G3, 2021, 11(11), jkab268
DOI: 10.1093/g3journal/jkab268
Advance Access Publication Date: 5 August 2021
Genome Report

Genome and transcriptome assemblies of the kuruma shrimp, *Marsupenaeus japonicus*

Satoshi Kawato [1], Koki Nishitsuji [2], Asuka Arimoto [2,†], Kanako Hisata [2], Mayumi Kawamitsu,[3] Reiko Nozaki,[1] Hidehiro Kondo [1], Chuya Shinzato [4], Tsuyoshi Ohira,[5] Noriyuki Satoh [2], Eiichi Shoguchi [2], and Ikuo Hirono [1,*]

[1]Laboratory of Genome Science, Tokyo University of Marine Science and Technology, Tokyo 108-8477, Japan
[2]Marine Genomics Unit, Okinawa Institute of Science and Technology Graduate University, Okinawa 904-0495, Japan
[3]DNA Sequencing Section, Okinawa Institute of Science and Technology Graduate University, Okinawa 904-0495, Japan
[4]Atmosphere and Ocean Research Institute, The University of Tokyo, Chiba 277-0882, Japan, and
[5]Faculty of Science, Department of Biological Sciences, Kanagawa University, Kanagawa 221-8686, Japan

*Corresponding author: Tokyo University of Marine Science and Technology, 4-5-7, Konan, Minato-ku, Tokyo 108-8477, Japan. Email: hirono@kaiyodai.ac.jp
†Present address: Marine Biological Laboratory, Graduate School of Integrated Sciences for Life, Hiroshima University, Onomichi, Hiroshima, Japan.

Abstract

The kuruma shrimp *Marsupenaeus japonicus* (order Decapoda, family Penaeidae) is an economically important crustacean that occurs in shallow, warm seas across the Indo-Pacific. Here, using a combination of Illumina and Oxford Nanopore Technologies platforms, we produced a draft genome assembly of *M. japonicus* (1.70 Gbp; 18,210 scaffolds; scaffold N50 = 234.9 kbp; 34.38% GC, 93.4% BUSCO completeness) and a complete mitochondrial genome sequence (15,969 bp). As with other penaeid shrimp genomes, the *M. japonicus* genome is extremely rich in simple repeats, which occupies 27.4% of the assembly. A total of 26,381 protein-coding gene models (94.7% BUSCO completeness) were predicted, of which 18,005 genes (68.2%) were assigned functional description by at least one method. We also produced an Illumina-based transcriptome shotgun assembly (40,991 entries; 93.0% BUSCO completeness) and a PacBio Iso-Seq transcriptome assembly (25,415 entries; 67.5% BUSCO completeness). We envision that the *M. japonicus* genome and transcriptome assemblies will serve as useful resources for the basic research, fisheries management, and breeding programs of *M. japonicus*.

Keywords: penaeidae; penaeid shrimp; kuruma prawn; *Penaeus japonicus*; *Marsupenaeus japonicus*; genome sequencing; genome annotation; transcriptome assembly

Introduction

The kuruma shrimp *Marsupenaeus japonicus* (order Decapoda, family Penaeidae) is an economically important crustacean that occurs across the Indo-West Pacific. In Japan, *M. japonicus* is especially highly prized as a seafood delicacy and has been a major fisheries and aquaculture target since the early 20th century. Artificial spawning and hatching of *M. japonicus* was achieved in 1933 (Hudinaga 1942), and an industrial-scale larval rearing technology was established by the 1960s (Liao 1985). Today, *M. japonicus* farming is practiced in several other countries in Asia and Europe, with China being the largest producer yielding 55,228 tons in 2018 (accessed April 2021;http://www.fao.org/fishery/statistics/).

Genomic resources of an organism provide powerful tools for investigating its basic biology, including economically important aspects such as growth, reproduction, and immunity. Genomic information is also important for developing molecular markers used in resource management and selective breeding programs. High-quality genome assemblies of three economically important shrimps [*Litopenaeus vannamei* (Zhang *et al.* 2019); *Penaeus monodon* (Uengwetwanit *et al.* 2021); *Fenneropenaeus chinensis* (Yuan *et al.* 2021)] have been recently reported. In contrast, the genomic resources of *M. japonicus* are severely limited. Although a previous study reported an Illumina-based draft genome assembly of *M. japonicus* (Yuan *et al.* 2018), this assembly is severely fragmented (2,434,740 scaffolds; scaffold N50: 912 bp), reflecting the repetitiveness and complexity of penaeid shrimp genomes.

Here, we present a high-quality draft genome assembly of *M. japonicus* generated using a combination of Illumina and Oxford Nanopore Technologies platforms. We also generated a transcriptome shotgun assembly and a PacBio Iso-Seq transcriptome assembly, which will complement the draft genome by capturing full-length transcript structures and genes missing from the genome assembly. We envision that the *M. japonicus* genome and transcriptome assemblies will serve as a useful resource for the basic research, fisheries management, and breeding programs of *M. japonicus*.

Materials and methods

General sequencing, assembly, and annotation strategy

We first generated an initial assembly using Illumina paired-end reads, and the assembly was scaffolded using Illumina paired-end

Received: April 07, 2021. **Accepted:** July 18, 2021

and mate-pair reads. We mapped RNA-seq reads onto the genome and took forward the mapped reads to generate a transcriptome shotgun assembly. The genome was further scaffolded using the transcriptome shotgun assembly, Iso-Seq cDNA sequences, and the RNA-seq reads. We improved this Illumina-based primary assembly by gap-filling using ONT long reads and scaffolding using ONT long reads, Illumina mate-pair reads, transcriptome shotgun assembly, and PacBio Iso-Seq cDNA sequences. See Supplementary Figure S1 for a summary of the assembly workflow.

To annotate protein-coding genes from the finished assembly, we combined multiple lines of evidence (*de novo* assembled transcripts, Iso-Seq cDNA sequences, RNA-seq reads, protein alignments, and *ab initio* predictions) into a consensus gene model set. We discarded low-quality predictions with no detectable similarity to known proteins. To maximize the recovery of gene models from transcript evidence, we again mapped the transcriptome shotgun assembly and Iso-Seq cDNA sequences to the genome. The gene models were again merged and filtered by homology search, deriving the final protein-coding gene model set. tRNA, rRNA, and other noncoding RNA genes were also predicted and were included in the annotation file. Supplementary Figure S2 summarizes the annotation workflow.

Marsupenaeus japonicus samples

For genome sequencing, we purchased a male *M. japonicus* (Ginoza2017, BioSample: SAMD00276454, 30 g body weight) from a commercial farm in Ginoza, Okinawa Prefecture, Japan, and kept it frozen at −80°C until use. For transcriptome analysis, we sequenced a total of 49 RNA samples, including 37 adult samples covering 18 different tissues and 12 larval samples covering nauplius, zoea, and mysis stages. The *M. japonicus* larvae were obtained from a hatchery on Kumejima Island, Okinawa Prefecture. See Supplementary Table S1 for a summary of the samples used in this study.

Illumina genomic DNA library preparation and sequencing

For Illumina sequencing, we prepared genomic DNA from muscle and sperm. We used the muscle DNA to construct a paired-end library and the mate-pair libraries, whereas the sperm DNA was used to prepare a paired-end library. We extracted genomic DNA using the phenol-chloroform-isoamyl alcohol extraction, followed by ethanol precipitation. The DNA preparations were resuspended in Tris-EDTA buffer and were frozen at −30°C until use. For pair-end sequencing, genomic DNA from muscle and sperm were sheared using a Covaris instrument, and paired-end libraries (600-bp nominal insert) were prepared using the KAPA Hyper Prep Kit Illumina platforms (KK8504, Roche). For mate-pair sequencing, genomic DNA from muscle was size-fractionated using SageELF (0.75% 1–18kb, SafeScience, ELD7510), and eleven mate-pair libraries (2, 3, 4, 5, 6, 7, 8, 10, 12, 14, and 15 kb nominal insert sizes) were prepared using the Nextera Mate Pair Library Prep Kit (FC-132-1001, Illumina). The pair-end and mate-pair libraries were sequenced for 2 × 150 cycles on an Illumina HiSeq4000 platform using Hiseq3000/4000 PE Cluster Kit cBot (PE-410-1001, Illumina) and HiSeq 3000/4000 SBS Kit (300 Cycles, FC-410-1003, Illumina). See Supplementary Table S2 for a summary of the sequencing reads generated in this study.

ONT library preparation and sequencing

For ONT sequencing, we extracted genomic DNA from muscle using the phenol-chloroform-isoamyl alcohol extraction, followed by isopropanol precipitation. The DNA was further purified with NucleoBond AXG 100 columns with the NucleoBond Buffer Set IV (Macherey-Nagel, Germany). ONT libraries were prepared with the Ligation Sequencing Kit (SQK-LSK109; Oxford Nanopore Technologies, UK) and were sequenced using multiple R9.4.1 flow cells on an ONT GridION platform. The raw .fast5 files were basecalled by Guppy v4.0.11 with the high-accuracy mode and quality filtering. Reads that were shorter than 1 kb were discarded using SeqKit v0.12.0 (Shen *et al.* 2016). The length-filtered ONT long reads were used in the downstream analyses. Supplementary Table S3 summarizes the filtered genomic reads generated in this study.

