The Israeli Journal of Aquaculture – Bamidgeh • ISSN 0792-156X • IJA.74.2022.1727480, 7 pages
CCBY-NC-ND-4.0 ● https://doi.org/10.46989/001c.37634

Single-molecule real-time sequencing of the full-length transcriptome of *Portunus pelagicus*

Baoquan Gao[1,2], Jianjian Lv[1,2], Xianliang Meng[1,2], Xianyun Ren[1,2], Ping Liu[1,2*]

[1] *Key Laboratory of Sustainable Development of Marine Fisheries, Ministry of Agriculture, Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao 266071, People's Republic of China*

[2] *Laboratory for Marine Fisheries and Aquaculture, Qingdao National Laboratory for Marine Science and Technology, Qingdao 266071, People's Republic of China*

Keywords: Blue swimming crab, full-length transcriptome, ISO-seq, PacBio sequencing, *Portunus pelagicus*

Abstract

Reconstruction and annotation of transcripts, particularly for a species without reference genome, plays a critical role in gene discovery, investigation of genomic signatures, and genome annotation in the pre-genomic era. This is the first study to use Single-molecule real-time (SMRT) sequencing for reporting the full-length transcriptome of *Portunus pelagicus*. Overall, 16.26 Gb of raw reads were obtained, including 7,068,387 subreads, with average length of 2,300 bp and N50 length of 3,594 bp. In total, 351,870 circular consensus sequences (CCS) reads were extracted, including 255,378 full-length non-chimeric (FLNC) reads with mean length of 3,423 bp.70,407 genes were obtained after eliminating redundant sequences, and 56,557 (80.33%) genes were annotated in at least one database, 17,267 (24.52%) genes were annotated in all of the seven databases. Further, 68,797 coding sequences (CDS) were identified, including 36,848 complete CDS. A total of 1,730 unigenes were predicted to be transcription factors (TFs). Finally, 11,894 long noncoding RNA (lncRNA) transcripts were predicted by different computational approaches and 147,262 single sequence repeat (SSR)s were obtained. The transcriptome data reported herein are bound to serve as a basis for future studies on *P. pelagicus*.

* Corresponding author. e-mail: liuping@ysfri.ac.cn, First author e-mail: gaobq@ysfri.ac.cn

Introduction

The blue swimming crab *Portunus pelagicus*, an economically important species, is widely distributed in the Indo-Pacific, particularly in tropical and subtropical waters. Due to overfishing, wild resources have witnessed rapid depletion since the 1970s (Ning et al., 2016). Accordingly, several studies have been conducted on *P. pelagicus* to investigate parasite infection, natural habitats, seedling raising, genetic breeding, and resource evaluation (Simon et al., 2003). Bishop *et al.* (1979) studied the morbid behavior of *P. pelagicus* parasitized by *Sacculina granifera* Boschma, 1973 (Cirripedia: Rhizocephala) and indicated that in comparison with noninfected crabs, infected crabs showed changes in their stance, defecation, and burying behaviors. Weng *et al.* (1992) examined biological differences between two populations of *P. pelagicus* in Queensland. They found that diverse environmental factors, such as food availability, affected the season of recruitment, molt, and gonad maturity. Furthermore, Liao et al. (2000) established the breeding technique of *P. pelagicus*, and they bred some juvenile crabs successfully. Sirawut *et al.* (2007) analyzed the genetic heterogeneity of *P. pelagicus* in Thailand to identify gene flow restriction. However, to date, the genetic and physiological processes of *P. pelagicus* remain to be reported at the molecular level, which can be attributed to the lack of genome and transcriptome information.

Several technologies have been applied for transcriptome sequencing (Yang et al., 2014), but most cannot assemble full-length (FL) transcripts. Single-molecule real-time (SMRT) sequencing developed by Pacific Biosciences (PacBio) overcomes this limitation by enabling the generation of kilobase-sized sequencing reads (Eid et al., 2009; Sharon et al., 2013). The analysis of FL transcriptome is associated with several advantages; for example, it enables the assessment of alternative splicing events, primary-precursor-mature RNA structures, and RNA processing (Jia et al., 2018).

