scientific data



OPEN

DATA DESCRIPTOR

# A high-quality chromosome-level genome assembly of Pacific whiteleg shrimp (*Penaeus vannamei*)

Jian Liao[1,2], Zhuoxin Lai[1,2], Chuanmeng He[1,2], Gaocong Li[3], Zhongdian Dong[1,2], Yusong Guo[1,2] ✉ & Zhongduo Wang[1,2] ✉

**The Pacific whiteleg shrimp (*Penaeus vannamei*, synonym *Lipopenaeus vannamei*) is an important aquaculture species, and its breeding improvement relies heavily on high-quality genomic resources. Here, we sequenced the genome of the Pacific whiteleg shrimp using PacBio and Hi-C technologies, resulting in a 1.7 Gb genome assembly with a Contig N50 of 57.65Kb and a Scaffold N50 of 605Kb. BUSCO analysis indicated that our genome assembly achieved a gene coverage of 84.6%, a figure considered high among crustacean reference genomes. In total, 907,237,377 bp of repetitive sequences (accounting for 50.48% of the assembled genome) and 22,923 protein-coding genes were identified. Utilizing Hi-C technology, we anchored 14,604 contigs onto 44 chromosomes, yielding a chromosome-level genome assembly with a contig N50 of 42,894,961 bp. In summary, our high-quality genome assembly provides valuable reference genomic resources for the improvement and breeding of elite cultivars of the Pacific whiteleg shrimp.**

## Background & Summary

The Crustacea, a subphylum within Arthropoda that dominates aquatic environments, holds significant importance in ecology and fisheries[1]. Specifically, the Penaeidae shrimp family stands as a crucial species in fisheries and aquaculture, boasting an extraordinarily diverse array of over 400 species, the majority of which possess economic value[2]. Consequently, this family has garnered significant attention from researchers[1,3].

The Pacific whiteleg shrimp, *Penaeus vannamei*, also known as *Litopenaeus vannamei*, is a critically important economic species that holds significant economic value in the aquaculture industry, serving as a high-quality source of protein for humans. Native to the eastern Pacific coast, it ranges from the Sonora region in northern Mexico, spanning Central and South America, and extending south to Tumbes in Peru[4]. These regions are characterized by warm climates with average annual water temperatures exceeding 20 °C, providing favorable environmental conditions for the growth of the Pacific whiteleg shrimp[2]. The whiteleg shrimp plays a specific role in ecosystems and has various applications in fisheries and aquaculture. In ecological systems, the whiteleg shrimp occupies a unique niche, primarily inhabiting coastal waters with muddy substrates and depths ranging from 0 to 70 meters. They are known for their adaptability to a wide range of salinities and temperatures, enabling them to thrive in diverse aquatic environments. Their dietary habits and rapid growth contribute to the dynamic balance of food webs, consuming algae and organic debris while serving as prey for larger fish and crustaceans. This interaction helps maintain biodiversity and enhances the overall productivity of aquatic ecosystems. The annual production of globally cultivated Pacific whiteleg shrimp (*P. vannamei*) has exceeded 6.5 million tons, with China ranking highest in terms of output. Currently, a portion of the genomic data for shrimp species such as *Marsupenaeus japonicus* and *Penaeus monodon* has been published[5]. Their total genome lengths are 1.94 Gb and 2.04 Gb, respectively, covering over 97% of the coding regions, and encompassing 16,716 and 18,100 genes,

[1]Key Laboratory of Aquaculture in South China Sea for Aquatic Economic Animals of Guangdong Higher Education Institutes, Fisheries College, Guangdong Ocean University, Zhanjiang, 524088, China. [2]Guangdong Provincial Key Laboratory of Aquatic Animal Disease Control and Healthy Aquaculture, College of Fisheries, Guangdong Ocean University, Zhanjiang, 524088, China. [3]College of Electronics and Information Engineering, Guangdong Ocean University, Zhanjiang, 524088, China. ✉e-mail: gysrabbit@163.com; wangzd@gdou.edu.cn

