetic Resources, Ministry of Education, Shanghai Ocean University, Shanghai, 201306, China. [3]Tropical Aquaculture Research and Development Center, South China Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Sanya, 572018, PR China. [4]School of Marine and Bioengineering, Yancheng Institute of Technology, Yancheng, 224051, China. [5]School of Pharmacy and State Key Laboratory for Quality Research of Chinese Medicines, (R & D Center) Lab. for Drug Discovery from Natural Resource, Macau University of Science and Technology, Taipa, Macau, China. [6]Joint Laboratory of Aquatic Biosynthesis and Green Prevention and Control (Zhejiang-Shanghai-Macao), Macau University of Science and Technology, Taipa, Macau, China. ✉e-mail: nzhou@must.edu.mo; sunshengming621416@163.com

	M. hainanense	*M. nipponense*	*E. carinicauda*
HiFi reads (Gb)	228.0	405.7 (Pacbio sequel)	203.3
Hi-C reads (Gb)	464.1	876.4	552.7
Genome size (Gb)	4.3	4.5	5.9
Contig N50 (kb)	1,176.8	231.2	235.5
BUSCO	93.8%	92.6%	92.9%
Gap length (Mb)	4.65	17.80	/
Repeat ratio	54.7%	54.6%	84.7%
Gene number	46,945	44,086	44,288
Average gene length (bp)	27,603.6	14,343.0	28,448.0
Functional gene number	38,274	39,317	31,238

Table 1. Statistics of the assembly and annotation results of *M. hainanense, M. nipponense* and *E. carinicauda*.

of genomic data limits in-depth study of the genetic basis of *M. hainanense* and its evolution adaptation for promoting its sustainable development and utilization.

In this study, we conducted the preliminary assembly of *M. hainanense* genome by using the Pacific Biosciences (PacBio) sequencing platform. Subsequently, we employed high-throughput chromosome conformation capture (Hi-C) technology to further scaffold the preliminary assembly. The assembled genome size is 4.3 Gb with a scaffold N50 of 87.0 Mb. These assembled sequences were anchored to 51 pseudo-chromosomes. A total of 46,945 protein-coding genes were annotated. This chromosome-level genome assembly serves as an important resource for gaining the in-depth understanding of the genome information related to *M. hainanense* as well as offering new resources for breeding programs and artificial interspecific hybridization between Macrobrachium species.

Methods

Sample collection and DNA extraction. A male *M. hainanense* was collected from Hainan province, China (110°10′N, 20°03′E). The total body length of this individual is 8.2 cm, and the body weight of this sample is 16.5 g. The muscle tissue from the sample was immediately flash-frozen in liquid nitrogen following dissection to preserve sample integrity for subsequent DNA extraction. Genomic DNA was isolated using the TIANamp Genomic DNA Kit (TIANGEN, Beijing, China) according to the manufacturer's protocol. The concentration of the DNA was determined by using the Qubit 2.0 Fluorometer (Life Technologies, USA), and the size of the extracted DNA was confirmed through 1% agarose gel electrophoresis.

Library construction and sequencing. For PacBio HiFi sequencing, six long-read libraries were constructed using the SMRTbell Express Template Prep Kit 2.0 for PacBio HiFi sequencing on the PacBio HiFi platform with PacBio's standard protocol (Pacific Biosciences, USA). All consensus sequences were generated using the CCS software (SMRT Link v9.0)[11]. Approximately 228.0 Gb of PacBio HiFi reads were produced (Table 1).

For Hi-C sequencing, the Hi-C libraries were prepared using the GrandOmics Hi-C kit and DpnII restriction enzyme (GrandOmics, China), following the standard protocol provided by the manufacturer. Sequencing was performed on the Illumina NovaSeq platform (Illumina, USA), which generated approximately 464.1 Gb of 150 bp paired-end reads. These Hi-C reads were employed for genome assembly and chromosome anchoring in *M. hainanense*.

For RNA sequencing, total RNA was extracted from mixed multiple tissues, including heart, hepatopancreas, gill, and muscle by using the TRIZOL Kit (Invitrogen, Carlsbad, CA, USA) in accordance with the manufacturer's instructions. RNA integrity and quality were assessed using the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA), and only samples with RIN (RNA Integrity Number) greater than 7.0 were selected for subsequent library preparation. The cDNA libraries were constructed using DNA nanoballs (DNBs) in accordance with the manufacturer's protocol for the DNBSEQ platform. Sequencing was conducted on the MGISEQ-2000 platform (MGI, BGI Shenzhen, China) to generate 150 bp paired-end reads.

