Fish and Shellfish Immunology 87 (2019) 346–359

Contents lists available at ScienceDirect

Fish and Shellfish Immunology

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

Full length article

Full-length transcriptome analysis of *Litopenaeus vannamei* reveals transcript variants involved in the innate immune system

Xiujuan Zhang, Guanyu Li, Haiying Jiang, Linmiao Li, Jinge Ma, Huiming Li, Jinping Chen*

Guangdong Key Laboratory of Animal Conservation and Resource Utilization, Guangdong Public Laboratory of Wild Animal Conservation and Utilization, Guangdong Institute of Applied Biological Resources, Guangzhou, Guangdong, 510260, China

ARTICLE INFO

Keywords:
The Pacific white shrimp
Litopenaeus vannamei
Isoform sequencing
Immune molecules

ABSTRACT

To better understand the immune system of shrimp, this study combined PacBio isoform sequencing (Iso-Seq) and Illumina paired-end short reads sequencing methods to discover full-length immune-related molecules of the Pacific white shrimp, *Litopenaeus vannamei.* A total of 72,648 nonredundant full-length transcripts (unigenes) were generated with an average length of 2545 bp from five main tissues, including the hepatopancreas, cardiac stomach, heart, muscle, and pyloric stomach. These unigenes exhibited a high annotation rate (62,164, 85.57%) when compared against NR, NT, Swiss-Prot, Pfam, GO, KEGG and COG databases. A total of 7544 putative long noncoding RNAs (lncRNAs) were detected and 1164 nonredundant full-length transcripts (449 UniTransModels) participated in the alternative splicing (AS) events. Importantly, a total of 5279 nonredundant full-length unigenes were successfully identified, which were involved in the innate immune system, including 9 immune-related processes, 19 immune-related pathways and 10 other immune-related systems. We also found wide transcript variants, which increased the number and function complexity of immune molecules; for example, toll-like receptors (TLRs) and interferon regulatory factors (IRFs). The 480 differentially expressed genes (DEGs) were significantly higher or tissue-specific expression patterns in the hepatopancreas compared with that in other four tested tissues (*FDR* < 0.05). Furthermore, the expression levels of six selected immune-related DEGs and putative IRFs were validated using real-time PCR technology, substantiating the reliability of the PacBio Iso-seq results. In conclusion, our results provide new genetic resources of long-read full-length transcripts data and information for identifying immune-related genes, which are an invaluable transcriptomic resource as genomic reference, especially for further exploration of the innate immune and defense mechanisms of shrimp.

1. Introduction

The Pacific white shrimp, *Litopenaeus vannamei*, is one of the most economically important shrimp species in the global aquaculture industry. The Food and Agriculture Organization of the United Nations published global production of this species from 2,688,901 tons in 2010 to 4,168,417 tons in 2016, becoming one of the fastest growing global foods (http://www.fao.org/home/en/). However, farmed shrimps are extremely vulnerable to various pathogens, which cause a number of problems in shrimp aquaculture; for example, three major shrimp pathogens, including acute hepatopancreatic necrosis disease (AHPND), white spot syndrome virus (WSSV) and bacteria in the genus Vibrio, result in massive mortality and devastating economic losses [1,2]. Lacking an adaptive immune system, shrimp only rely on innate immunity, which includes cellular defenses (phagocytosis, encapsulation, nodule) and humoral defenses, such as the prophenoloxidase system, hemolymph clotting mechanism and the release of antimicrobial peptides to defense against invading microbes and protect against pathogen infections [3–5]. Understanding the innate immune system of shrimp and revealing their immune responses against invading pathogens might contribute to developing strategies for the prevention and treatment of these disease, which is essential for the shrimp aquaculture industry.