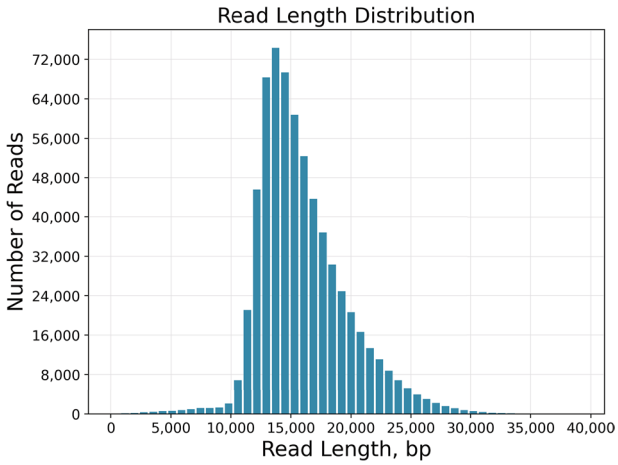


**Fig. 1** The HiFi Length Distribution Chart of the Genomic Sequencing of *P. vannamei*.

respectively[5]. Although the genome of the Pacific whiteleg shrimp strain Kehai 1 has been released, its quality still requires significant improvement. The previously published genome of the Pacific whiteleg shrimp strain Kehai 1, which was sequenced and assembled using second-generation sequencing technologies (Illumina and 454 platforms), yielded a genome size of 1.66 Gb with a Contig N50 of only 57.65Kb and a Scaffold N50 of 605Kb.A recent study has utilized PacBio long-read data from 14 publicly available genomic libraries (totaling 131.2 Gb) to make minor improvements to the genome of the Pacific whiteleg shrimp (*P. vannamei*)[6]. The refined genome has a size of 2.055 Gb, with an N50 of 40.14 Mb, an L50 of 21, and the longest scaffold of 65.79 Mb[6]. In contrast, the present study primarily employed third-generation sequencing technology, which offers longer read lengths, combined with Hi-C technology, to assemble a chromosome-level genome of the Pacific whiteleg shrimp strain Xinghai 1 (including 44 chromosomes). This strain is an elite cultivar developed by our research team. The assembled genome achieved a Contig N50 of 214Kb and a Scaffold N50 of 42.9 Mb, with a genome size of 1.7 Gb. The further refinement of the genome of the Pacific whiteleg shrimp will enable us to conduct a more in-depth analysis of its growth, development, reproduction mechanisms, and adaptive strategies to the environment. For instance, we can investigate how certain genes regulate the growth rate of Pacific whiteleg shrimp or analyze whether specific genetic variations are directly associated with its remarkable disease resistance and stress tolerance. By deeply exploring these genomic characteristics, we will not only gain a more comprehensive understanding of the biological traits of Pacific whiteleg shrimp but also provide solid scientific support for future aquaculture practices, genetic improvement programs, and disease prevention and control strategies. The genome of *P. vannamei* provided in this study facilitates future research in functional genomics, the development of genetic markers, and the application of genome editing technologies such as CRISPR/Cas9.

## Methods

**Sample collection and genomic sequencing.** The samples of *P. vannamei* were derived from the new breed "Xinghai 1" with outstanding ammonia nitrogen resistance developed by Guangdong Ocean University, Zhanjiang, city, Guangdong Province, China. To eliminate the influence of surface microorganisms, we performed surface cleaning using 75% ethanol and sodium hypochlorite solution, respectively[7]. High-quality genomic DNA was extracted from muscle tissue using a modified CTAB method[1]. RNase A was used to remove RNA contaminants. The quality and quantity of the extracted DNA were examined using a NanoDrop 2000 spectrophotometer (NanoDrop Technologies, Wilmington, DE, USA) and electrophoresis on a 0.8% agarose gel, respectively.

During the PacBio sequencing process, we performed sequencing utilizing the Pacific Biosciences Sequel II platform, generating high-quality HiFi reads acquired through the circular consensus sequencing (CCS) mode. The entire sequencing endeavor yielded 55.39 Gb of single-molecule real-time (SMRT) long-read data, attaining a coverage of approximately 30.8×. In parallel, we successfully constructed Hi-C libraries employing the MboI restriction enzyme and sequenced these libraries on the Illumina Novaseq/MGI-2000 platform. Subsequently, in this phase of our investigation, we amassed approximately 236.56 Gb of data, achieving a remarkable coverage of 136.6×. Additionally, we sequenced the samples using a BGI platform, yielding 106.41 Gb of raw sequence data, equivalent to 59.2× coverage of the genome. These comprehensive datasets will provide invaluable insights for our subsequent genomic analysis and research endeavors.