Genome assembly and quality evaluation. The initial genome assembly was generated using hifiasm (v0.19.8, parameters: -l 3)[12]. This step resulted in the assembly of a graph representing the *M. hainanense* genome sample with a total size of 8.2 Gb and the detailed contigs N50 value of 1.0 Mb. The sequenced Hi-C reads were aligned onto above assembled contigs of samples by employing Bowtie2 (v2.2.5, parameters:–very-sensitive -L 30–score-min L,-0.6,-0.2–end-to-end –reorder)[13]. Whole valid pair information of chromosome linkage was calculated by HiC-Pro (v2.8.0)[14] with default parameters based on the alignment results. This information was then used to anchor contigs into primary chromosomes by using Juicer (v1.5)[15] and 3d-DNA (v170123, parameters: -m haploid -r 2)[16]. Further adjustments were made using Juicebox (v1.11.08)[17] to rectify mis-assemblies and remove redundancies.

The final assembly anchored to 51 chromosomes with the scaffold N50 of 87.0 Mb (Figs. 1, 2). In these 51 chromosomes, about 4.65 Mb gaps were contained. The completeness of this chromosome assembly was assessed using BUSCO, with 93.8% of complete and 92.2% of single-copy BUSCOs identified, indicating high assembly accuracy and completeness (Table 1).

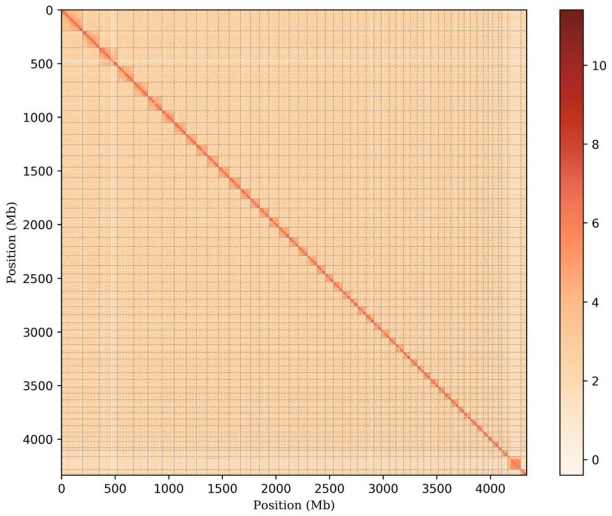

Fig. 1 The heatmap of *M. hainanense* genome assembly.