Illumina RNA sequencing

We extracted total RNA using RNAiso Plus (Takara, Japan), followed by isopropanol precipitation. The RNA was resuspended in nuclease-free water and was stored at −80°C until use. Paired-end RNA-seq libraries, with an empirical fragment size of 600–700 bp, were prepared using the KAPA mRNA HyperPrep Kit Illumina Platforms (KK8581, Roche) and were sequenced on the Hiseq4000 platform using the Hiseq3000/4000 PE Cluster Kit cBot (PE-410-1001, Illumina) and HiSeq 3000/4000 SBS Kit (300 Cycles, FC-410-1003, Illumina).

PacBio Iso-Seq sequencing and analysis

Four RNA samples (male hepatopancreas, male hemocytes, and two larval samples) were selected for PacBio Iso-Seq sequencing. We synthesized full-length cDNA using the SMARTer Pico PCR cDNA Synthesis Kit (Takara, 634928) and size-selected four fractions (>5, 3–5, 2–3, and 1–2 kb nominal sizes) using SageELF (0.75% 1–18 kb, SafeScience, ELD7510). The libraries were sequenced on a PacBio RS II platform using a SMRT Cell v3 8Pac (PacBio, 100-171-800) and the DNA Sequencing Kit 4.0 v2 (PacBio, 100-612-400).

The raw subread bam files were processed using ccs v3.4.0, and the resulting ccs bam files were processed with lima v1.10.0 to remove primer sequences. This was followed by Iso-Seq3 refine, cluster, and polish functions in Iso-Seq3 v3.1.2, generating the final assembly. Before submission to DDBJ/ENA/NCBI, we manually removed duplicated entries and trimmed primer sequences remaining in the assembly.

Genome size estimation and ploidy analysis

We processed the raw Illumina paired-end reads using Fastp v0.20.1 with relaxed parameters. For genome size estimation and ploidy analysis, we combined the trimmed paired-end reads from both libraries. We used KMC v3.1.1 (Kokot *et al.* 2017) to tally the occurrence of 17, 21, 23, 25, 27, 29, 31, 33, and 35-mers. The resulting *k*-mer histograms were used for genome size estimation using GenomeScope 2.0 (Ranallo-Benavidez *et al.* 2020). We adopted the *k*-mer size of $k = 23$, which yielded the highest model fit (Supplementary Table S4). We used the same 23-mer index for ploidy analysis using SmudgePlot v0.2.3dev (Ranallo-Benavidez *et al.* 2020).

For polishing the assembly, we error-corrected the trimmed reads using Tadpole in BBtools v38.86 (Bushnell *et al.* 2017).

De novo assembly of Illumina genomic DNA libraries

The Illumina paired-end reads from the sperm library were trimmed with fastp v0.19.4 (Chen *et al.* 2018); here, we applied stringent parameters (length_required 140, qualified_quality_phred 20, unqualified_percent_limit 10, n_base_limit = 0,

low_complexity_filter) with an aim to reduce the complexity of the dataset input to the assembler (Supplementary Table S3).

The raw mate-pair reads were processed using Trimmomatic v0.36 (Bolger *et al.* 2014) and NextClip v1.3.1 (Leggett *et al.* 2014). The filtered mate-pair reads were used for the scaffolding and misassembly detection of the initial assembly.

The filtered paired-end reads from the sperm library were *de novo* assembled using SPAdes v3.13.0 (Nurk *et al.* 2013). For producing a homozygous reference genome assembly, the initial assembly ("SPAdes" in Supplementary Table S5) was processed by Redundans v0.14a (Pryszcz and Gabaldón 2016) using paired-end reads from the muscle library (preprocessed by Fastp v0.19.6; Supplementary Table S3) and the filtered mate-pair reads from all libraries ("Redundans" in Supplementary Table S5). Misassembled sequences were broken by REAPR v1.0.18 (Hunt *et al.* 2013) using the filtered mate-pair reads from 04 to 15 k libraries, generating respective broken assemblies for each mate-pair library ("REAPR_04k" to "REAPR_15k" in Supplementary Table S5).

Transcriptome shotgun assembly

The raw reads were processed using Fastp v0.12.6. The trimmed reads were aligned to the "REAPR_06k" assembly using HISAT2 v2.1.0. We processed the alignments with SAMtools (Li *et al.* 2009) and BEDtools bamtofastq v2.25.0 (Quinlan and Hall 2010) to extract paired-end reads that mapped to the genome. The retained paired-end reads were *de novo* assembled using Trinity v2.8.6. In addition, the *in silico*-normalized reads generated by Trinity were *de novo* assembled using rnaSPAdes v3.13.0 (Bushmanova *et al.* 2019) and Trans-ABySS v2.0.1 (Robertson *et al.* 2010). We used a series of different k-mers (31-, 41-, 51-, 61-, 71-, 81-, 91-, 101-, 111-, and 121-mers) for Trans-ABySS. The assemblies were concatenated into a single file, and coding sequences were predicted using TransDecoder v5.5.0 (https://github.com/TransDecoder/TransDecoder/). The predicted coding sequences labeled as "complete" by TransDecoder were extracted and clustered using cd-hit-est v4.8.1 (Li and Godzik 2006), with the following settings: -c 0.98 -G 0 -aS 0.25 -d 0 -T 0 -M 0. The transcripts harboring the longest CDS in each cluster were retained for generating the final, nonredundant transcriptome assembly. We assessed the assembly completeness using BUSCO v4.1.4 using the arthropoda_odb10 dataset in transcriptome mode.

Scaffolding of genome assembly using transcriptome data

We chose the "REAPR_15k" assembly for further scaffolding because it showed the highest scaffolding efficiency in a preliminary analysis. The transcriptome shotgun assembly and Iso-Seq cDNA sequences were aligned to the genome using BLAT (Kent 2002), and the alignment was used for scaffolding with L_RNA_scaffolder (Xue *et al.* 2013). Onto the L_RNA_scaffolder output, we mapped the *in silico*-normalized RNA-seq reads (generated by Trinity) using HISAT2 (Kim *et al.* 2019b) and forwarded the alignment to P_RNA_Scaffolder (Zhu *et al.* 2018) for scaffolding using paired-end information. We gap-filled the assembly using Sealer v2.1.0 (Paulino *et al.* 2015) and corrected small assembly errors using NtEdit v1.3.1 (Warren *et al.* 2019). In the end, scaffolds shorter than 2 kb were discarded ("Illumina primary assembly" in Supplementary Table S5).

Assembly improvement using ONT long reads, Illumina mate-pair reads, and transcriptome data

We improved the Illumina primary assembly (Supplementary Table S5) by scaffolding and gap filling using ONT long reads, Illumina mate-pair reads, transcriptome shotgun assembly, and Iso-Seq cDNA sequences. For use in the following analyses, the raw Illumina mate-pair reads were trimmed using NxTrim v0.4.3-6eb8d5e (O'Connell *et al.* 2015), reads shorter than 60 bases were removed using Fastp v0.20.1, and low-entropy reads were removed by BBduk in BBTools v38.86 (Bushnell *et al.* 2017) (Supplementary Table S3). We used minimap2 v2.17-r941 (Li 2018) for mapping ONT long reads and Illumina paired-end and mate-pair reads.

The Illumina primary assembly was gap-filled by TGS-GapCloser v1.1.1 (Xu *et al.* 2020) using ONT long reads, followed by polishing using NtEdit v1.3.2 (Warren *et al.* 2019). The polished assembly was broken into contigs using split.scaffolds.to.contigs.pl (https://github.com/MadsAlbertsen/miscperlscripts/blob/master/split.scaffolds.to.contigs.pl). The broken contigs were subjected to the first round of iterative scaffolding by the following programs: LRScaf v1.1.11 (Qin *et al.* 2019), using ONT long reads; BESST v2.2.8 (Sahlin *et al.* 2014), using Illumina mate-pair reads from all libraries; L_RNA_Scaffolder, using the transcriptome shotgun assembly and Iso-Seq cDNA sequences (aligned by BLAT v36). Each scaffolding was followed by gap-filling by TGS-GapCloser v1.1.1 and polishing by NtEdit v1.3.2. After the iterative scaffolding, misassembled sequences were broken by REAPR v1.0.18 using 10-kb mate-pair reads, and haplotigs were removed by purge_haplotigs v1.1.1 (Roach *et al.* 2018) using ONT long reads. This was followed by a second round of iterative scaffolding (using LRScaf, BESST, and L_RNA_Scaffolder; each scaffolding run was followed by gap-filling by TGS-GapCloser and polishing by NtEdit) and a purge_haplotigs run. The third round of iterative scaffolding consisted of scaffolding by LRScaf v1.1.11 and BESST v2.2.8, with each scaffolding run followed by gap-filling by TGS-GapCloser v1.1.1 and polishing by NtEdit v1.3.2. Finally, the assembly was polished by HyPo v1.0.3 (Kundu *et al.* 2019) using the trimmed, error-corrected paired-end reads from the muscle library.