**HiFi statistics.** HiFi sequences refer to those that have been sequenced three or more times within the same Zero-Mode Waveguide (ZMW) well and corrected through mutual calibration of multiple sequencing results, achieving an accuracy of over 99%. In this study, a thorough analysis of the quantity and length of HiFi sequences was conducted. The total number of base pairs (bp) for HiFi reads amounted to 55,391,405,363 bp, encompassing a total of 3,406,777 reads with an average length of 16,259 bp. Notably, the Reads N50 length was 16,343 bp. Furthermore, the GC content of these reads reached 33.5%. To provide a more intuitive representation of the characteristics of HiFi sequences, sequence length distribution and sequence quality distribution plots were generated, as shown in Figs. 1 and 2, respectively.

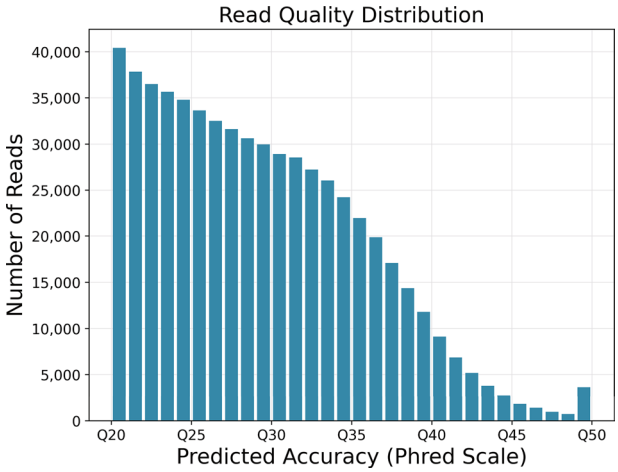


**Fig. 2** The HiFi Quality Distribution Chart of the Genomic Sequencing of *P. vannamei*.

| Category | Number of BUSCOs | Percentage |
|---|---|---|
| Complete BUSCOs (C) | 857 | 84.6% |
| Complete and single-copy BUSCOs (S) | 840 | 82.9% |
| Complete and duplicated BUSCOs (D) | 17 | 1.7% |
| Fragmented BUSCOs (F) | 57 | 5.6% |
| Missing BUSCOs (M) | 99 | 9.8% |
| Total BUSCO groups searched | 1013 | 100.0% |

**Table 1.** Assessment of completeness of *P. vannamei* assembly.

| Content | Mapping rate | Coverage |
|---|---|---|
| second-generation reads | 95.04% | 88.79% |
| HiFi reads | 99.90% | 72.16% |
| ONT reads | 99.96% | 95.14% |

**Table 2.** Genomic assessment of *P. vannamei*.

**Genome assembly and evalution.** First, we estimated the genome size of the sample using the Sysmex CyFlow® Cube6 flow cytometer, predicting a genome size of approximately 1.7 Gb. Subsequently, we utilized the MECAT2[8] assembler to conduct an initial assembly of the sequencing data, resulting in a Contig genome with a total length of 1.97 Gb, where the Contig N50 value was 190Kb. Following Hi-C scaffolding, the total length of the genome shortened to 1.79 Gb, while the Contig N50 value increased to 200Kb, achieving a high scaffolding completenessof 95.09%. The final genome assembly had a total length of 1.79 Gb and a Contig N50 value of 200Kb.

To eliminate CCS reads containing residual PacBio adapter sequences, the HiFiAdapterFilt software was employed[9]. Subsequently, clean PacBio reads were assembled using the default settings in Hicanu v2.0[10]. The purge_dup software was then utilized to remove haplotypic and consecutively overlapping sequences[11]. To assess the comprehensiveness of the assembly, we utilized the Benchmarking Universal Single-Copy Orthologs v5.1.2 (BUSCO) tool, leveraging orthologous genes from the Arthropoda database in OrthoDB 10. The results indicated a BUSCO assessment completeness of 84.6%. Specifically, the counts for Complete BUSCOs, Complete and single-copy BUSCOs, Complete and duplicated BUSCOs, Fragmented BUSCOs, and Missing BUSCOs are summarized in Table 1. Additionally, a back-mapping analysis was conducted, revealing a mapping rate of 95.04% and coverage of 88.79% for second-generation reads, 99.90% mapping rate and 72.16% coverage for HiFi reads, and 99.96% mapping rate and 95.14% coverage for ONT reads. Furthermore, the QV value was determined to be 28.307 (Table 2).