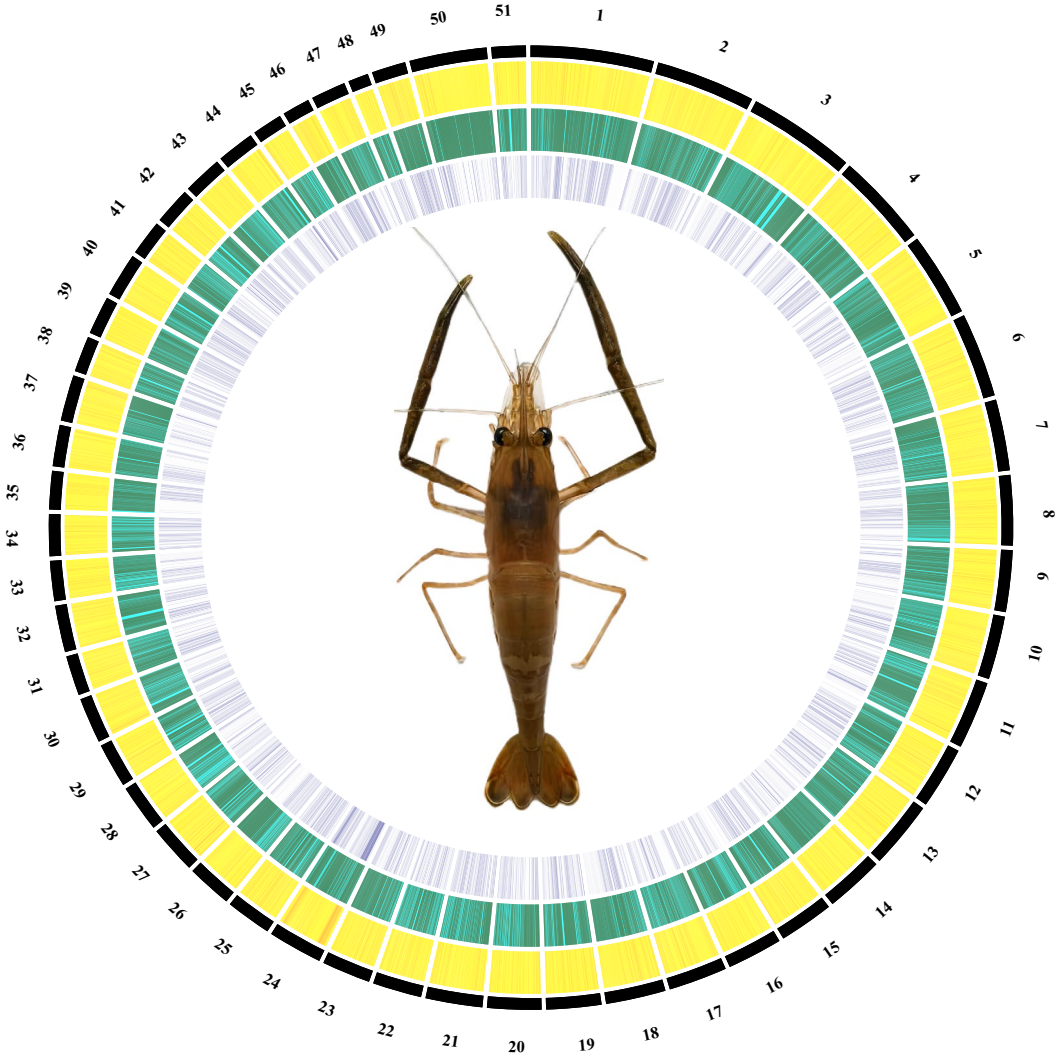

Fig. 2 Circos figure of 51 anchored chromosomes for *M. hainanense* genome. From the outside to the inside, the length and number of 51 chromosomes, gene density distribution (bin = 1 Mb), repetitive sequence density (bin = 1 Mb), GC content (bin = 1 Mb) and *M. hainanense* view.

Type	Repbase TEs		TE protiens		*De novo*		Combined TEs	
	Length (Mb)	% in genome	Length (Mb)	% in genome	Length (Mb)	% in genome	Length (Mb)	% in genome
DNA	338.8	7.7	43.2	1.0	562.9	12.9	766.0	17.5
LINE	148.5	3.4	163.1	3.7	391.0	8.9	637.5	14.6
SINE	1.6	0.037	0	0	1.3	0.03	2.9	0.07
LTR	120.9	2.8	149.4	3.4	392.3	9.0	619.1	14.1
Other	0.6	0.01	0	0	0	0	0.6	0.01
Unknown	0	0	0	0	818.3	18.7	818.3	18.7
Total	494.2	11.3	355.1	8.1	1950.3	44.5	2396.8	54.7

Table 2. Statistics of the repeat sequences of *M. hainanense*.

Repeat element annotation. Repetitive elements in the anchored chromosome-level genome assemblies were identified using both *de novo* and homology-based approaches. For *de novo* identification, RepeatModeler (v1.0.8)[18] was used to construct a *de novo* repeat library, and LTR-FINDER (v1.0.6)[19] was applied specifically to identify long terminal repeat (LTR) retrotransposons. A custom repeat library was then generated using RepeatMasker (v4.0.623)[20] in combination with the Repbase TE library (v21.01)[21]. Tandem repeats were detected using Tandem Repeats Finder (v4.07, parameters: 2 7 7 80 10 50 2000 -d -h)[22]. The custom repeat library from *de novo* prediction, along with known repeat elements from Repbase, was used to annotate the genome using RepeatMasker (v4.0.623)[20] and RepeatProteinMask (v4.0.623)[20] with default parameters. The centromere and telomere sequences were identified by the quarTeT software (v1.1.7)[23]. In total, approximately 2.4 Gb of repetitive sequences were identified, accounting for 54.7% of the genome. A total of 97 telomeres were predicted (Table 2). Among them, 92 telomeres were contained in both ends of 46 chromosomes. Entire 51 chromosomes contained the centromeres (Fig. 3).