The polished assembly was further curated. Mitogenomic contamination was identified using BLASTN (-perc_identity 90) by querying the *M. japonicus* mitogenome sequence (NCBI Reference Sequence: NC_007010.1) (Yamauchi *et al.*, 2004), and two possible mitogenomic sequences showing abnormally high coverage were removed. Next, scaffolds matching at least one of the following criteria were removed: shorter than 2 kb; GC content over 70% or below 10% (which might represent bacterial contamination or assembly artifacts); gap content above 60%. Gaps occurring at scaffold ends were trimmed to expose the first contig; similarly, if a scaffold ends with a short contig (up to 100 bp) but is followed by a gap, the contig and gap were trimmed to expose the next contig. We mapped the trimmed, error-corrected paired-end reads from the muscle and sperm libraries against the assembly and calculated the read coverage (two libraries combined) of each scaffold using SAMtools coverage. Scaffolds with coverage below 30 or mean depth below 20 were discarded. To rescue genic regions, we queried the discarded scaffolds against the transcriptome shotgun assembly and recovered one scaffold, which had been discarded due to a high gap content. This scaffold was broken into contigs and returned to the assembly. We scanned for bacterial ribosomal DNA sequences by Barrnap v0.9 (https://github.com/tseemann/barrnap) and queried the positive hits against the

NCBI nonredundant nucleotide database using BLASTN (Accessed February 12, 2021). All the hits were found to be eukaryotic sequences and therefore were retained in the assembly. The completeness of the finished assembly was assessed by BUSCO v4.1.4 (Simão *et al.* 2015) using the arthropoda_odb10 dataset.

Mitogenome assembly and annotation

ONT long reads longer than 5 kb (Supplementary Table S3) were *de novo* assembled by Flye v2.8-b1674 (Kolmogorov *et al.* 2019), and a circular mitogenome contig was recovered by querying the reference *M. japonicus* mitogenome (NCBI Reference Sequence: NC_007010.1) against the assembly using BLASTN. The contig was rearranged, using SeqKit (Shen *et al.* 2016), to start at the same position (the first base of the *tRNA-Ile* gene) as the reference. The contig was then polished by Pilon v1.23 (Walker *et al.* 2014) fed with a Minimap2 alignment of error-corrected paired-end reads from the muscle library. We manually annotated the mitogenome based on the reference *M. japonicus* mitogenome (NCBI Reference Sequence: NC_007010.1).

Repeat identification

To characterize repeat elements in the *M. japonicus* genome, we generated a custom repeat library using RepeatScout v1.0.5 (Benson 1999), its associated scripts (build_lmer_table, filter-stage-1.prl, and filter-stage-2.prl), and RepeatMasker v4.1.1 (http://www.repeatmasker.org). The Mj_TUMSAT_v1.0 assembly was broken into contigs using split.scaffolds.to.contigs.pl script, and 18,434 contigs (998,779,913 bp) were subsampled using SeqKit. 14-mers occurring in the subsampled contigs were tallied using build_lmer_table script. The resulting 14-mer table was used as input for RepeatScout v1.0.5, and the output was parsed with filter-stage-1.prl script to retain candidate repeat elements. The candidate repeat elements were queried against the subsampled contigs with RepeatMasker v4.1.1, using RMBLAST v2.10.0+ as the search engine. Entries occurring no more than 3 times were discarded using filter-stage-2.prl script. The final repeat library was annotated using RepeatClassifier v2.0, a supplementary tool of RepeatModeler2 (Flynn *et al.* 2020) bundled in the Dfam TE Tools Container v1.2 (https://github.com/Dfam-consortium/TETools). The annotated repeat library was used to characterize the repeat contents of the finished assembly using RepeatMasker with RMBLAST as the search engine.

Prediction and functional annotation of protein-coding genes

To predict the protein-coding genes from the *M. japonicus* genome assembly, we combined cDNA alignments, RNA-seq read alignments, protein alignments, and *ab initio* predictions.

We newly generated a transcriptome shotgun assembly to maximize cDNA alignment-based genome annotation. The raw RNA-seq reads were trimmed using Fastp v0.20.1, and the trimmed reads were *de novo* assembled using Trinity v2.11.0 (Grabherr *et al.* 2011). The assembly (Trinity transcripts) consists of 1,810,475 contigs (N50: 763; Max: 38,355, min: 163). The Trinity transcripts and Iso-Seq cDNA sequences were used as input of DNA alignment-based gene prediction by PASA v2.4.1 (Haas *et al.* 2003), using BLAT v36 as aligner and TransDecoder v5.5.0 for CDS prediction.

To generate a genome-guided transcriptome assembly, we mapped the trimmed RNA-seq reads using HISAT2 v2.2.1 and reconstructed the transcripts using StringTie v2.1.4 (Pertea *et al.* 2015).

The RNA-seq alignments (generated by HISAT2) were used as the hints for *ab initio* gene prediction on the soft-masked genome by BRAKER2 v2.1.5 (Brůna *et al.* 2021), which internally runs GeneMark-ES v4.62_lic (Borodovsky and Lomsadze 2011), and Augustus v3.3.3 (Stanke *et al.* 2006).

To prepare protein-to-genome alignments, we downloaded 694 entries of NCBI Identical Protein Groups associated with *M. japonicus* (NCBI: txid27405; downloaded December 17, 2020). The proteins were queried to the genome, using TBLASTN, to identify

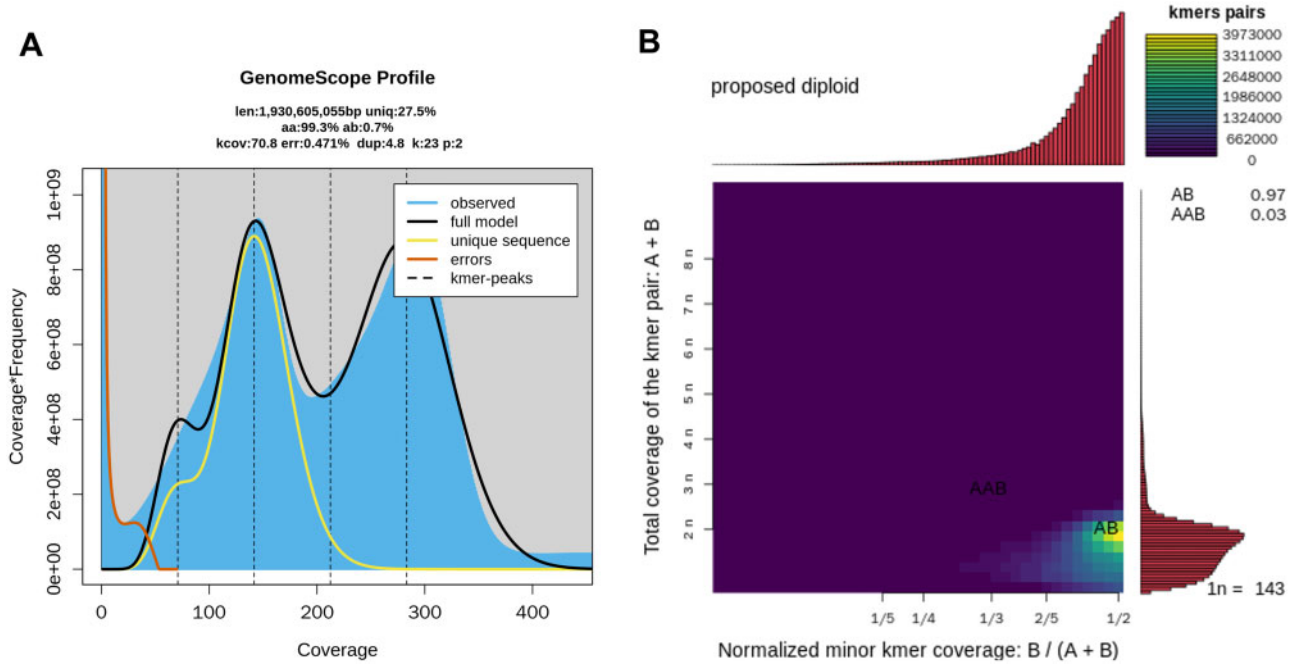

Figure 1 Characterization of the *M. japonicus* genome. (A) GenomeScope2.0 analysis using a 23-mer table built from trimmed Illumina paired-end reads (384 Gb). The *M. japonicus* genome size was estimated to be 1.93 Gb, with repetitive fraction occupying 1.40 Gb (72.5%) and a heterozygosity of 0.72%. (B) SmudgePlot ploidy analysis using the same 23-mer table with (A). No evidence of whole genome duplication was observed.