In this study, we obtained the FL transcriptome of *P. pelagicus* for the first time using the PacBio SMRT sequencing technology. With the transcriptome sequencing data, we performed transcript functional annotation, coding sequence (CDS) prediction, long noncoding RNA (lncRNA) prediction, and single sequence repeat (SSR) analysis. We believe that the data reported herein will serve as a foundation for future studies on the genetic evolution, genetic breeding, and physiological mechanisms of *P. pelagicus*.

Materials and Methods

Sample Collection and RNA Sample Preparation

We captured 30 healthy *P. pelagicus* (108.6 g ± 8 g) from offshore waters of Xiamen, China, and reared them in an indoor closed tank (10,000-L seawater; temperature, 21°C; salinity, 30 ppt; Ph,8.0) for 7 days. After 7 days, the following tissues were extracted from six randomly picked crabs (three male and three female) and immediately frozen in liquid nitrogen: hemocyte, eyestalk, muscle, hepatopancreas, heart, stomach, gill, and thymus. The same tissue of different individuals was mixed. Total RNA was extracted separately using TRIzol. RNA samples were assessed. RNA samples from different tissues were mixed equally, and a mixed pool sample was used for single molecule FL transcriptome sequencing.

Library Preparation and SMRT Sequencing

To obtain first- and second-strand cDNA, mRNA was reverse transcribed using the SMARTer™ PCR cDNA Synthesis Kit. Then, >4 kb size selection was performed with the BluePippin™ Size-Selection System, and >4 kb cDNA was mixed in equal amounts with no-size-selection cDNA. SMRTbell™ hairpin adapters were ligated after a round of PCR and after amplified products were end-repaired. On exonuclease digestion, a cDNA library was obtained. Finally, SMRT sequencing was performed on the PacBio Sequel platform.

Data Processing

The raw sequencing reads of the cDNA libraries were analyzed by using the SMRT Link (V5.1) pipeline. The high-quality circular consensus sequences (CCSs) were extracted from the BAM subread file using the CCS function with the following parameters: minFullPass = 1 and minPredictedAccuracy = 0.80. By searching for the poly(A) tail signal and 5′ and 3′ cDNA primers in ROIs, FL and FL non-chimeric (FLNC) transcripts were identified. Iterative clustering for error correction was used to obtain FLNC consensus isoforms. Further, polished FLNC consensus isoforms were adjusted with Illumina short-read RNA-seq reads using LoRDEC to increase consensus accuracy (Salmela et al., 2014). CD-HIT was used to remove repetitive and identical sequences to obtain high-quality FLNC transcripts (Fu et al., 2012).

Gene function annotations

Gene function was annotated by BLAST v2.2.26 (Altschul et al.,1997) based on the following databases: NR (Li et al., 2002), GO (Ashburner et al., 2002), NT, Pfam, KOG/COG (Tatusov et al., 2002), KEGG (Kanehisa et al., 2002), and Swiss-Prot (Amos et al., 2000).

Gene Structure Predictions

The ANGEL (Shimizu et al., 2006) pipeline was used to identify protein-coding sequences (CDS) from cDNAs. Animal TFDB 2.0 was used to analyze transcription factors (TFs) (Zhang et al., 2006). CNCI (Sun et al., 2013), CPC (Kong et al., 2007), Pfam (Finn et al., 2016), and PLEK (Li et al., 2014) were used to predict the coding potential of transcripts. SSRs within the transcriptome were identified using MISA (http://pgrc.ipk-gatersleben.de/misa.html).

Results

Overall, 16.26 Gb of subreads were obtained, including 7,068,387 subreads, with an average length of 2,300 bp and N50 length of 3,594 bp. In total, 351,870 CCS reads were extracted, including 255,378 FLNC reads with a mean length of 3,423 bp. Further, 127,864 consensus reads were obtained with a mean length of 3,400 bp. On correcting polished transcripts using Illumina short reads, 127,864 high-quality transcripts with a mean length of 3,416 bp were obtained, which were subjected to subsequent analyses. A total of 70,407 genes were obtained after eliminating redundant sequences, among which 1,900 were 1–500 bp, 1,172 genes (1.67%) were 500–1000 bp, 4,592 genes (6.52%) were 1000–2000 bp, 18,420 genes (26.14%) were 2000–3000 bp, and 44,323 genes (62.95%) were >3000 bp **(Table 1)**.