Gene prediction and functional annotation. Gene prediction was performed using a combination of homology-based and transcriptome-based methods. Protein sequences from three related species, including *Penaeus vannamei* (GCF_042767895.1), *Daphnia pulex* (GCF_021134715.1), *and M. nipponense* (GCF_015104395.2), were downloaded from the NCBI database and aligned to the assembled genome using TBLASTn (v2.14.7, e-value 10^{-5})[24] for homology-based annotation. Gene structures were further refined using GeneWise (v2.2.0)[25](parameters:–blast_eval 1e-5–align_rate 0.5-extend_len 500) based on the TBLASTn alignments. Concurrently, RNA-seq reads were mapped to the genome using HISAT (v2.2.1)[26], and transcript assemblies were generated using Cufflinks (v2.2.1, http://cole-trapnell-lab.github.io/cufflinks/). The final non-redundant gene set was produced by integrating the homology- and transcriptome-based predictions with MAKER (v3.01.2-beta)[27] (parameters: max_dna_len = 300000, min_contig = 500, pred_flank = 500, AED_threshold = 1, split_hit = 30000, single_exon = 1, single_length = 250, tries = 2). In total, 46,945 protein-coding genes were predicted, with an average mRNA_length of 27603.6 bp and 5.0 exons per gene. Functional annotations were assigned by aligning predicted genes to the SwissProt[28], TrEMBL[29], and KEGG[30] databases, while Gene Ontology (GO) terms were added using InterProScan (v5.11- 55.0)[31].

Data Records

The chromosome-level genome assembly of *M. hainanense* have been deposited in the Genome Warehouse[32] in National Genomics Data Center with accession numbers: PRJCA047697[33]. The raw reads and genome assembly of *M. hainanense* are available at NCBI with accession numbers: PRJNA1172361[34,35]. The annotation coding sequences and protein sequences were deposited at Figshare with doi number 10.6084/m9.figshare.27600606[36]. The raw reads of PacBio HiFi and MGI sequencing were also uploaded at the NCBI with accession numbers SRR31037707[37] and SRR31037706[38]. The mixed RNA-seq reads have been deposited to NCBI with accession number: SRR33222360[39].

Technical Validation

We constructed 51 chromosomes of *M. hainanense* by using high-quality Hi-C assemblies. The HiFi reads were aligned onto the genome assembly by using Minimap2. The mapping ratio is 99.97%. We also performed the assessment of the completeness of the genome assembly by using Benchmarking Universal Single-Copy Orthologs (BUSCO) v.5.2.2 with the Arthropoda reference (1,013 single-copy orthologs; arthropoda_odb10). A total of 93.8% (950) BUSCOs were identified as complete. About 92.2% (934) BUSCOs were single-copy, 1.6% (16) were duplicated, and 3.1% (31) were fragmented. Taken together, these results confirmed the high quality of this genome assembly.

Data availability

The chromosome-level assembled genome is available under CNCB[32] BioProject PRJCA047697 (https://ngdc.cncb.ac.cn/gwh/Assembly/102277/show)[33]. All raw reads in this study and the genome assembly are available under NCBI BioProject PRJNA1172361[34,35], which includes PacBio HiFi reads (SRR31037707)[37], MGI short reads (SRR31037706)[38] and the mixed RNA-seq reads (SRR33222360)[39]. The annotation coding sequences and protein sequences are available at Figshare with doi number https://doi.org/10.6084/m9.figshare.27600606[36].