**Hi-C scaffolding.** During the initial processing phase, we successfully eliminated low-quality raw reads with quality scores below 20 and lengths shorter than 30 bp using FASTP v0.20.0, while simultaneously removing adapter sequences. Subsequently, we employed the BOWTIE2 v2.3.2 tool with specific parameters (-end-to-end–very-sensitive -L 30) to accurately align the purified reads with the contig assembly[12]. Next, we utilized the Juicer v1.5 tool to construct a draft genome using Hi-C read data and performed misjoin correction, ordering, and orientation adjustments using 3D-DNA v180922 software with the parameter (-r 0)[13,14]. By integrating Hi-C data with the contig-level assembly, we successfully generated a chromosome-level assembly. Furthermore, we optimized these contigs, resulting in 44 scaffolds with a total size of 1,701,849,523 bp, including a maximum scaffold

| | Total(bp) | Contig Num | Contigs N50(bp) | Scaffold Num | Scaffold N50(bp) |
|---|---|---|---|---|---|
| Initial Assembly Results | 1,967,314,555 | 29,250 | 188,690 | | |
| Draft Genome Sequence Before Hi-C Assisted Assembly | 1,789,786,463 | 20,029 | 201,332 | | |
| Chromosome-anchored Sequence After Hi-C Assisted Assembly | 1,701,849,523 | 14,604 | 214,247 | 44 | 42,894,961 |
| Final Assembly Results | 1,789,786,463 | 20,029 | 201,332 | 5,469 | 42,332,664 |

**Table 3.** statistics of genome assembly and Hi-C scaffolding information for *P. vannamei*.

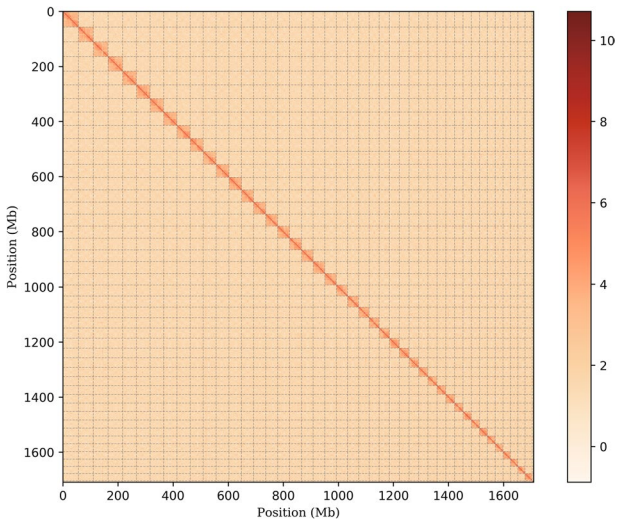


**Fig. 3** Hi-C interaction heatmap for *P. vannamei* genome (resolution: 250Kb). The colors ranging from light to dark represent increasing interaction strengths, with deeper colors indicating stronger interactions. The x-axis and y-axis represent the base positions on the genome. The 44 rows or columns of blocks in the figure correspond to the 44 chromosomes of *P. vannamei*.

N50 size of 42,894,961 bp (detailed in Table 3). Finally, we anchored, ordered, and oriented these scaffolds using Hi-C data, obtaining 44 chromosomes (Chr1-Chr44) that encompassed over 95.09% of the assembled sequences, as visually presented in Fig. 3. The genome characteristics of *P. vannamei* were displayed in Fig. 4.

**Repeat annotation.** To accurately identify tandem repeats and interspersed repeats (i.e., transposon elements), we employed a comprehensive approach combining homology-based analysis with de novo prediction. For the homology-based analysis, we leveraged two powerful tools, RepeatMasker v4.1.2 (with parameters: -nolow -no_is -norna -parallel 2) and RepeatProteinMask v1.36 (with parameters: -engine ncbi -noLowSimple -pvalue 0.0001) (http://www.repeatmasker.org), to precisely predict transposable elements (TEs) within the *P. vannamei* genome, utilizing the known TE protein database and the RepBase library v21.12[15]. For de novo prediction, we utilized RepeatModeler v2.0.2a and LTR_FINDER v1.0.5[16] to successfully construct a novel repeat sequence library for the *P. vannamei* genome. Subsequently, we employed RepeatMasker to search and meticulously classify the newly constructed repeat library. During this process, the Tandem Repeat Finder (TRF)[17] played a pivotal role in effectively identifying tandem repeats, while RepeatMasker specialized in recognizing non-dispersed repeat sequences. Finally, through an in-depth analysis of genomic annotations, we discovered that transposable elements occupy a significant proportion of the *P. vannamei* genome, approximately 50.48% (as detailed in Table 4).