A total of 56,557 (80.33%) genes were annotated in at least one database, and 17,267 (24.52%) genes were annotated in all seven databases. Further, 35,216 (50.02%) genes were annotated in GO, 50,915 (72.32%) in NR, 48,356 (68.68%) in KEGG, 42,680 (60.62%) in Swiss-Prot, 39,648 (56.31%) in KOG, 35,216 (50.02%) in Pfam, and 30,987 (44.01%) in NT **(Table 1)**. On functional annotation, genes were found to be associated with 53 GO terms. In the present study, a total of 35,216 unigenes were assigned to 53 sub-categories of GO terms belonging to the following three main categories: cellular component, molecular function, and biological process, which included 18, 10, and 25 sub-categories, respectively. The most abundant GO terms in the biological process category were cellular process (44.68%), metabolic process (41.33%), and single-organism process (31.97%). The most abundant GO terms in the cellular component category were cell (28.99%), cell part (28.99%), and organelle (24.49%). The most enriched GO terms in the molecular function category were binding (64.37%), catalytic activity (45.06%), and transporter activity (7.29%) **(Figure 1A)**. Sequence alignment based on the NR database revealed that approximately 18,490 (36.32%) genes were aligned to *Hyalella azteca*, 2,129 (4.18%) to *Zootermopsis nevadensis*, and 1,192 (2.34%) to *Scylla paramamosain* **(Figure 1B)**. With regard to KOG annotation, genes were divided into 26 subcategories, such as function R (general function prediction only), J (translation, ribosomal structure, and biogenesis), and O (post-translational modification) **(Figure 1C)**. Further, 360 pathways were derived from the KEGG database, with "metabolism" being the most prevalent (118, 32.78%). The highest number of genes

was assigned to "signal transduction" (6,131, 12.68%), followed by "cardiovascular diseases" (4,928, 10.19%) **(Figure 1D)**.

Table 1 Reads and annotation statistics for Iso-seq transcripts

Type	*Subreads*	*Consensus reads*	*Transcripts corrected by Illumina RNA-seq data*	*CD-HIT transcripts*
Read number and length distribution				
Total_read_number	7,068,387	127,864	127,864	70,407
Average_length (bp)	2,300	3,400	3,416	-----
<500 bp	1,886,402	2,657	2,658	1,900
500–1,000 bp	592,170	1,934	1,925	1,172
1,000–2,000 bp	932,161	12,800	12,613	4,592
2,000–3,000 bp	1,342,378	36,830	36,047	18,420
>3,000 bp	2,315,276	73,643	74,621	44,323
Functional Annotation				
GO	35,216 (50.02%)			
NR	50,915 (72.32%)			
KEGG	48,356 (68.68%)			
Swiss-Prot	42,680 (60.62%)			
KOG	39,648 (56.31%)			
Pfam	35,216 (50.02%)			
NT	30,987 (44.01%)			
Annotated_in_all	17,267 (24.52%)			
At_least_in_one	56,557 (80.33%)			
Structural				
CDS	68,797			
TF	1,730			
SSR	147,262			
lncRNA	11,894			

Using ANGEL, 68,797 CDS were identified, including 36,848 complete CDS. In total, 1,730 unigenes were predicted to be TFs by searching against all the TFs in animal TFDB 2.0. Overall, 11,894 lncRNA transcripts were predicted by all four computational approaches,86.55% lncRNAs were <4,000 nt long, and most were 2,000–3,000 nt long. Further, 127,864 high-quality transcripts were subjected to SSR analysis, which led to the identification of 147,262 SSRs. The number of mono-, di-, tri-, tetra-, penta-, and hexa-nucleotides was 44,125, 29,914, 21,152, 2,119, 458 and 210, respectively.

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://ngdc.cncb.ac.cn/, CRA006387 https://figshare.com/articles/dataset/SMRT_sequencing_of_full-length_transcriptome_of_Portunus_pelagicus/19358729.