The first step in investigating the key molecules of the innate immune system of *L. vannamei* is to acquire the information on the nucleotide sequences of the genes. Whole-genome sequencing and assembly combined with transcriptome data would be an efficient way to systematically characterize gene models. However, the *L. vannamei* genome is large and contains highly repetitive sequences [6,7], presenting significant difficulties in the entire genome sequencing, and there is still no available genome reference. Homology-based cDNA cloning was mainly used to amplify full-length sequences of the

[*] Corresponding author.
E-mail address: Chenjp@giabr.gd.cn (J. Chen).

https://doi.org/10.1016/j.fsi.2019.01.023
Received 28 November 2018; Received in revised form 9 January 2019; Accepted 13 January 2019
Available online 21 January 2019

shrimp's immune genes; for example, LvToll2 and LvToll3 (Toll-like receptors), MyD88 (myeloid differentiation factor 88), TRAF6 (tumor necrosis factor receptor-associated factor 6), LvTLC1 (L-type lectin), LvRelish, LvLac (laccases) and LvIL-16L (Interleukin-16-like) [8–12]. Recently, transcriptome scale characterization of genes was conducted in *L. vannamei* with high-throughput next-generation sequencing [13–15]. Although next-generation sequencing can achieve a very high throughput for transcriptomes, the limitation of short read length restricts the yield of full-length genes. However, the third generation long-read sequencing platform can overcome this difficulty.

In comparison with short-read sequencing, the methodological advantages of PacBio Isoform sequencing (Iso-Seq) mainly include better completeness in sequence both the 5′ and 3' ends of the cDNA molecules and greater accuracy in producing isoform-level transcripts. Recently, PacBio Iso-Seq technology provided a better alternative sequencing to obtain full-length cDNA molecules, which has been successively used in multiple species, such as human cell lines and tissues [16], rabbit [17], and primates [18]. However, full-length isoform transcripts and alternative splicing events still are not reported in shrimp, especially for the innate immune system.

To analyze characteristics of the innate immune molecules from *L. vannamei*, we combined PaBio Iso-Seq and Illumina short-read sequencing to obtain comprehensive full-length transcriptomes for isoform transcripts identification and quantification to generate a high-confidence isoform dataset. Subsequently, the function annotation of these full-length transcripts was systematically conducted with well-curated databases. Long noncoding RNAs (lncRNAs) and alternative splicing events were then detected. Most importantly, identification of immune-related molecules and tissue expression profiles were perfectly performed. Herein, we not only provide a valuable resource of a comprehensive full-length transcript set for the genomic reference of shrimp, but we also systematically characterized the complexity of the innate immune system of *L. vannamei*, further investigating the pathways of molecules in shrimp.

2. Materials and methods

2.1. Animals and RNA sample preparation

Twelve healthy shrimp were sampled from commercial *L. vannamei* farm in Guangdong province, China. Five main tissues (the hepatopancreas, heart, muscle, cardiac stomach and pyloric stomach) from each shrimp individual were sheared under MS222 anesthesia to minimize stress and preserved in liquid nitrogen until RNA extraction.

2.2. RNA extraction and quality evaluation

Total RNA from each tissue sample was extracted using the RNeasy Kit (Qiagen, Germany). The RNA purity was checked using the NanoPhotometer® spectrophotometer (IMPLEN, CA, USA). The RNA concentration was measured using the Qubit® RNA Assay Kit in the Qubit® 2.0 Fluorometer (Life Technologies, CA, USA). The RNA integrity was assessed using the RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, CA, USA). The RNA quality criteria for the RNA samples was RIN > 7.0 (RNA Integrity Number) and 2.0 < OD 260/280 < 2.2. Qualified RNAs were used for PacBio library preparation, Illumina library construction and sequencing. All the sequencing works were conducted at Nextomics Biosciences CO., LTD (Wuhan, China).

2.3. PacBio library construction and sequencing

To construct the library for PacBio sequencing, qualified RNA from five tissues, including the cardiac stomach, heart, hepatopancreas, muscle, and pyloric stomach, were mixed in equal amounts. The mixed RNA sample was reverse-transcribed using the SMARTer® PCR cDNA Synthesis Kit. The PCR amplification was performed using the KAPA HiFi PCR Kits. Size selection of the PCR product was performed using BluePippin Size Selection system (Sage science, USA), and the fragments with the length of 0.5–6 kb were retained. Then, the PCR product for the SMRTbell library was constructed using the SMRTbell template pre kit 1.0. A total of 30 ng of the library for each SMRT cell was sequenced using polymerase 2.0 and chemistry on the PacBio Sequel platform with 10 h of movie times. The sample information was in turn registered as BioProject (https://submit.ncbi.nlm.nih.gov/subs/bioproject/) with accession number PRJNA475443 and BioSample (https://submit.ncbi.nlm.nih.gov/subs/biosample/) with accession number SAMN09389149. Subsequently, raw consensus sequence generated by PacBio Iso-Seq sequencing was deposited into the NCBI Sequence Read Archive (SRA) with accession number SRP151627.