**Gene prediction and functional annotation.** To accurately annotate the gene structure of *P. vannamei*, we employed three complementary strategies: ab initio annotation, homology-based prediction, and RNA sequencing-assisted prediction. For homology prediction, we utilized Exonerate v2.2.0 to perform an exhaustive alignment of the *P. vannamei* genome sequence with protein sequences from closely related species[18], including previously published sequences from *P. vannamei*, *P. chinensis*, *P. monodon*, and *P. japonicus*. Concurrently, we generated ab initio gene structures using Augustus v3.3 and Genescan v1.0[19,20]. Additionally, we integrated RNA-seq data from *P. vannamei* and aligned it to the repeat-masked genome to precisely identify splice sites and exonic regions. Subsequently, we utilized MAKER v3.00 to comprehensively integrate all these data[21]. To further refine the gene structure based on transcriptome data, we also employed the PASA tool[22]. Finally, we successfully constructed a comprehensive gene set containing 22,923 genes (detailed in Table 5).

Regarding functional annotation, we employed the Blastp algorithm (with parameters: -e 1e-5) to conduct a thorough comparison of the integrated gene set against multiple authoritative databases, including SwissProt37, KEGG38, TrEMBL39, GO Ontology (GO)40, and the NR database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz)[23–25]. Furthermore, we utilized PfamScan and InterProScan (v5.35–74.0) to conduct an in-depth search for protein