Table 1 Assembly statistics of *M. japonicus* and other penaeid shrimp genomes

Species		*M. japonicus*		*P. monodon*		*L. vannamei*	*F. chinensis*
Estimated genome size (Gb)	1.93[a]–2.25[b]	2.80[b]–2.85[b]	2.09[b]–2.45[c]	1.74[d]–1.88[d]			
Assembly	Mj_TUMSAT_v1.0	Mjap_WGS_v1	NSTDA_Pmon_1	Pmod26D_v1	ASM378908v1	ASM1692082v1	
Accession No.	GCA_017312705.1	GCA_002291165.1	GCA_015228065.1	GCA_007890405.1	GCA_003789085.1	GCA_016920825.1	
Reference	This study	Yuan *et al.* (2018)	Uengwetwanit *et al.* (2021)	Van Quyen *et al.* (2020)	Zhang *et al.* (2019)	Yuan *et al.* (2021)	
Used platforms	Illumina, ONT	Illumina	PacBio, Illumina, Hi-C	Illumina, ONT	PacBio, Illumina	PacBio, Illumina, Hi-C	
Contig statistics	Contig number	31,741	2,937,051	1,185,601	1,270,376	50,608	49,643
	Total length	1,669,100,725	1,624,778,290	2,000,783,471	1,604,130,837	1,618,035,433	1,554,870,548
	Contig N50	113,129	682	45,084	1781	57,611	59,002
	Largest contig	862,579	16,221	1,387,722	67,204	739,420	829,035
Scaffold statistics	Scaffold number	18,210	2,434,740	26,876	1,211,364	4683	8457
	Total length	1,704,994,957	1,660,270,162	2,394,347,767	1,632,388,725	1,663,581,301	1,581,041,647
	Scaffold N50	234,949	912	44,862,054	1,981	605,555	28,972,424
	Largest scaffold	1,343,847	29,048	65,869,259	93,158	3,458,385	45,565,908
	Gap content (%)	2.11	2.14	16.44	1.73	2.74	1.66
	GC (%)	34.48	39.26	36.63	38.93	36.69	37.07
	Assembly size/Estimated genome size (%)	76–88	74–87	84–86	57–58	68–79	84–91
BUSCO genome completeness (BUSCO v4.1.4, arthropoda_odb10)	Complete	946 (93.4%)	143 (14.2%)	848 (83.7%)	879 (86.8%)	890 (87.8%)	860 (84.9%)
	Complete and single-copy	918 (90.6%)	113 (11.2%)	830 (81.9%)	872 (86.1%)	835 (82.4%)	826 (81.5%)
	Complete and duplicated	28 (2.8%)	30 (3.0%)	18 (1.8%)	7 (0.7%)	55 (5.4%)	34 (3.4%)
	Fragmented	29 (2.9%)	355 (35.0%)	48 (4.7%)	66 (6.5%)	46 (4.5%)	61 (6.0%)
	Missing	38 (3.7%)	515 (50.8%)	117 (11.6%)	68 (6.7%)	77 (7.7%)	92 (9.1%)

[a] Estimated genome size obtained from this study.
[b] Estimated genome size calculated by converting the values from Swathi *et al.* (2018) with the following formula: 0.978 Gb/pg DNA (Doležel *et al.* 2003).
[c] Estimated genome size by Zhang *et al.* (2019).
[d] Estimated genome size by Yuan *et al.* (2021).

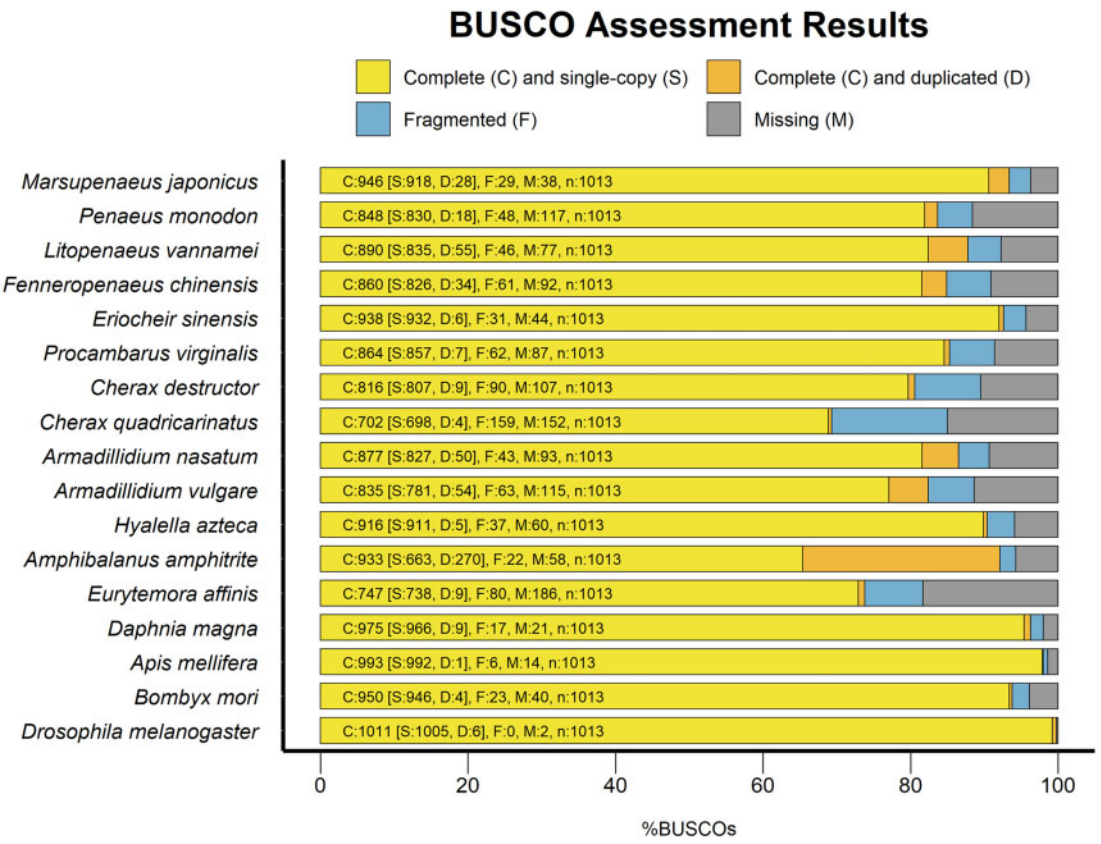

Figure 2 BUSCO scores of 17 representative arthropod genomes. The BUSCO scores of 17 representative arthropod genomes were calculated by BUSCO v4.1.4 (Simão *et al.* 2015) using the arthropoda_odb10 dataset. The GenBank accession numbers for each species are: *M. japonicus*, GCA_017312705.1; *P. monodon*, GCF_015228065.1; *L. vannamei*, GCF_003789085.1; *F. chinensis*, GCA_016920825.1; *Eriocheir sinensis*, GCA_013436485.1; *Procambarus virginalis*, GCA_002838885.1; *Cherax destructor*, GCA_009830355.1; *Cherax quadricarinatus*, GCA_009761615.1; *Armadillidium nasatum*, GCA_009176605.1; *Armadillidium vulgare*, GCA_004104545.1; *Hyalella azteca*, GCA_000764305.3; *Amphibalanus amphitrite*, GCA_009805615.1; *Eurytemora affinis*, GCF_000591075.1; *Daphnia magna*, GCF_003990815.1; *Apis mellifera*, GCF_003254395.2; *Bombyx mori*, GCF_000151625.1; *Drosophila melanogaster*, GCF_000001215.4. The values for *M. japonicus*, *P. monodon*, *L. vannamei*, and *F. chinensis* are identical to Table 1 but appear in this Figure for comparison.

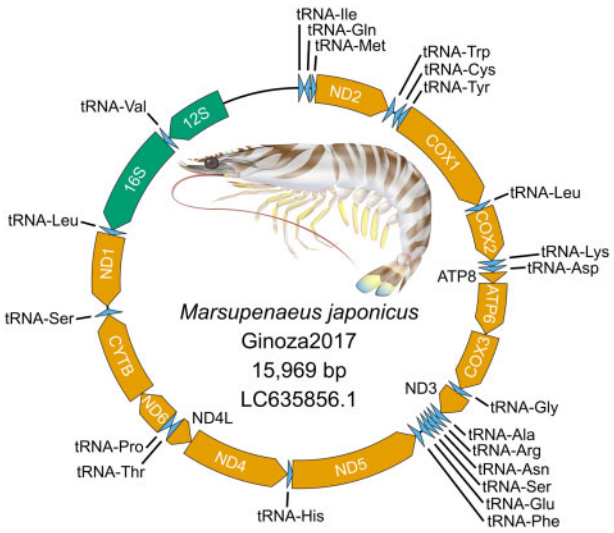

Figure 3 *M. japonicus* mitochondrial genome. Arrows indicate genes and their transcriptional orientations: orange, protein-coding genes; light blue, transfer RNA genes; green: ribosomal RNA genes.

one-to-one matches of the proteins and corresponding scaffolds. We then ran Exonerate (Slater and Birney 2005) for each protein-scaffold pair and merged the results into a single file.

Transcript evidences (PASA assemblies and StringTie predicted transcripts), protein evidences (TransDecoder CDS predicted on PASA assemblies and Exonerate protein alignments), and *ab initio* predictions (generated by BRAKER2) were merged using EVidenceModeler v1.1.1 (Haas *et al.* 2008) to derive a consensus gene model set.

We sought to filter out low-quality predicted gene models with no detectably similarity to known protein sequences. To this end, we first picked up the longest isoform from each locus using gffread v0.12.3 (Pertea and Pertea 2020), GenomeTools v1.6.1 (Gremme *et al.* 2013), and custom scripts. The nonredundant proteins were queried, using MMSeqs2 v12.113e3, against the UniRef50 protein database (49,410,134 entries; downloaded December 20, 2020) and the accessioned protein products annotated on the RefSeq genome assemblies of *P. monodon* (GCF_015228065.1_NSTDA_Pmon_1_protein.faa; 32,900 entries; downloaded December 1, 2020) and *L. vannamei* (GCF_003789085.1_ASM378908v1_protein.faa; 33,273 entries; downloaded December 1, 2020). Gene models no significant hits (*e*-value cutoff: 1e−10) were regarded as low-quality and discarded. We noticed that the gene coding for penaeidin-II (an antimicrobial peptide; GenBank Accession no. AMH87234) was missing from the gene model set, and the gene was manually added based on an Exonerate EST-to-genome alignment.

