Marine Genomics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Marine Genomics

journal homepage: www.elsevier.com/locate/margen

Genomic resources and comparative analyses of two economical penaeid shrimp species, *Marsupenaeus japonicus* and *Penaeus monodon*

Jianbo Yuan[a,b], Xiaojun Zhang[a,b,*], Chengzhang Liu[a], Yang Yu[a], Jiankai Wei[c], Fuhua Li[a,b], Jianhai Xiang[a,b,**]

[a] Key Laboratory of Experimental Marine Biology, Institute of Oceanology, Chinese Academy of Sciences, Qingdao 266071, China
[b] Laboratory for Marine Biology and Biotechnology, Qingdao National Laboratory for Marine Science and Technology, Qingdao 266071, China
[c] Ministry of Education Key Laboratory of Marine Genetics and Breeding, College of Marine Life Sciences, Ocean University of China, Qingdao 266003, China

ARTICLE INFO

Keywords:
Penaeid shrimps
Genome
Assembly
Annotation

ABSTRACT

Penaeid shrimps are among the most economically important crustaceans, which provide an important global food source. These species exhibit complex body plans and novelties, such as segments and appendages, which render them interesting organisms for developmental biology study of crustaceans. However, limited genomic resources have been put forward for the researches of them. Here, we report the genome sequencing and draft assembly of two economically important penaeid shrimp species, *Marsupenaeus japonicus* and *Penaeus monodon*. A total of 132.86 Gb and 132.83 Gb sequencing data was obtained in the two shrimp species. The genome assembly, a total length of 1.94 Gb and 2.04 Gb in *M. japonicus* and *P. monodon*, respectively, covers more than 97% of coding regions. We further identified 626 Mb (34.96%) and 833 Mb (46.68%) repeats, 16,716 and 18,100 genes in these two genomes, respectively. We also identified Hox genes that are important to their body plans. These data will provide valuable resources for the study of selective breeding and some plastic biological characters of penaeid shrimps, including molting, lobstering, brooding eggs and sensitization in humans.

1. Introduction

Penaeid shrimps belong to Penaeidae, a family of marine crustaceans, which includes many economical important species, such as Pacific whiteleg shrimp *Litopenaeus vannamei*, kuruma prawn *Marsupenaeus japonicus* and the giant tiger prawn *Penaeus monodon* (Koyama et al., 2010; Wilson et al., 2000; Farfante and Kensley, 1997). These species are the subject of commercial fisheries, which makes them as the valuable internationally traded commodity in aquaculture (FAO, Yearbook of Fisheries Statistics Summary Tables, 2013). Penaeid shrimps exhibit complex body plans and novelties, such as segments, appendages and lateral line-like sense organs on the antennae (Farfante and Kensley, 1997), thus, the research of them may be important for developmental biology study of crustaceans. However, to our knowledge, except for the low coverage sequencing and draft assembly of *L. vannamei* (Yu et al., 2015), *Exopalaemon carinicauda* (Yuan et al., 2017), *Parhyale hawaiensis* (Kao et al., 2016), and *Neocaridina denticulata* (Kenny et al., 2014), none of the shrimp genomes has been ultimately completed because of the large genome size and highly repetitive sequences (Yu et al., 2015; Abdelrahman et al., 2017).

Here, we provide genome sequences of two penaeid shrimps, *M. japonicus* and *P. monodon*. We performed draft genome assemblies, gene structure and repetitive sequences predictions for these two species. These data can be used for comparative genomics analyses, and provide valuable resources for shrimp genetics and breeding.