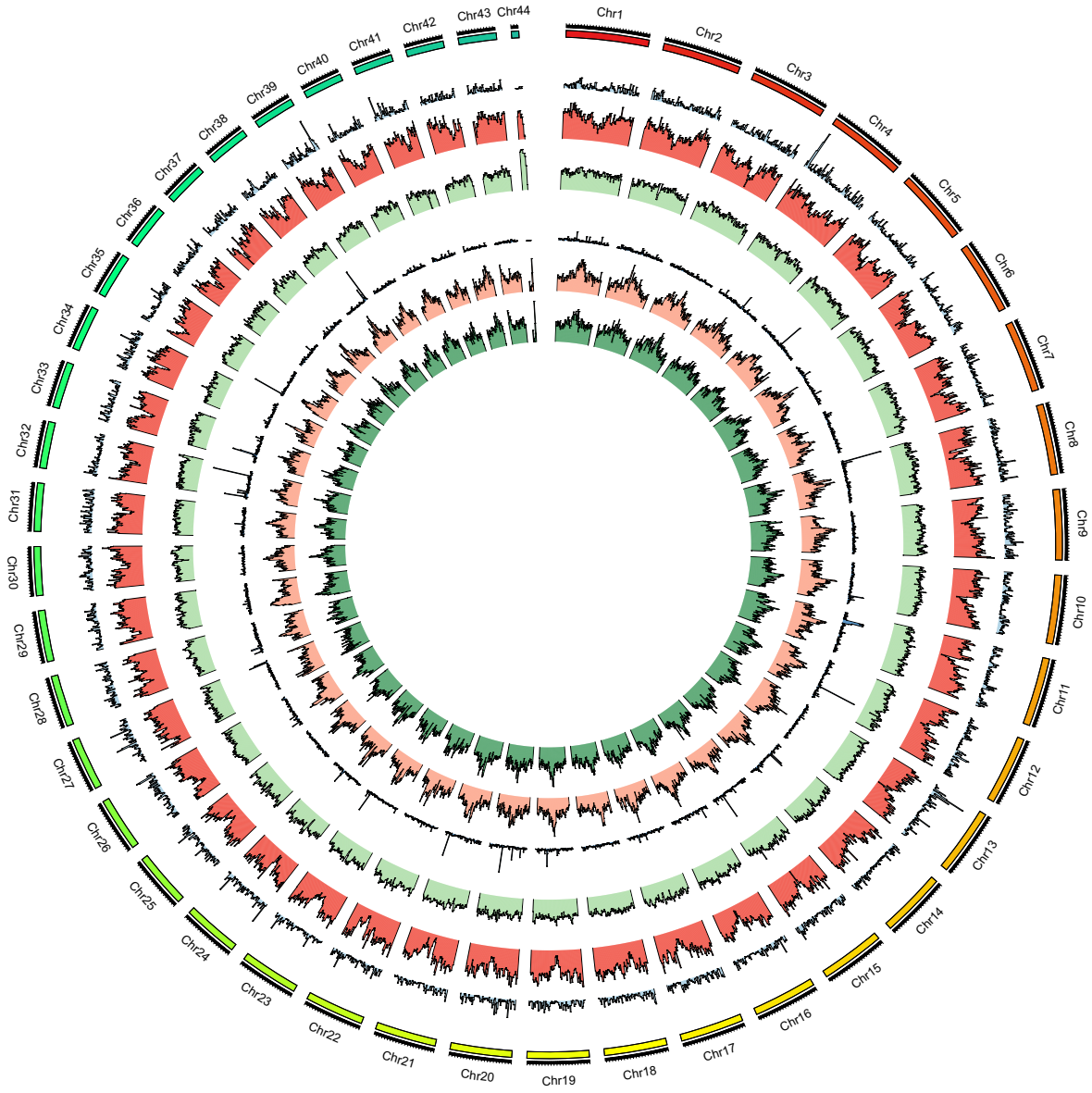


**Fig. 4** Genomic Characteristics of *P. vannamei*. The outermost circle (in clockwise direction) represents different chromosomal sequences, the second circle represents gene density, the third circle represents GC content, the fourth circle represents repetitive sequence content, the fifth circle represents single nucleotide polymorphism (SNP) density distribution information, and the sixth circle represents insertion and deletion (INDEL) density distribution information.

| | RepeatMasker TEs | | RepeatProteinMask TEs | | *De novo* | | Combined TEs | |
|---|---|---|---|---|---|---|---|---|
| **Type** | **Length (bp)** | **% in genome** | **Length (bp)** | **% in genome** | **Length (bp)** | **% in genome** | **Length (bp)** | **% in genome** |
| DNA | 644745290 | 35.88 | 613468 | 0.03 | 175931913 | 9.79 | 698990228 | 38.9 |
| LINE | 411247926 | 22.88 | 34770340 | 1.93 | 58530285 | 3.26 | 418585759 | 23.29 |
| SINE | 338294 | 0.02 | 0 | 0 | 218591 | 0.01 | 551857 | 0.03 |
| LTR | 268301763 | 14.93 | 3740032 | 0.21 | 30979646 | 1.72 | 290713258 | 16.18 |
| Other | 54869 | 0 | 0 | 0 | 0 | 0 | 54869 | 0 |
| Unknown | 8849748 | 0.49 | 777 | 0 | 39450397 | 2.2 | 47529615 | 2.64 |
| Total TE | 808943410 | 45.01 | 39121678 | 2.18 | 302306802 | 16.82 | 907237377 | 50.48 |

**Table 4.** Repeat sequence classification result statistics for *P. vannamei* genome.

| Gene set | Number | Average gene length (bp) | Average CDS length (bp) | Average exon per gene | Average exon length (bp) | Average intron length (bp) |
|---|---|---|---|---|---|---|
| De novo/AUGUSTUS | 48963 | 10858.18 | 1091.3 | 4.12 | 265.2 | 3135.41 |
| De novo/Genscan | 53609 | 19377.74 | 1155.69 | 4.98 | 232.06 | 4578.21 |
| homo/*P.vannamei* | 25660 | 9172.1 | 1170.55 | 4.89 | 239.16 | 2054.65 |
| homo/*P.chinensis* | 22908 | 15058.56 | 1351.2 | 5.64 | 239.45 | 2952.21 |
| homo/*P.monodon* | 24229 | 11487.57 | 1141.7 | 4.71 | 242.54 | 2790.66 |
| homo/*P.japonicus* | 23156 | 16374.16 | 1340.08 | 5.7 | 235.3 | 3201.97 |
| trans.orf/RNAseq | 10995 | 27755.62 | 1633.97 | 7.99 | 391.81 | 3520.81 |
| MAKER | 27694 | 17635.96 | 1424.36 | 5.8 | 344.9 | 3260.42 |
| PASA | 22923 | 22023.6 | 1549.98 | 6.65 | 341.71 | 3387.21 |