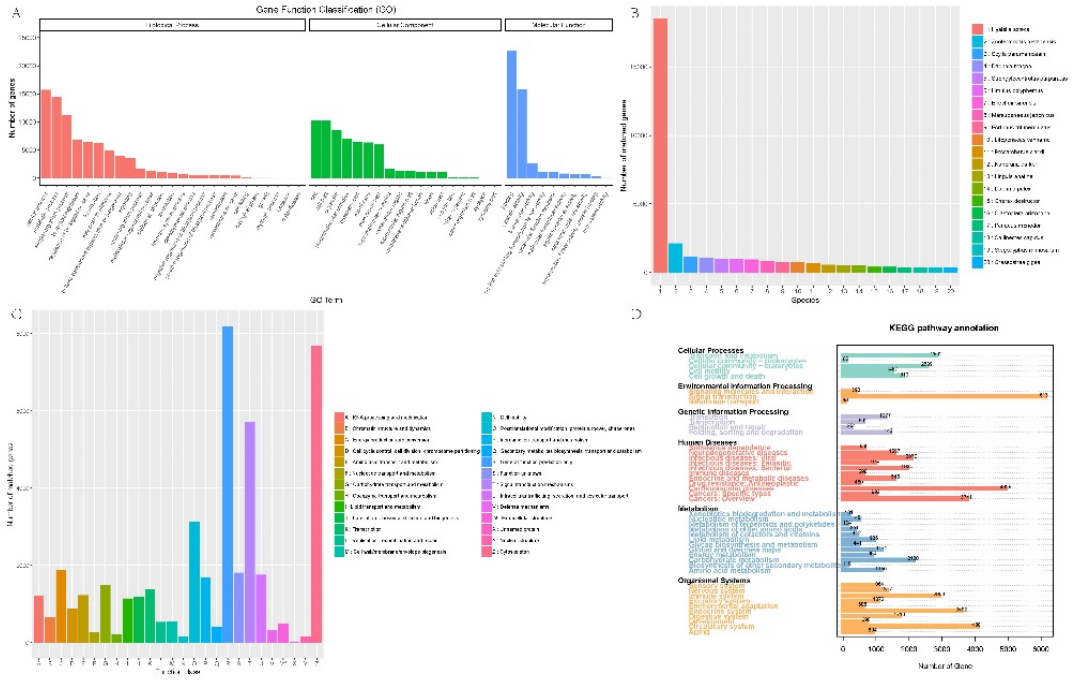

Figure 1 Characterization of functional annotation. (A) GO functional annotations. (B) NR functional annotations. (C) KOG functional

Discussion

Transcriptome reconstruction and annotation, particularly for species without a reference genome, has improved significantly with the development of sequencing techniques and plays a critical role in gene discovery, deep exploration of genomic signatures, and genome annotation in the pre-genomic era. Besides, it could be used to analyze alternative splicing events and the primary-percussor-mature RNAs structures to help better understand RNA processing. To the best of our knowledge, this is the first study to use SMRT sequencing for reporting the FL transcriptome of *P. pelagicus*. The transcriptome data reported herein are bound to serve as a basis for future studies on *P. pelagicus*. Moreover, our findings should support further studies on chromosome-level genome sequencing of *P. pelagicus* and other swimming crab species.

Acknowledgments

This study was supported by National Marine Genetic Resource Center, Central Public-interest Scientific Institution Basal Research Fund, YSFRI, CAFS (NO.20603022018024), and Central Public-interest Scientific Institution Basal Research Fund CAFS (NO. 2020TD46).

References

Altschul, S. F., Madden, T.L., Schaffer, A. A., Zhang, J.N., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*,25, 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Bairoch, A., Apweiler, R., 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res,* 28(1): 45-48. https://doi.org/10.1093/nar/28.1.45