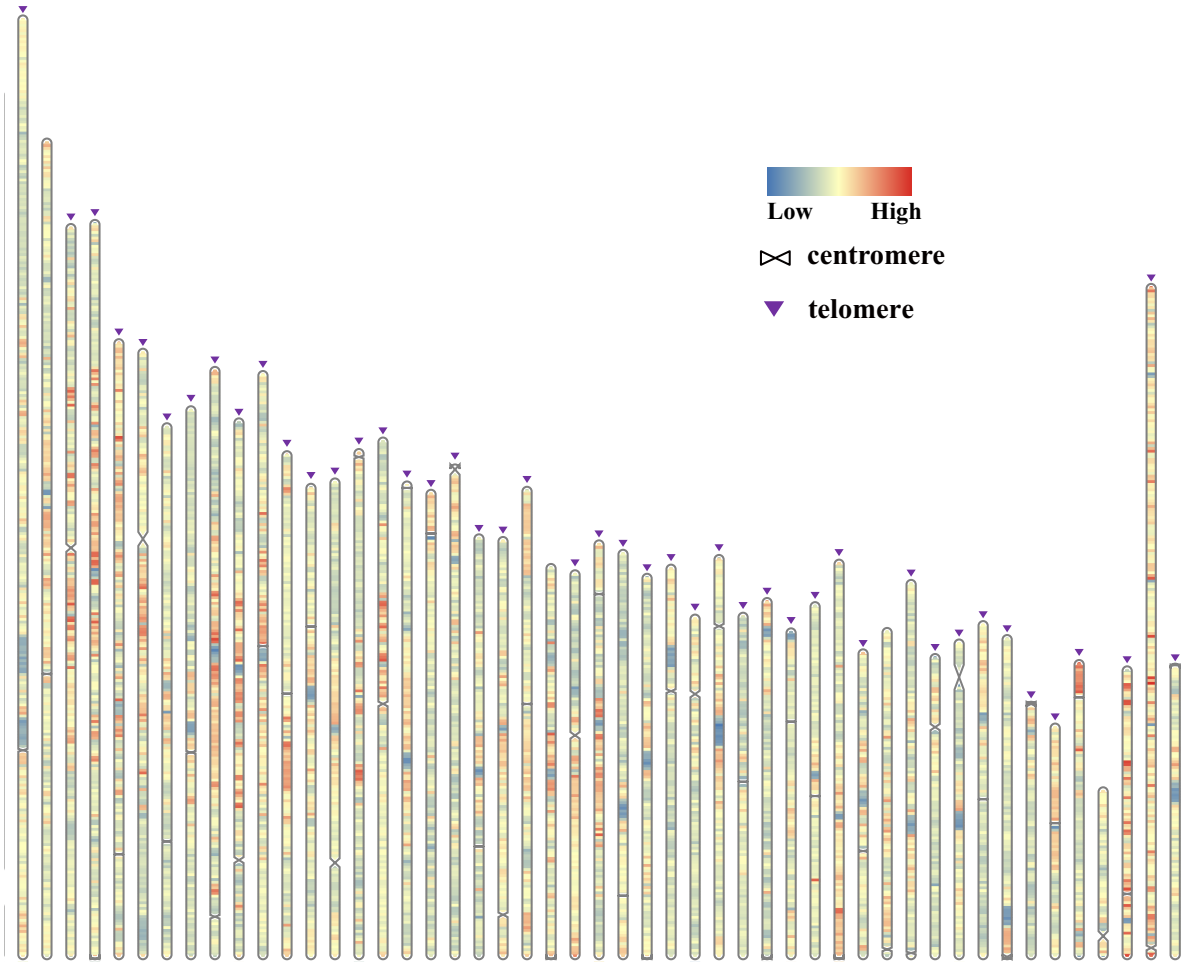

Fig. 3 Repeat distribution, telomere and centromere locations in *M. hainanense* genome.

Code availability

All software, pipelines and scripts used for the genome assembly and gene annotation were performed according to their manuals and protocols of the applied bioinformatics software. No specific code has been developed for this study.

Received: 13 November 2024; Accepted: 22 October 2025;
Published online: 05 December 2025

References

1. Behringer, D. & Duermit, E. Crustaceans, one health and the changing ocean. *Journal of Invertebrate Pathology* **186**, 107500 (2020).
2. Rahi, L. *et al.* Distributed under Creative Commons CC-BY 4.0 Candidate genes that have facilitated freshwater adaptation by palaemonid prawns in the genus *Macrobrachium*: identification and expression validation in a model species (*M. koombooloomba*). *Peer J* **5**, e2977 (2017).
3. Rahi, M. *et al.* The Molecular basis of freshwater adaptation in prawns: insights from comparative transcriptomics of three *Macrobrachium* Species. *Genome Biology and Evolution* **11**, 1002–1018 (2019).
4. Mossolin, E. C. & Bueno, S. L. S. Relative growth of the second pereiopod in *Macrobrachium olfersi* (Wiegmann, 1836) (Decapoda, Palaemonidae). *Crustaceana* **76**, 363–376 (2003).
5. Wong, J. T. Y. Abbreviated larval development of *Macrobrachium hainanense* (Parisi, 1919) reared in the laboratory (Decapoda, Caridea, Palaemonidae). *Crustaceana* **56**, 18–30 (1989).
6. Cai, Y. *et al.* The freshwater prawns of the genus *Macrobrachium* Bate, 1868, of Thailand (Crustacea: Decapoda: Palaemonidae). *Journal of Natural History* **38**, 581–640 (2004).
7. Mantel, S. K. & Dudgeon, D. Effects of *Macrobrachium hainanense* predation on benthic community functioning in tropical Asian streams. *Freshwater Biology* **49**, 1306–1319 (2004).
8. Wang, K. W. *et al.* Preliminary study on the biology and artificial cultivation of *Macrobrachium hainanense* in Oujiang River. *Freshwater Fisheries* **5**, 9–11 (1990).