**Table 5.** Statistical analysis of protein coding genes for *P. vannamei* genome.

| Type | Number | Percent(%) |
|---|---|---|
| Total | 22923 | |
| InterPro | 14037 | 61.24 |
| GO | 13216 | 57.65 |
| KEGG_ALL | 19935 | 86.97 |
| KEGG_KO | 7879 | 34.37 |
| Swissprot | 13456 | 58.7 |
| TrEMBL | 16297 | 71.09 |
| NR | 21148 | 92.26 |
| Annotated | 21180 | 92.4 |

**Table 6.** Functional annotation of protein-coding genes for *P. vannamei* genome.

| Data type | Mapping rate (%) | Average sequencing depth | Coverage (%) | Coverage (>=5X,%) | Coverage (>=10X,%) | Coverage (>=20X,%) |
|---|---|---|---|---|---|---|
| BGI | 95.04 | 51.14 | 88.79 | 69.69 | 60.69 | 52.27 |
| HiFi | 99.90 | 27.69 | 72.16 | 52.36 | 41.59 | 29.27 |
| ONT | 99.96 | 4.98 | 95.14 | 45.45 | 4.46 | 0.77 |

**Table 7.** Assessment of assembly integrity and sequencing uniformity.

structural domains based on the PFAM and InterPro protein databases, respectively[26]. After an exhaustive annotation process, we found that 92.4% of the predicted protein-coding genes received precise functional annotations (detailed in Table 6).

## Data Records

The genome project of the Pacific whiteleg shrimp (*P. vannamei*) and its associated samples have been successfully submitted to the NCBI database, acquiring the designated BioProject ID (PRJNA1091545) and BioSample ID (SAMN40597728). Concurrently, we have uploaded the Illumina sequencing data (SRR28527408)[27], PacBio sequencing data (SRR28525434 - SRR28525438)[27–32], and Hi-C sequencing data (SRR28447356)[33] to NCBI. Sequence Read Archive (SRA) project number was SRP497752[34]. Additionally, the complete genome sequence has also been successfully submitted to NCBI with the accession number JBFNAF000000000[35]. To access the genome annotation files, please visit the Figshare platform for download[36].

## Technical Validation

The completeness of the *P. vannamei* genome was assessed using BUSCO with the tetrapoda_odb10 (parameters: -m genome -l tetrapoda_odb10)[37]. The assembled genome exhibited approximately 84.6% complete BUSCO genes, with 82.9% being complete and single copy, 1.7% being complete and duplicated, 5.6% being fragmented, and 9.8% being missed (Table 1). To evaluate the integrity of the assembly and the uniformity of sequencing coverage, second- and third-generation data from the study species were selected. The assembled genome was aligned back using alignment tools BWA[38] and minimap2[39]. The alignment rate, the degree of genome coverage, and the distribution of coverage depth of the reads were then statistically analyzed to assess the integrity of the assembly and the uniformity of sequencing coverage. The results are presented in Table 7. Collectively, all the aforementioned results indicate that the *P. vannamei* strain Xinghai 1 genome we have obtained is of relatively high quality among crustacean species.

## Usage Notes

All data analyses were rigorously conducted in accordance with the manuals and protocols provided by the published bioinformatics tools. Detailed information on the software versions and specific parameter settings has been comprehensively described in the "Methods" section.

## Code availability

In this work, no specific codes or scripts were utilized. The commands employed for data processing were all executed in accordance with the manuals and protocols provided by the respective software packages.

Received: 29 October 2024; Accepted: 19 February 2025;
Published online: 26 February 2025