Jia, D., Wang, Y.X., Liu, Y.H., Hu, J., Guo, Y.Q., Gao, L.L., Ma, R.Y., 2018. SMRT sequencing of the full-length transcriptome of flea beetle *Agasicles hygrophila* (Selman and Vogt). *Scientific Reports*,8:2197. https://doi.org/10.1038/s41598-018-20181-y
Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al., 2009. Real-time DNA sequencing from single polymerase molecules. *Science,* 323, 133–138. https://doi.org/10.1126/science.1162986
Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., et al., 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, 44:279-285. https://doi.org/10.1093/nar/gkv1344
Fu, L.M., Niu, B.F., Zhu, Z.W., Wu, S.T., Li, W.Z., 2012. CD-HIT: accelerated for clustering the next generation sequencing data. *Bioinformatics*, 28(23):3150. https://doi.org/10.1093/bioinformatics/bts565
Weng, H.T., 1992. The sand crab (*Portunus pelagicus* (Linnaeus)) populations of two different environments in Queensland. *Fisheries Research*, 13:407–422. https://doi.org/10.1016/0165-7836(92)90061-W
Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., Hattori, M., 2004. The KEGG resource for deciphering the genome. *Nucleic acids research,* 32(1): 277-280. https://doi：10.1093/nar/gkh063
Kong, L., Zhang, Y., Ye, Z.Q., Liu, X.Q., Zhao, S.Q., Wei, L., et al., 2007. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res*, 36: 345-349. https://doi.org/10.1093/nar/gkm391
Li, W., Jaroszewski, L., Godzik, A., 2002. Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics,* 18(1): 77-82. https://doi.org/10.1093/bioinformatics/18.1.77
Li, A. M., Zhang, J.Y., Zhou, J.Y., 2014. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. *BMC Bioinformatics,* 15:311. https://doi.org/10.1186/1471-2105-15-311
Liao, Y.Y., Zeng, J.,2000. The studied on artificial breeding of *Portunus pelagicus* in Spring. *Marine Sciences*,24(12):10-15
Michael, A., Catherine, A. B., Judith, A. B., David, B., Heather, B., Cherry, J.M., et al., 2000. Gene Ontology: tool for the unification of biology. *Nature genetics,* 25(1):25-29. https://doi.org/10.1038/75556
Ning, J.J., Du, F.Y., Li, Y.F., Gu, Y.G., Wang L.G., 2016. Dietry composition and trophic position of blue swimmer crab (*Portunus pelagicus*) in Honghai Bay. *Haiyang Xuebao*, 38(10):62-69. https://doi.org/10.1126/science.1158441
Bishop, R. K. Cannon, L. R. G., 1979. Morbid behaviour of the commercial sand crab, *Portunus pelagicus* (L.), parasitized by Sacculina granifera Boschma, 1973 (Cirripedia: Rhizocephala). *Journal of Fish Diseases*, 2(2):131-144. https://doi.org/10.1111/j.1365-2761.1979.tb00150.x
Salmela, L., Rivals, E., 2014. LoRDEC: accurate and efficient long read error correction. *Bioinformatics,* 30(24): 3506-3514. https://doi.org/10.1093/bioinformatics/btu538
Sharon, D., Tilgner, H., Grubert, F., Snyder, M., 2013. A single-molecule long-read survey of the human transcriptome. *Nature Biotechnology,* 31, 1009-1014. https://doi.org/10.1038/nbt.2705
Shimizu, K., Adachi, J., Muraoka, Y., 2006. ANGLE: a sequencing errors resistant program for predicting protein coding regions in unfinished cDNA. *Journal of bioinformatics and computational biology,* 4(3):649-664.
Lestang,S.D., Hall,N., Potter,I.C., 2003. Changes in Density, Age Composition, and Growth Rate of *Portunus Pelagicus* in a Large Embayment in Which Fishing Pressures and Environmental Conditions Have Been Altered. *Journal of Crustacean Biology*, 23, (4):908-919. https://doi.org/10.1651/C-2376
Sirawut, K., Kannika, K., Bavornlak, K., Piamsak, M., 2007. Genetic Heterogeneity of the Blue Swimming Crab *(Portunus pelagicus*) in Thailand Determined by AFLP Analysis. *Biochemical Genetics volume*, 45:725–736. https://doi.org/10.1007/s10528-007-9110-1
Sun, L., Luo, H.T., Bu, D.C., Zhao, G.G., Yu, K.T., Zhang, C.H., et al., 2013. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. *Nucleic Acids Research,* 41(17):166. https://doi.org/10.1093/nar/gkt646
Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., et al., 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4(1): 41. https://doi.org/10.1186/1471-2105-4-41

Weng, H.T., 1992. The sand crab (*Portunus pelagicus* (Linnaeus)) populations of two different environments in Queensland. *Fisheries Research*, 13:407–422. https://doi.org/10.1016/0165-7836(92)90061-W

Yang, F., Huang, L., Zhang, A., 2014. High-throughput transcriptome sequencing technology and its applications in Lepidoptera. *Acta Entomologica Sinica*. 57, 991–1000.

Zhang, H. M., Liu, T., Liu, C. J., Song, S.Y., Zhang., X.T., Liu, W., et al., 2015. AnimalTFDB 2.0: A resource for expression, prediction and functional study of animal transcription factors. *Nucleic acids research*, 43,76-81. https://doi.org/10.1093/nar/gku887