EVidenceModeler integrates predicted gene models from multiple sources to produce high-quality CDS predictions, but the

Table 2 Transcriptome assembly statistics of *M. japonicus*, *P. monodon*, and *L. vannamei*

Species		*M. japonicas*		*P. monodon*	*L. vannamei*
Accession no.	ICRK00000000.1	ICRJ00000000.1	GGLH00000000.1	GGUK00000000.1	
Reference	This study	Huerlimann *et al.* (2018)	Zeng *et al.* (2018)		
Used platforms	Illumina	PacBio	Illumina	PacBio	
Transcript number		40,991	25,415	236,085	47,260
Total length (bp)		91,091,355	64,458,168	225,712,489	162,735,114
N50 (bp)		2983	2926	1431	3978
Maximum length (bp)		38,693	9298	32,161	36,453
GC (%)		45.83	42.18	44.25	44.76
BUSCO transcriptome completeness	Complete	942 (93.0%)	684 (67.5%)	965 (95.3%)	776 (76.7%)
(BUSCO v4.1.4, arthropoda_odb10)	Complete and ingle-copy	893 (88.2%)	364 (35.9%)	487 (48.1%)	349 (34.5%)
	Complete and duplicated	49 (4.8%)	320 (31.6%)	478 (47.2%)	427 (42.2%)
	Fragmented	8 (0.8%)	21 (2.1%)	9 (0.9%)	31 (3.1%)
	Missing	63 (6.2%)	308 (30.4%)	39 (3.8%)	206 (20.2%)

Table 3 The *M. japonicus* genome annotation statistics

Number of predicted protein-coding genes		26,381
Maximum coding exon count		170
Median coding exon count		5
Average coding exon count		6.51
Median coding exon length (bp)		151.0
Average coding exon length (bp)		230.7
Median coding intron length (bp)		555
Average coding intron length (bp)		2539
BUSCO completeness of the predicted protein-coding genes	Complete	959 (94.7%)
(BUSCO v4.1.4, arthropoda_odb10)		
	Complete and single-copy	931 (91.9%)
	Complete and duplicated	28 (2.8%)
	Fragmented	19 (1.9%)
	Missing	35 (3.4%)
Number of predicted noncoding RNA genes	rRNA	81
	tRNA	2314
	snRNA	133
	scRNA	83
	snoRNA	69
	Nuclear RNase P	6
	Guide RNA	1

resulting gene models do not contain UTR information. To rescue the UTR features, the gene models were subjected to two rounds of PASA updates with the input of the Trinity transcripts, transcriptome shotgun assembly, and Iso-Seq cDNA sequences.

We sought to maximize the recovery of genes that could be predicted by transcript evidence. The shotgun transcriptome assembly and Iso-Seq cDNA sequences were aligned to the genome by Minimap2, and isoforms were collapsed by cDNA_Cupcake v24.3.0 (https://github.com/Magdoll/cDNA_Cupcake). The collapsed transcripts were structurally annotated into gene models by SQANTI3 v3.3 (https://github.com/ConesaLab/SQANTI3) (Tardaguila *et al.* 2018), and the gene models from all the sources were merged by TAMA (https://github.com/GenomeRIK/tama) (Kuo *et al.* 2020). ORFs were predicted by TransDecoder v5.5.0; here, we included homology search (queried using MMSeqs2 against UniRef50, GCF_015228065.1_NSTDA_Pmon_1_protein.faa, and GCF_003789085.1_ASM378908v1_protein.faa; *e*-value cutoff: 1e−10) as ORF retention criteria. The TransDecoder-predicted proteins were again queried using MMSeqs2 against UniRef50, GCF_015228065.1_NSTDA_Pmon_1_protein.faa, and GCF_003789085.1_ASM378908v1_protein.faa, and proteins with no significant hits (*e*-value cutoff: 1e−03) were discarded. We additionally discarded a total of six protein-coding genes on scaffold_01936 (GenBank Accession no. BOPN01001936), because they overlapped with a nuclear ribosomal RNA gene cluster coding for 18S, 5.8S, and 28S rRNAs (predicted by cmscan; see "Noncoding RNA annotation").

To assign functional annotation to the *M. japonicus* protein-coding genes, we used the Automated Assignment of Human Readable Descriptions (AHRD) pipeline v3.3.3 (https://github.com/groupschoof/AHRD) and KEGG Automatic Annotation Server (KAAS) v2.1 (accessed June 2021; https://www.genome.jp/kaas-bin/kaas_main) (Moriya *et al.* 2007).

To produce AHRD inputs, we queried the *M. japonicus* predicted protein sequences, using MMSeqs2 (*e*-value cutoff: 1e−03), against the Swiss-Prot database (563,972 entries;

downloaded December 17, 2020), GCF_015228065.1_NSTDA_Pmon_1_protein.faa, GCF_003789085.1_ASM378908v1_protein.faa, and, and publicly available *M. japonicus* soproteins downloaded from the NCBI database (802 entries; downloaded February 18, 2021). We adopted the functional descriptions that met at least two of the three criteria in the AHRD quality code. The KAAS analysis was run with default parameters for eukaryotic proteomes (query type pep; program BLAST; method BBH; genes data set hsa, mmu, rno, dre, dme, cel, ath, sce, ago, cal, spo, ecu, pfa, cho, ehi, eco, nme, hpy, bsu, lla, mge, mtu, syn, aae, mja, ape).

Noncoding RNA gene prediction

tRNA genes were predicted using tRNAScan-SE 2.0 (Chan *et al.* 2021). Predicted tRNA genes labeled as "pseudo" or overlapping with protein-coding regions were discarded. Other noncoding RNAs were predicted by Infernal cmscan v1.1.4 (Nawrocki and Eddy 2013), with the Rfam 14.5 database as reference. We used infernal-tblout2gff.pl script (https://github.com/nawrockie/jiffy-infernal-hmmer-scripts/blob/master/infernal-tblout2gff.pl) to convert cmscan output to GFF format. We first discarded spurious or ambiguous hits (*e.g.*, tRNAs, bacterial RNA families, histone H3 UTR, microRNAs, and so on.) and extracted the genomic coordinates matching the retained RNA families. The sequences from each RNA family were clustered by cd-hit-est v4.8.1 (-c 0.98 -s 0.95), and the nonredundant sequences were aligned by MAFFT v7.481 (Katoh *et al.* 2019) (accessed June 2021; https://mafft.cbrc.jp/alignment/server/). The resulting alignments were manually inspected to remove sequences of aberrant length. The remaining sequences were queried, using BLASTN, against the NCBI nonredundant nucleotide databases (Accessed June 2021), and sequences suspected to be too degenerate were discarded manually. The remaining predicted RNA genes were projected to the final annotation file.

Construction of GFF3 annotation file

The predicted protein-coding, tRNA, and other noncoding RNA genes were integrated into a GFF3-format annotation file using custom scripts. Functional annotations of the protein-coding genes were incorporated as attributes (column 9) of the corresponding CDS features. We used GenomeTools v1.6.1 (Gremme *et al.* 2013) for sorting gff3 features.

Structural annotation of Iso-Seq cDNA sequences

The Iso-Seq cDNA sequences were aligned to the genome by Minimap2 and collapsed into nonredundant isoforms by cDNA_Cupcake v24.3.0. The nonredundant isoforms were structurally annotated by SQANTI3 v3.3 with the final *M. japonicus* genome annotation as reference.

Comparison with other arthropod genome assemblies

We downloaded a total of 18 arthropod genome assemblies from the NCBI Genome database (The International Silkworm Genome Consortium 2008; Eyun *et al.* 2017; Gutekunst *et al.* 2018; Poynton *et al.* 2018; Yuan *et al.* 2018, 2021; Becking *et al.* 2019; Chebbi *et al.* 2019; Lee *et al.* 2019; Wallberg *et al.* 2019; Zhang *et al.* 2019; Kim *et al.* 2019a; Tan *et al.* 2020; Van Quyen *et al.* 2020; Larkin *et al.* 2021). The BUSCO scores were calculated by BUSCO v4.1.4 (Simão *et al.* 2015) using the arthropoda_odb10 dataset. A BUSCO analysis plot was drawn using an R script generated by generate_plot.py (https://gitlab.com/ezlab/busco/-/blob/master/scripts/generate_plot.py) (Simão *et al.* 2015).

Simple repeat contents of 16 representative arthropod genomes were estimated using RepeatMasker v4.1.1, with the following options: -noint -no_is -norna -gff -xsmall -a.

Comparison of protein-coding gene metrics in three penaeid shrimp genomes

We compared the protein-coding gene metrics (exon counts, exon lengths, intron lengths, and gene lengths, all excluding UTRs) of *M. japonicus*, *L. vannamei*, and *P. mondon*. We downloaded the NCBI RefSeq genome annotations for *P. monodon* (GCF_015228065.1_NSTDA_Pmon_1_genomic.gff; downloaded December 1, 2020) and *L. vannamei* (GCF_003789085.1_ASM378908v1_genomic.gff; downloaded December 1, 2020). We picked up the longest transcript from each locus using gffread v0.12.3, GenomeTools v1.6.1, and custom scripts. The results were visualized using ggplot2 (Wickham 2016).

Orthologous gene clustering analysis

The nonredundant proteomes of three penaeid shrimps (generated in "Comparison of protein-coding gene metrics in three penaeid shrimp genomes") were used for ortholog analysis using OrthoFinder2 v2.5.1 (Emms and Kelly 2019). We used VennDiagram (Chen and Boutros 2011) for visualization.