## References

1. Zhang, X. *et al*. Penaeid shrimp genome provides insights into benthic adaptation and frequent molting. *Nat Commun* **10**, 356 (2019).
2. Albalat, A., Zacarias, S., Coates, C. J., Neil, D. M. & Planellas, S. R. Welfare in Farmed Decapod Crustaceans, With Particular Reference to Penaeus vannamei. *Front. Mar. Sci.* **9**, 886024 (2022).
3. Rothlisberg, P. C. Aspects of penaeid biology and ecology of relevance to aquaculture: a review. *Aquaculture* **164**, 49–65 (1998).
4. Briggs, M., Funge-Smith, S., Subasinghe, R. & Phillips, M. *Introductions and Movement of Penaeus Vannamei and Penaeus Stylirostris in Asia and the Pacific*. 1–79 (2004).
5. Yuan, J. *et al*. Genomic resources and comparative analyses of two economical penaeid shrimp species, Marsupenaeus japonicus and Penaeus monodon. *Marine Genomics* **39**, 22–25 (2018).
6. Perez-Enriquez, R., Juárez, O. E., Galindo-Torres, P., Vargas-Aguilar, A. L. & Llera-Herrera, R. Improved genome assembly of the whiteleg shrimp *Penaeus* (*Litopenaeus*) *vannamei* using long- and short-read sequences from public databases. *Journal of Heredity* **115**, 302–310 (2024).
7. Li, H. *et al*. A chromosome-level genome assembly of Sesamia inferens. *Sci Data* **11**, 134 (2024).
8. Xiao, C.-L. *et al*. MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *Nat Methods* **14**, 1072–1074 (2017).
9. Sim, S. B., Corpuz, R. L., Simmonds, T. J. & Geib, S. M. HiFiAdapterFilt, a memory efficient read processing pipeline, prevents occurrence of adapter sequence in PacBio HiFi reads and their negative impacts on genome assembly. *BMC Genomics* **23**, 157 (2022).
10. Koren, S. *et al*. Canu: scalable and accurate long-read assembly via adaptive *k* -mer weighting and repeat separation. *Genome Res* **27**, 722–736 (2017).
11. Guan, D. *et al*. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).
12. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).
13. Dudchenko, O. *et al*. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
14. Durand, N. C. *et al*. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems* **3**, 95–98 (2016).
15. Jurka, J. *et al*. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110**, 462–467 (2005).
16. Xu, Z. & Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research* **35**, W265–W268 (2007).
17. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**, 573–580 (1999).
18. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 31 (2005).
19. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* **268**, 78–94 (1997).
20. Stanke, M. *et al*. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Research* **34**, W435–W439 (2006).
21. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
22. Haas, B. J. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31**, 5654–5666 (2003).
23. Ashburner, M. *et al*. Gene Ontology: tool for the unification of biology. *Nat Genet* **25**, 25–29 (2000).
24. Bairoch, A. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research* **28**, 45–48 (2000).
25. Kanehisa, M. *et al*. KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36**, D480–D484 (2007).
26. Blum, M. *et al*. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Research* **49**, D344–D354 (2021).
27. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR28527408 (2024).
28. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR28525434 (2024).
29. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR28525435 (2024).
30. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR28525436 (2024).
31. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR28525437 (2024).
32. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR28525438 (2024).
33. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRR28447356 (2024).
34. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP497752 (2024).
35. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_042767895.1 (2024).
36. Liao, J. *Penaeus vannamei*_genome. *figshare* https://doi.org/10.6084/m9.figshare.27179964 (2024).
37. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods in molecular biology (Clifton, N.J.)* **1962**, 227–245 (2019).
38. Chong, C.-F., Li, Y.-C., Wang, T.-L. & Chang, H. Stratification of Adverse Outcomes by Preoperative Risk Factors in Coronary Artery Bypass Graft Patients: An Artificial Neural Network Prediction Model.
39. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

## Acknowledgements
This work was funded by Construction Project of Modern Seed Industry Park for Whiteleg Shrimp of Guangdong Province from the Department of Agriculture and Rural Affairs of Guangdong Province, the National Key Research and Development Program of China (Grant No. 2024YFD2401802 & 2024YFD2401803), the Guangdong Provincial Ordinary University Youth Innovative Talent Project in 2024 (Grant No. 2024KQNCX134), the Guangdong Provincial Special Fund Project for Talent Development Strategy in 2024 (Grant No. 2024R3005), the Guangdong Ocean University Scientific Research Startup Funding Project (Grant No. 060302022312), and Guangdong Provincial Field Observation and Research Station for Marine Ecosystem in Hanjiang River Estuary - Nanao Island Area Open Fund (Grant No. HNS202407). Additionally, we would like to express our gratitude to Engineer Li Yijun from Hainan Xinbang Marine Biotechnology Co., Ltd., China, for providing the samples for this project.

## Author contributions
Y.G. and Z.W. conceived the project. Z.L., M.H. and G.L. collected the samples. J.L., G.L. and Z.D. performed the genome assembly, gene annotation and other bioinformatics analysis. J.L., Y.G. and Z.W. wrote and revised the manuscript. G.L. and Z.D. 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 Y.G. or Z.W.

**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.