2. Data description

2.1. Sample preparation and sequencing

The nomenclature of the two penaeid shrimps, *M. japonicus* and *P. monodon*, was referred to the (ITIS) database (https://www.itis.gov/) and previous researches (Koyama et al., 2010; Wilson et al., 2000; Farfante and Kensley, 1997). The DNA was extracted from muscle of male adults using a TIANamp Marine Animal DNA Kit (TIANGEN, Beijing, China) (Table 1). Two paired-end DNA libraries with insert size of 230 bp and 500 bp were constructed following the standard Illumina operating procedure (Illumina, San Diego, CA). The paired-end

Abbreviations: MIGS, Minimum Information about a Genome Sequence; CEGMA, core eukaryotic genes mapping approach; TGICL, TIGR Gene Indices clustering tools; NCBI, National Center for Biotechnology Information; TEs, transposable elements; SSR, simple sequence repeats; SNP, single-nucleotide polymorphisms; Indels, short insertion/deletion

* Correspondence to: X. Zhang, Institute of Oceanology, Chinese Academy of Sciences, 7, Nanhai Road, Qingdao 266071, China.
** Correspondence to: J. Xiang, Institute of Oceanology, Chinese Academy of Sciences, 7, Nanhai Road, Qingdao 266071, China.
E-mail addresses: xjzhang@qdio.ac.cn (X. Zhang), jhxiang@qdio.ac.cn (J. Xiang).

https://doi.org/10.1016/j.margen.2017.12.006
Received 29 September 2017; Received in revised form 16 December 2017; Accepted 16 December 2017

Table 1
General information of *M. japonicus* and *P. monodon*.

Items	Description
General feature of classification	
Investigation type	Eukaryote
Classification	Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia; Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Crustacea; Malacostraca; Eumalacostraca; Eucarida; Decapoda; Dendrobranchiata; Penaeoidea; Penaeidae; Penaeus
Project name	Whole genome sequencing of *Marsupenaeus japonicus* and *Penaeus monodon*
Geographic location	*M. japonicus*: Nanning, Guangxi, China *P. monodon*: Shenzhen, Guangzhou, China
Latitude, longitude	*M. japonicus*: 21.83°N/108.29°E *P. monodon*: 22.35°N/114.18°E
Collection date	2015-07
Environment (biome)	Water body (ENVO:00000063)
Environment (material)	Sea water (ENVO: 00002149)
Sequencing method	Illumina HiSeq2500; Paired-end (2 × 150)
MIGS-specific mandatory descriptors	
Ploidy	Diploid
Number of replicons	*M. japonicas*: 2n = 86 chromosomes; *P. monodon*: 2n = 88 chromosomes
Estimated genome size	*M. japonicus*: 2.28 Gb *P. monodon*: 2.59 Gb
Reference of biomaterial	(Koyama et al., 2010; Wilson et al., 2000; Farfante and Kensley, 1997)
Assembly method	*De novo* assembly
Assembly program	SOAPdenovo2

Table 2
Summary of the genome assembly of two penaeid shrimp species.

	M. japonicus		*P. monodon*	
	Contig	Scaffold	Contig	Scaffold
Number:	5,632,117	3,719,281	7,106,289	4,985,320
Total length (bp):	1,924,054,682	1,942,550,811	1,882,378,599	2,035,458,477
Longest (bp):	16,221	1,606,464	12,599	1,275,042
Shortest (bp):	100	100	100	100
N50 (bp):	416	937	301	786
N90 (bp):	159	189	138	144
> 2 kb:	154,376	97,798	118,142	74,634

sequencing was performed on the Illumina HiSeq2500 platform with read length of 150 bp. The raw sequencing data were trimmed to filter out low-quality data and adapter contaminates by using the NGS QC Toolkit with the parameters of "2 A-c 10" (Patel and Jain, 2012). Finally, we collected the clean data of the two penaeid shrimps (Table S1).

2.2. Estimation of genome size, polymorphism, and repetitiveness

Genome size was estimated based on the K-mer depth distribution according to previous researches (Li et al., 2010). A major peak was observed around K-mer depth of 45 and 47 in *M. japonicus* and *P. monodon*, respectively, which corresponds to homozygous regions (Fig. S1). Genome size was estimated to be 2.28 Gb and 2.59 Gb in *M. japonicus* and *P. monodon*, respectively, which was similar to the results (C-value of 2.83 pg and 2.53 pg) from Animal Genome Size Database (www.genomesize.com/). Besides, a high proportion of K-mers with depth higher than 200 × (47.73% and 50.92% in *M. japonicus* and *P. monodon*, respectively) indicated the presence of abundant repetitive sequences.