Results and discussion

Marsupenaeus japonicus genome sequencing, characterization, and assembly

We sequenced the *M. japonicus* genome using Illumina HiSeq4000 and ONT GridION platforms, generating a total of 723 Gb (564 Gb Illumina paired-end reads, 144 Gb Illumina mate-pair reads, and 14.6 Gb ONT long reads; Supplementary Table S2). Collectively, these sequencing data cover 375 times the estimated haploid genome size (1.93 Gb) of *M. japonicus*.

GenomeScope 2.0 analysis using Illumina paired-end reads ($k = 23$; model fit = 92.7%; Supplementary Table S4) yielded an estimated genome size of 1.93 Gb (Figure 1A), which is slightly smaller than the previously reported value [2.30 pg by Swathi *et al.* (2018), which equals 2.25 Gb assuming 0.978 Gb/pg (Doležel *et al.* 2003)]. The estimated repeat content of 72.5% and heterozygosity of 0.72% indicates that *M. japonicus* genome is highly repetitive and heterozygous (Figure 1A). SmudgePlot analysis detected no evidence of polyploidization (Figure 1B).

Error correction of Illumina paired-end reads by Tadpole reduced low-coverage k-mers and increased the coverage of homo- and heterozygous k-mer peaks, suggesting that error correction effectively purged erroneous k-mers from the data set (Supplementary Figure S3). However, this also introduced a drop of the heterozygous peak relative to the homozygous peak, which would lead to an underestimation of heterozygosity rate. We thus used the uncorrected reads for genome size estimation and ploidy analysis and used the error-correcting reads for assembly polishing.

To obtain the *M. japonicus* draft genome assembly, we first assembled filtered Illumina paired-end reads from the sperm library (93.6 Gb) and scaffolded the contigs using Illumina mate-pair reads and transcriptome data. The resulting assembly ("Illumina primary assembly" in Supplementary Table S5) consisted of 22,425 scaffolds, with the N50 of 209,993 bp and gap rate of 34.0%. The Illumina primary assembly was further improved using ONT long reads, Illumina mate-pair reads, and transcriptome data, followed by manual curation.

The final assembly (Mj_TUMSAT_v1.0) contains 18,210 scaffolds with a total length of 1,704,994,957 bp and 34.48% GC content (Table 1). Assuming a genome size of 1.93 Gb, Mj_TUMSAT_v1.0 covers 88% of the *M. japonicus* genome; however, 95.28% of muscle and 95.60% of sperm Illumina paired-end reads were successfully mapped back to the assembly. The BUSCO completeness (arthropoda_odb10 dataset; $n = 1066$) was 93.4%, indicating that the completeness of the genome assembly is high. Mj_TUMSAT_v1.0 significantly improved contiguity and BUSCO score over the previously published *M. japonicus* genome assembly [Mjap_WGS_v1.0 (Yuan *et al.* 2018) in Table 1]. Although Mj_TUMSAT_v1.0 had a lower scaffold N50 than PacBio long read-based genome assemblies of other species, the contig N50 was comparable, and the BUSCO score was higher (Table 1). The BUSCO score of Mj_TUMSAT_v1.0 was comparable to other representative arthropod genome assemblies, both in terms of completeness and low redundancy (Figure 2).

We also assembled a complete *M. japonicus* mitogenome (GenBank Accession No. LC635856, Figure 3) from ONT long reads. The *M. japonicus* Ginoza2017 mitogenome is 15,969 bp and encodes 13 protein-coding genes, 22 tRNA genes, and two rRNA genes. Pairwise BLASTN alignment showed that the *M. japonicus* Ginoza2017 mitogenome is 99.51% identical to the reference *M. japonicus* genome (NCBI Reference Sequence: NC_007010.1).

In summary, Mj_TUMSAT_v1.0 significantly improves the genomic resources of *M. japonicus* with a draft genome assembly with decent contiguity and high completeness.

Repeat identification

To characterize the repeat landscape of the *M. japonicus* genome, we constructed a *M. japonicus*-specific repeat library (4335 entries; total: 1,865,850 bp: maximum: 11,554 bp; average: 430.4 bp; minimum: 51 bp; see Supplementary Table S6 for a summary of annotated repeat elements; repeat library available as Supplementary File Mj_TUMSAT_v1.0.repeats.fna). RepeatMasker analysis using the *M. japonicus*-specific repeat library masked 54% of Mj_TUMSAT_v1.0, indicating that the kuruma shrimp is repeat-rich (Supplementary Table S7). A majority of the interspersed elements were unclassified (occupying 14.81% of the genome),

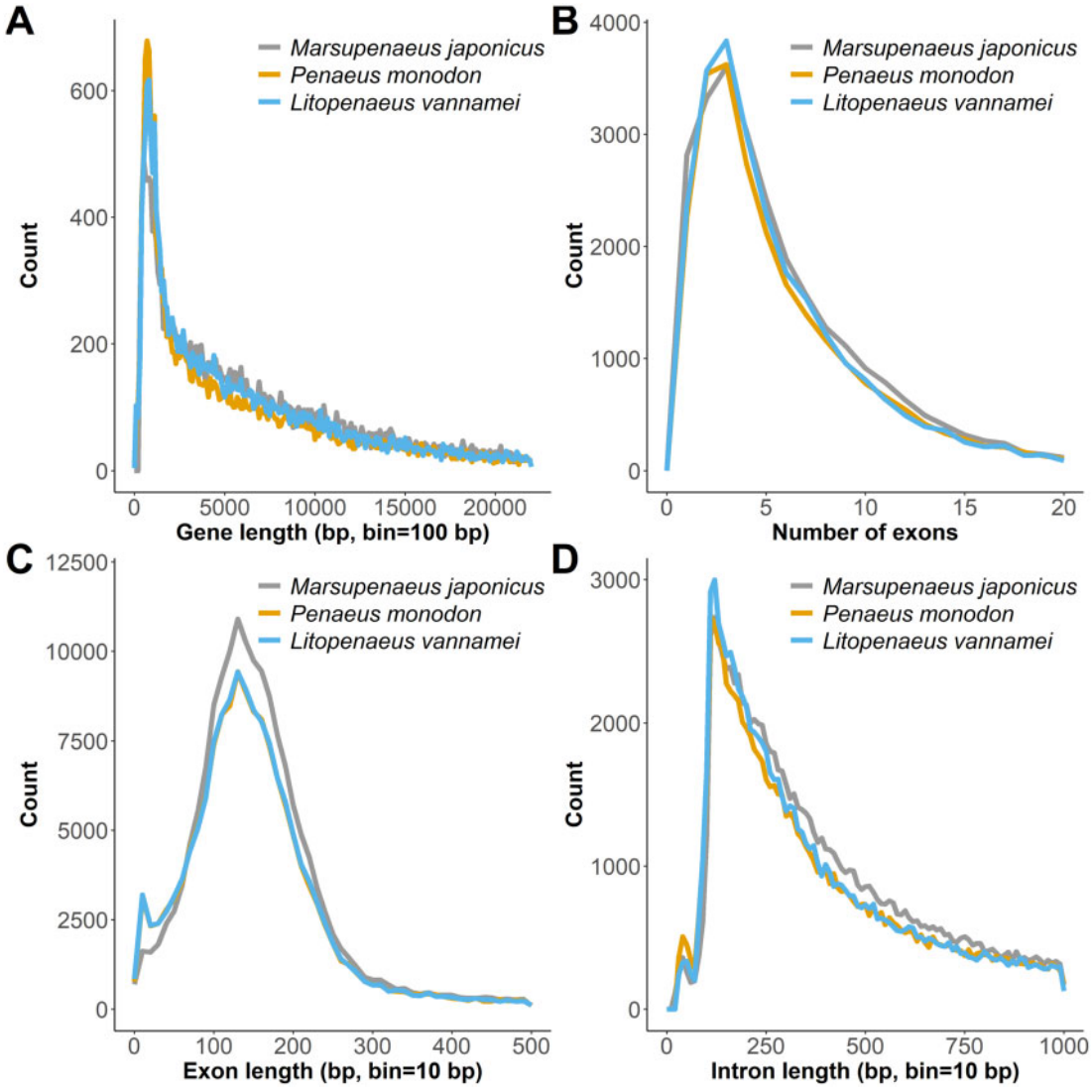

Figure 4 Protein-coding gene metrics of three penaeid shrimp genomes. The distributions of (A) gene lengths, (B) exon counts, (C) exon lengths, and (D) intron lengths were calculated for the protein-coding genes (excluding UTRs) of *M. japonicus* (26,381 genes), *P. monodon* (24,079 genes), and *L. vannamei* (24,974 genes).