9. Fu, H. *et al*. Artificial interspecific hybridization between *Macrobrachium* species. *Aquaculture* **232**, 215–223 (2004).
10. Wang, K. W. *et al*. Discussion on the development and utilization of *Macrobrachium hainanense* seed resources in the middle and lower reaches of Oujiang River. *Journal of Hydroecology* **5**, 33–35 (1992).
11. Rhoads, A. & Au, K. F. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics* **13**, 278–289 (2015).
12. Cheng, H. *et al*. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nature Methods* **18**, 1–6 (2021).
13. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357 (2012).
14. Servant, N. *et al*. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology* **16**, 259 (2015).
15. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Systems* **3**, 95–98 (2016).
16. Dudchenko, O. *et al*. *De novo* assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
17. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Systems* **3**, 99–101 (2016).
18. Abrusán, G. *et al*. TEclass—a tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* **25**, 1329–1330 (2009).
19. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* **35**, W265–W268 (2007).
20. Tarailo-Graovac, M. & Chen, N. Using RepeatMasker to identify repetitive elements in genomic sequences. *Current Protocols in Bioinformatics* **25**, 4.10. 11–14.10. 14 (2009).
21. Jurka, J. *et al*. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research* **110**, 462–467 (2005).
22. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573–580 (1999).
23. Lin, Y. *et al*. quarTeT: a telomere-to-telomere toolkit for gap-free genome assembly and centromeric repeat identification. *Horticulture Research* **10**, uhad127 (2023).
24. Mount, D. W. Using the basic local alignment search tool (BLAST). *Cold Spring Harbor Protocols* **2007**, pdb. top17 (2007).
25. Birney, E. *et al*. GeneWise and genomewise. *Genome Research* **14**, 988–995 (2004).
26. Kim, D. *et al*. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* **37**, 907–915 (2019).
27. Cantarel, B. L. *et al*. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research* **18**, 188–196 (2008).
28. Boeckmann, B. *et al*. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* **31**, 365–370 (2003).
29. Kulikova, T. *et al*. The EMBL nucleotide sequence database. *Nucleic Acids Research* **32**, D27–D30 (2004).
30. Ogata, H. *et al*. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **27**, 29–34 (1999).
31. Hunter, S. *et al*. InterPro: the integrative protein signature database. *Nucleic Acids Research* **37**, D211–D215 (2009).
32. The Updated Genome Warehouse: Enhancing Data value, Security, and Usability to Address Data Expansion. Genomics Proteomics Bioinformatics **10**; 23 (1): qzaf010. [PMID=39977364] (2025) may.
33. *NGDC Genome Warehouse*. https://ngdc.cncb.ac.cn/gwh/Assembly/102277/show (2025).
34. *NCBI Bioproject* https://identifiers.org/ncbi/bioproject:PRJNA1172361 (2025).
35. *NCBI Genbank*. https://identifiers.org/ncbi/insdc.gca:GCA_044707065.1 (2025).
36. Bian, C. Genome and annotation of *Macrobrachium hainanense*. *Figshare* https://doi.org/10.6084/m9.figshare.27600606 (2025).
37. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31037707 (2025).
38. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR31037706 (2025).
39. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR33222360 (2025).

Acknowledgements

This study was supported by National Key Research and Development Program of China (2023YFE0205100), Macau Science and Technology Development Fund (FDCT0102/2023/AMJ), the Key R&D Program of Hainan Province, China (ZDYF2024XDNY163).

Author contributions

S.S. and C.B. conceived this project; Y.J. and Q.Y. collected and identified the samples; C.B., R.H. and W.H. assembled the genome and performed the annotation set; S.M., R.H. and N.Z. analyzed the data. C.B., N.Z. and S.S. wrote 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 N.Z. or S.S.