The sequence polymorphism rate was calculated based on single-nucleotide polymorphisms (SNPs) and short insertion/deletion (Indels) according to previous researches (Kao et al., 2016). Burrows-Wheeler Aligner (BWA) was used to measure the level of heterozygosity by aligning sequencing reads to the genome (Li and Durbin, 2010). SAMtools was used to call SNPs and Indels from the alignment results (Li et al., 2009). Finally, 2,969,278 SNPs and 637,450 Indels were detected in the *M. japonicus* genome, yielding a sequence polymorphism rate of 0.19%. Besides, a sequence polymorphism rate of 0.21% (3,562,719 SNPs and 711,744 Indels) was detected in the *P. monodon* genome.

2.3. Genome assembly

A *de novo* assembly procedure was performed on the clean reads using SOAPdenovo2 with the k value set from 31 to 99 (Luo et al., 2012). And the assembly was improved by using L_RNA_scaffolder, which can use long single-end RNA-seq reads to order, orient and combine genomic fragments into larger sequences (Xue et al., 2013). Finally, a total length of 1.94 Gb scaffolds with N50 length of 937 bp were produced in *M. japonicus*; and for *P. monodon*, 2.04 Gb scaffolds with N50 length of 786 bp were obtained (Table 2), which was comparable to that of *L. vannamei* (Yu et al., 2015) and *N. denticulata* (Kenny et al., 2014).

2.4. Estimation of genome completeness

We collected the transcriptome data of two shrimp species from NCBI SRA database (accession no. of *M. japonicus*: SRX2030618; and accession no. of *P. monodon*: SRX110649, SRX110651, SRX110652, SRX1333495, SRX1333568, SRX1333569, SRX1333570, SRX757561). The transcriptome data were assembled by Trinity (Haas et al., 2013), and removed isoforms by TGICL (Pertea et al., 2003). There were 80,444 and 89,473 unigenes assembled in *M. japonicus* and *P. monodon*, respectively (Supplementary materials 2 and 3, Table S2). We downloaded 2885 ESTs of *M. japonicus* and 424 complete genes of *P. monodon* from NCBI GenBank, and compared it with the *de novo* assembled unigenes. > 90% of these sequences were covered by unigenes,

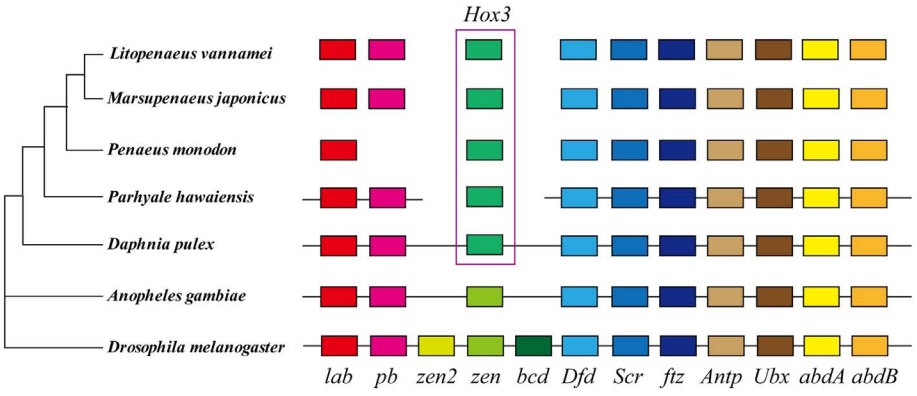

Fig. 1. Hox gene cluster of penaeid shrimps. The relevant information was referred and modified to (Yuan et al., 2017). The lines that connect each gene indicates they are synteny in the same scaffold. *zen2*, *zen* and *bcd* are three homologous genes in *D. melanogaster*, but only one homologous gene (*Hox3*) was identified in crustaceans.