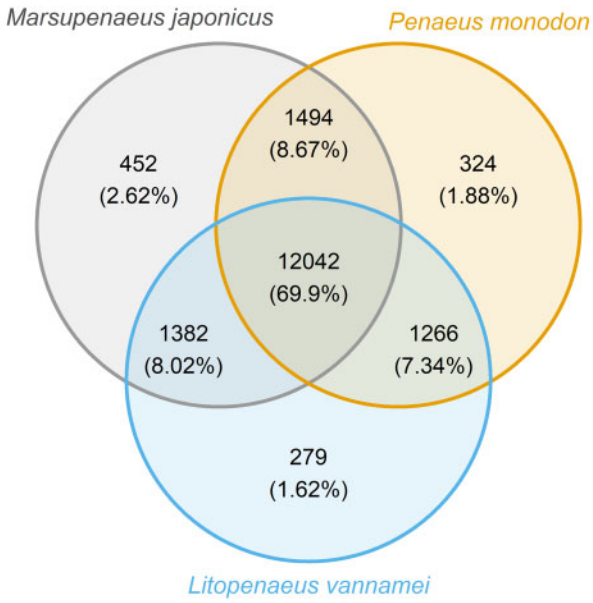

Figure 5 Orthogroup analysis of three penaeid shrimp genomes. A Venn diagram showing the number of unique and shared orthogroups among *M. japonicus*, *P. monodon*, and *L. vannamei*.

followed by retroelements (7.45%) and DNA transposons (0.06%). Simple repeats covered 27.44% of the assembly. We found similar levels of simple repeats in *L. vannamei* (34.34%), *P. monodon* (21.64%), and *F. chinensis* (31.00%), consistent with previous reports (Zhang *et al.* 2019; Uengwetwanit *et al.* 2021; Yuan *et al.* 2021) (Supplementary Table S8). Among the other arthropods analyzed, only pill bugs (*Armadillidium vulgare*, 26.37%; *A. nasatum*, 20.29%) harbored similar proportions of simple repeats (Supplementary Table S8). Overall, the *M. japonicus* genome, similar to other penaeid shrimps, is repeat-rich and is characterized by an unusually high content of simple repeats.

Marsupenaeus japonicus transcriptome sequencing and assembly

To obtain a comprehensive transcript dataset and aid genome scaffolding, we sequenced 49 RNA samples using Illumina HiSeq4000 platform and obtained a total of 346 Gb sequencing reads (Supplementary Table S2). To eliminate possible contamination, we *de novo* assembled the reads that successfully mapped back to the *M. japonicus* genome (average mapping rate: 87.25%; max: 94.26%; min: 79.22%). The transcriptome shotgun assembly consisted of 40,991 transcripts with a BUSCO completeness of 93.0% (Table 2). This BUSCO completeness was close to or above those of other penaeid shrimps (Huerlimann *et al.* 2018; Zeng *et al.* 2018) (Table 2).

We also sequenced four samples using the PacBio Iso-Seq platform, generating 25,415 transcripts with a BUSCO completeness of 67.5% (Table 2). We mapped the Iso-Seq cDNA sequences to the genome, collapsed redundant isoforms, and structurally annotated the mapped transcripts using SQANTI3 pipeline. A total of 7486 unique genes (6487 annotated, 999 novel) and 16,160 isoforms were mapped to the genome. The full SQANTI3 report and nonredundant, genome-corrected Iso-Seq cDNA sequences are provided as Supplementary Files (IsoSeq_sqanti_report.pdf, IsoSeq.aligned.nr.corrected.fna).

Genome annotation and ortholog inference

A total of 26,381 protein-coding, 2314 tRNA, 81 rRNA, and 292 other noncoding RNA genes were predicted on the *M. japonicus* genome (Table 3; Supplementary File Mj_TUMSAT_v1.0.gff3). An average *M. japonicus* protein-coding gene (excluding untranslated regions) spanned 15,490 bp containing 6.51 exons, with an average exon length of 230.7 bp and average intron length of 2539 bp (Table 3, Figure 4). The distributions of gene lengths, exon counts, exon lengths, and intron lengths were similar to those of *L. vannamei* and *P. monodon* (Figure 4). The translated *M. japonicus* predicted gene models yielded 94.7% BUSCO completeness (arthropoda_odb10 dataset; $n = 1066$) (Table 3), closely matching the completeness of the genome mode. This indicates that the predicted gene models capture most of the coding capacity of the *M. japonicus* genome. The number of predicted *M. japonicus* protein-coding genes was slightly larger than those in the NCBI RefSeq annotations for other penaeid shrimp genomes (*L. vannamei*: 24,974 genes; *P. monodon*: 24,079 genes). We compared the proteomes of *M. japonicus*, *L. vannamei*, and *P. monodon* using Orthofinder2 (Figure 5). A total of 17,239 orthogroups were identified, of which 15,370 orthogroups (89.2%) contained *M. japonicus* genes. 20,683 *M. japonicus* genes (78.4%) were clustered in orthogroups shared with at least one other shrimp species. Eight thousand and twenty-seven genes were conserved as single-copy genes across the three species. These observations indicate that most *M. japonicus* predicted genes have putative orthologs in other penaeid shrimp genomes and further substantiates the quality of the predicted *M. japonicus* gene models.

Conclusions

Here, we present a high-quality draft genome assembly and two transcriptome assemblies of the kuruma shrimp *M. japonicus*, an economically important crustacean. The 1.70 Gb draft genome assembly covers 88% of the estimated genome size (1.93 Gb) and is characterized by a high BUSCO score (93.4% complete). We predicted a total of 26,381 protein-coding gene models with a BUSCO completeness of 94.7%. The two transcriptome datasets complement the genome assembly by capturing complex transcript structures. We expect that our datasets will serve as a valuable resource for basic research, fisheries management, and breeding programs for *M. japonicus* as well as comparative genomics of penaeid shrimps and other crustaceans.

Data availability

The assembled sequences presented in this study are available in DDBJ/ENA/NCBI under the following accession numbers; genome assembly: GCA_017312705.1; mitogenome: LC635856.1; transcriptome shotgun assembly: ICRK00000000.1; PacBio Iso-Seq assembly: ICRJ00000000.1. The raw reads are available in DDBJ/ENA/NCBI under the accession numbers DRA011460, DRA011716, and DRA011525.

The supplementary figures, supplementary tables, and supplementary files (custom scripts: Mj_TUMSAT_v1.0.html and Mj_TUMSAT_v1.0.md; GFF3 annotation file: Mj_TUMSAT_v1.0.gff3; predicted transcripts: Mj_TUMSAT_v1.0.transcripts.fna; predicted CDS: Mj_TUMSAT_v1.0.cds.fna; translated CDS: Mj_TUMSAT_v1.0.prot.faa; summary table for the functional annotations of the predicted protein-coding genes: Mj_TUMSAT_v1.0.annotation.tsv; *M. japonicus*-specific repeat library: Mj_TUMSAT_v1.0.repeats.fna;

SQANTI3 report for the Iso-Seq cDNA sequences: IsoSeq_sqanti_report.pdf; genome-corrected, nonredundant Iso-Seq cDNAs: IsoSeq.aligned.nr.corrected.fna) are available on figshare: https://doi.org/10.25387/g3.15043521.

Acknowledgements

We thank the DNA sequencing section and the Scientific Computing & Data Analysis section of OIST for excellent technical support.Computation for part of this study ("PacBio Iso-Seq sequencing and analysis", "De novo assembly of Illumina genomic DNA libraries", "Transcriptome shotgun assembly", and "Scaffolding of genome assembly using transcriptome data" in Materials and Methods) was performed on the Sango high-performance computing cluster at OIST.We also thank Dr. Aiko Shitara, who played played a key role in launching the TUMSAT-OIST collaboration on this project.

Funding

This research was supported by funding from the Okinawa Institute of Science and Technology to the Marine Genomics Unit (NS), Grants-in-Aid for Scientific Research from the Japan Society for Promotion of Science (JSPS) (JSPS KAKENHI Grant Numbers JP15H02462 and JP19H00949) and Science and Technology Research Partnership for Sustainable Development from the Japan Science and Technology Agency (SATREPS Grant Number JPMJSA1806) to the Laboratory of Genome Science (IH), and Grant-in-Aid for JSPS Research Fellow from JSPS (JSPS KAKENHI Grant Number JP19J21518) to SK.

Conflicts of interest

The authors declare that there is no conflict of interest.

Literature cited

Becking T, Chebbi MA, Giraud I, Moumen B, Laverré T, *et al.* 2019. Sex chromosomes control vertical transmission of feminizing Wolbachia symbionts in an isopod. PLoS Biol. 17:e3000438.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27:573–580.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30:2114–2120.

Borodovsky M, Lomsadze A. 2011. Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. Curr Protoc Bioinformatics. 35:4.6.1–4.6.10.

Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genomics Bioinformatics. 3:1.

Bushmanova E, Antipov D, Lapidus A, Prjibelski AD. 2019. rnaSPAdes: a *de novo* transcriptome assembler and its application to RNA-Seq data. GigaScience. 8:giz100.

Bushnell B, Rood J, Singer E. 2017. BBMerge—accurate paired shotgun read merging via overlap. PLoS One. 12:e0185056.

Chan PP, Lin BY, Mak AJ, Lowe TM. 2021. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. bioRxiv. doi:10.1101/614032 (Preprint posted June 17, 2021).

Chebbi MA, Becking T, Moumen B, Giraud I, Gilbert C, *et al.* 2019. The genome of *Armadillidium vulgare* (Crustacea, Isopoda) provides insights into sex chromosome evolution in the context of cytoplasmic sex determination. Mol Biol Evol. 36:727–741.

Chen H, Boutros PC. 2011. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC Bioinformatics. 12:35.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 34:i884–i890.

Doležel J, Bartoš J, Voglmayr H, Greilhuber J. 2003. Nuclear DNA content and genome size of trout and human. Cytometry A. 51:127–128.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20:238.

Eyun S, Soh HY, Posavi M, Munro JB, Hughes DST, *et al.* 2017. Evolutionary history of chemosensory-related gene families across the Arthropoda. Mol Biol Evol. 34:1838–1862.

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, *et al.* 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci USA. 117:9451–9457.