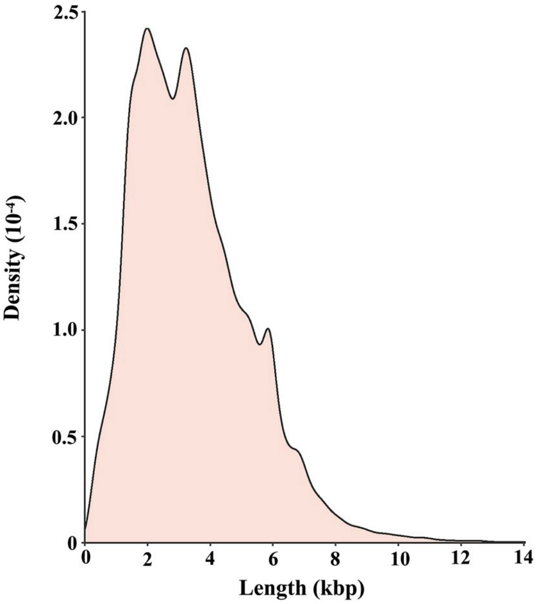

Fig. 1. Length distribution of unigenes generated by the PacBio Iso-Seq analysis in *L. vannamei*.

2.4. Illumina library construction and sequencing

The Illumina library was prepared using the NEBNext UltraTM RNA Library Prep Kit (E7530L) for Illumina (NEB, USA). Briefly, polyadenylated RNA was isolated and fragmented into ~200 bp fragments. The first-strand cDNA was synthesized using random hexamer-primers, which was followed by the synthesis of the second strand. The purified and repaired double-stranded cDNA fragments were selected by size. The amplified mRNA libraries were finally sequenced on the Illumina HiSeq X Ten platform for generating 150 bp paired-end reads. Raw reads were subjected to quality filtering using the NGS QC Toolkit v2.3.3 [19]. The first five bases from the 5' end of the read were trimmed, and the low quality bases > 20% or ambiguous bases > 1% were removed.

2.5. Error correction of PacBio Iso-Seq reads

According to the PacBio's protocol, raw polymerase reads were first processed using the SMRTlink 5.0 software. Briefly, post-filter polymerase reads were obtained after the adapter and the low-quality data were removed. The circular consensus sequence (CCS) was generated from the subread BAM files, also known as the reads of insert (ROI) [20]. All the ROIs were further classified into full-length (FL) and non full-length (nFL) transcript sequences according to whether the 5′ primer, 3′ primer and poly A tail were simultaneously observed. We

Table 1
Description of Iso-Seq in *L. vannamei* by PacBio Sequel platform.

species	SMRT cells	Polymerase reads	N50	ROIs	FLNC reads	Consensus transcripts	High-quality consensus transcripts	Low-quality consensus transcripts	Mean length (bp)	Unigenes
L. vannamei	3	1,06,08,982	2532	5,95,166	4,98,084	2,57,047	44,534	2,12,513	2545	72,648

Table 2
Statistics of unigene transcripts annotation with different databases.

Database	NO. transcripts annotated	Annotated rate (%)
NO.unigene	72,648	–
NR	59,293	81.62
NT	30,440	41.90
SwissProt	50,246	69.16
KOG	45,665	62.86
GO	41,058	56.52
KEGG	56,011	77.10
Pfam	41,058	56.52
Total	62,164	85.57

employed three-step strategies of error correction for improving the accuracy of full-length transcripts produced by the PacBio Iso-Seq platform. Firstly, the circle sequencing with > 1 passes provided the opportunity for CCSs of self-correction. Secondly, full-length, non-chimeric (FLNC) reads were subjected to non-redundant and cluster treatment by ICE Quiver algorithm and Arrow polishing with nFL sequence, herein high-quality and polished full-length consensus sequence were produced. Finally, these polished consensus sequences were further subjected to correct and remove redundancy with Illumina short reads using LoRDEC tool [21] and CD-Hit program with –c 0.95 parameter [22], respectively. After above three times corrections, nonredundant, nonchimeric full-length unigenes (isoform level) with high accuracy were yielded for subsequent analysis.