Gremme, G Steinbiss, SKurtz, S. 2013. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. IEEE/ACM Trans Comput Biol Bioinform. 10:645–656.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, *et al.* 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 29:644–652.

Gutekunst J, Andriantsoa R, Falckenhayn C, Hanna K, Stein W, *et al.* 2018. Clonal genome evolution and rapid invasive spread of the marbled crayfish. Nat Ecol Evol. 2:567–573.

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK. Jr, *et al.* 2003. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 31:5654–5666.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, *et al.* 2008. Automated eukaryotic gene structure annotation using evidence modeler and the program to assemble spliced alignments. Genome Biol. 9:R7.

Hudinaga M. 1942. Reproduction, development and rearing of *Penaeus japonicus* Bate. Jap J Zool. 10:305–393.

Huerlimann R, Wade NM, Gordon L, Montenegro JD, Goodall J, *et al.* 2018. *De novo* assembly, characterization, functional annotation and expression patterns of the black tiger shrimp (*Penaeus monodon*) transcriptome. Sci Rep. 8:13553.

Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, *et al.* 2013. REAPR: a universal tool for genome assembly evaluation. Genome Biol. 14:R47.

Katoh K, Rozewicki J, Yamada KD. 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform. 20:1160–1166.

Kent WJ. 2002. BLAT—the BLAST-like alignment tool. Genome Res. 12:656–664.

Kim J-H, Kim H, Kim H, Chan B, Kang S, *et al.* 2019a. Draft genome assembly of a fouling barnacle, *Amphibalanus amphitrite* (Darwin, 1854): the first reference genome for Thecostraca. Front Ecol Evol. 7:465.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019b. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 37:907–915.

Kokot M, Dlugosz M, Deorowicz S. 2017. KMC 3: counting and manipulating k-mer statistics. Bioinformatics. 33:2759–2761.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 37:540–546.

Kundu R, Casey J, Sung W-K. 2019. HyPo: super fast & accurate polisher for long read genome assemblies. bioRxiv.doi: 10.1101/2019.12.19.882506 (Preprint posted December 20, 2019).

Kuo RI, Cheng Y, Zhang R, Brown JWS, Smith J, *et al.* 2020. Illuminating the dark side of the human transcriptome with long read transcript sequencing. BMC Genomics. 21:751.

Larkin A, Marygold SJ, Antonazzo G, Attrill H, dos Santos G, *et al.* 2021. FlyBase: updates to the *Drosophila melanogaster* knowledge base. Nucleic Acids Res. 49:D899–D907.

Lee B-Y, Choi B-S, Kim M-S, Park JC, Jeong C-B, *et al.* 2019. The genome of the freshwater water flea *Daphnia magna*: a potential use for freshwater molecular ecotoxicology. Aquat Toxicol. 210:69–84.

Leggett RM, Clavijo BJ, Clissold L, Clark MD, Caccamo M. 2014. NextClip: an analysis and read preparation tool for Nextera Long Mate Pair libraries. Bioinformatics. 30:566–568.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 34:3094–3100.

Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 22:1658–1659.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, *et al.* 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25:2078–2079.

Liao I-C. 1985. A brief review of the larval rearing techniques of penaeid prawns. in Proceedings of the First International Conference on the Culture of Penaeid Prawns/Shrimps, 4-7 December 1984, Iloilo City, Philippines, Aquaculture Department, Southeast Asian Fisheries Development Center, Iloilo City, Phillipines.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res. 35:W182–W185.

Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 29:2933–2935.

Nurk S, Bankevich A, Antipov D, Gurevich A, Korobeynikov A, *et al.* 2013. Assembling genomes and mini-metagenomes from highly chimeric reads. In: M Deng, R Jiang, F Sun, X Zhang, editors, Research in Computational Molecular Biology. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 158–170.

O'Connell J, Schulz-Trieglaff O, Carlson E, Hims MM, Gormley NA, *et al.* 2015. NxTrim: optimized trimming of Illumina mate pair reads. Bioinformatics. 31:2035–2037.

Paulino D, Warren RL, Vandervalk BP, Raymond A, Jackman SD, *et al.* 2015. Sealer: a scalable gap-closing application for finishing draft genomes. BMC Bioinformatics. 16:230.

Pertea G, Pertea M. 2020. GFF Utilities: GffRead and GffCompare [version 2; peer review: 3 approved]. F1000Research. 9: 304.

Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, *et al.* 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 33:290–295.

Poynton HC, Hasenbein S, Benoit JB, Sepulveda MS, Poelchau MF, *et al.* 2018. The Toxicogenome of *Hyalella azteca*: a model for sediment ecotoxicology and evolutionary toxicology. Environ Sci Technol. 52:6009–6022.

Pryszcz LP, Gabaldón T. 2016. Redundans: an assembly pipeline for highly heterozygous genomes. Nucleic Acids Res. 44:e113.

Qin M, Wu S, Li A, Zhao F, Feng H, *et al.* 2019. LRScaf: improving draft genomes using long noisy reads. BMC Genomics. 20:955.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26:841–842.

Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. 11:1432.

Roach MJ, Schmidt SA, Borneman AR. 2018. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. BMC Bioinformatics. 19:460.

Robertson G, Schein J, Chiu R, Corbett R, Field M, *et al.* 2010. *De novo* assembly and analysis of RNA-seq data. Nat Methods. 7:909–912.

Sahlin K, Vezzi F, Nystedt B, Lundeberg J, Arvestad L. 2014. BESST—efficient scaffolding of large fragmented assemblies. BMC Bioinformatics. 15:281.

Shen W, Le S, Li Y, Hu F. 2016. SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. PLoS One. 11: e0163962.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31: 3210–3212.

Slater GSC, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 6:31.

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, *et al.* 2006. AUGUSTUS: *ab initio* prediction of alternative transcripts. Nucleic Acids Res. 34:W435–W439.

Swathi A, Shekhar MS, Katneni VK, Vijayan KK. 2018. Genome size estimation of brackishwater fishes and penaeid shrimps by flow cytometry. Mol Biol Rep. 45:951–960.

Tan MH, Gan HM, Lee YP, Grandjean F, Croft LJ, *et al.*. 2020. A Giant Genome for a Giant Crayfish (Cherax quadricarinatus) With Insights Into *cox1* Pseudogenes in Decapod Genomes. Front Genet. 11:201 10.3389/fgene.2020.00201PMC: 32211032

Tardaguila M, de la Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, *et al.* 2018. SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. Genome Res. 28:396–411.

The International Silkworm Genome Consortium, 2008. The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. Insect Biochem Mol Biol. 38:1036–1045.

Uengwetwanit T, Pootakham W, Nookaew I, Sonthirod C, Angthong P, *et al.* 2021. A chromosome-level assembly of the black tiger shrimp (*Penaeus monodon*) genome facilitates the identification of growth-associated genes. Mol Ecol Resour. 21:1620–1640.

Van Quyen D, Gan HM, Lee YP, Nguyen DD, Nguyen TH, *et al.* 2020. Improved genomic resources for the black tiger prawn (*Penaeus monodon*). Mar Genomics. 52:100751.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, *et al.* 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 9:e112963.

Wallberg A, Bunikis I, Pettersson OV, Mosbech M-B, Childers AK, *et al.* 2019. A hybrid *de novo* genome assembly of the honeybee, *Apis mellifera*, with chromosome-length scaffolds. BMC Genomics. 20: 275.

Warren RL, Coombe L, Mohamadi H, Zhang J, Jaquish B, *et al.* 2019. ntEdit: scalable genome sequence polishing. Bioinformatics. 35: 4430–4432.

Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. New York, NY: Springer-Verlag.

Xu M, Guo L, Gu S, Wang O, Zhang R, *et al.* 2020. TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. GigaScience. 9:giaa094

Xue W, Li J-T, Zhu Y-P, Hou G-Y, Kong X-F, *et al.* 2013. L_RNA_scaffolder: scaffolding genomes with transcripts. BMC Genomics. 14:604.

Yamauchi MM, Miya UM, Machida JR, Nishida M. 2004. PCR-Based Approach for Sequencing Mitochondrial Genomes of Decapod Crustaceans, with a Practical Example from Kuruma Prawn (*Marsupenaeus japonicus*). Marine Biotechnology. 6:419–429.

Yuan J, Zhang X, Liu C, Yu Y, Wei J, *et al.* 2018. Genomic resources and comparative analyses of two economical penaeid shrimp species, *Marsupenaeus japonicus* and *Penaeus monodon*. Mar Genomics. 39:22–25.

Yuan J, Zhang X, Wang M, Sun Y, Liu C, *et al.* 2021. Simple sequence repeats drive genome plasticity and promote adaptive evolution in penaeid shrimp. Commun Biol. 4:186.

Zeng D, Chen X, Peng J, Yang C, Peng M, *et al.* 2018. Single-molecule long-read sequencing facilitates shrimp transcriptome research. Sci Rep. 8:16920.

Zhang X, Yuan J, Sun Y, Li S, Gao Y, *et al.* 2019. Penaeid shrimp genome provides insights into benthic adaptation and frequent molting. Nat Commun. 10:356.

Zhu B-H, Xiao J, Xue W, Xu G-C, Sun M-Y, *et al.* 2018. P_RNA_scaffolder: a fast and accurate genome scaffolder using paired-end RNA-sequencing reads. BMC Genomics. 19:175.

Communicating editor: R. Houston