2.6. Function annotation of unigenes

For comprehensive functional annotation, the unigenes were searched against the following seven databases, including NCBI non-redundant protein sequence (NR), NCBI nonredundant nucleotide sequence (NT), Swiss-Prot, protein family (Pfam), Gene ontology (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) and Cluster of Orthologous Groups of proteins (COG). The BLAST software with an E-value $< 1 \times 10^{-10}$ was used in the NT database analysis. The Diamond BLASTX methods [23] with an E-value $< 1 \times 10^{-10}$ were analyzed in NR, COG, Swiss-Prot and KEGG annotations. The Hmmscan procedure was used in the Pfam database, and GO function categories were performed using the WEGO method [24].

2.7. LncRNA prediction

The unigenes with a length of over 200 nt were used for LncRNA prediction with the following pipelines: 1) Firstly, PLEK SVM classifier with default parameters of –minlength 200 [25] and coding-non-coding Index (CNCI) with default settings [26] were performed to assess coding potential; 2) Secondly, coding potential calculator (CPC) searched the transcripts with the NCBI eukaryotic protein database with an E-value $< 1E^{-10}$ setting [27]; 3) Finally, we translated each transcript in all three possible frames and used the Pfam Scan to identify the occurrence of any of the known protein family domains documented in the Pfam database with default parameters of -E 0.001 –domE 0.001 [28]. As a result, all isoforms transcripts with coding potential were filtered, and those without coding potential were our candidate set of lncRNAs.

2.8. Alternative splicing analysis

To detect alternative splicing (AS) events of *L. vannamei*, the error-corrected isoform transcripts (unigenes) were further processed using the Coding GENome reconstruction tool (Cogent v3.1, https://github.com/Magdoll/Cogent). Briefly, Cogent first creates the k-mer profile of the non-redundant transcripts by calculating pairwise distances and clusters transcripts into families based on their k-mer similarity. Then each transcript family is further reconstructed into one or several unique transcript models (named UniTransModels) using a De Bruijin graph method. Finally, unigenes were first mapped to the UniTransModels using the methods [29]. Splicing junctions for the transcripts that were mapped to the same splicing junctions were examined, and transcripts with the same splicing junctions were collapsed. Therefore, the unigenes with different splicing junctions that exhibited insertions/deletions (the alignment gaps) of more than 50 bp were identified as transcription isoforms of the UniTransModels according to the methods [30]. AS events were detected with SUPPA using default settings (https://github.com/comprna/SUPPA) [31].

2.9. The differential expression analysis

Using full-length isoform transcripts yielded from the SMRT Iso-Seq analysis as reference sequences, the unigene expression levels among the various tissues of *L. vannamei* were further analyzed based on the short reads datasets generated by the Illumina sequencing platform. Extracts of five tissues (muscle, heart, hepatopancreas, cardiac stomach, and pyloric stomach) from three *L. vannamei* individuals were separately used as examples of the analysis.

The expression value from the Illumina reads of each sample was determined with RSEM using default parameters [32]. Briefly, clean data from the Illumina sequencing were mapped back onto the reference sequences, and readcount values of the unigenes for each sample were obtained. To eliminate the effects of the sequencing depth and transcript length, all the readcounts were transformed into FPKM values (expected number of Fragments Per Kilobase of transcript sequence per Millions base pairs sequenced).

To detect the differential expression genes of immune-related molecules, we used the hepatopancreas as the experimental tissue, and other four tissues (muscle, heart, cardiac stomach, and pyloric stomach) were used as control samples. Therefore, we considered four combinations for comparison, including hepatopancreas vs. muscle, hepatopancreas vs. heart, hepatopancreas vs. cardiac stomach and hepatopancreas vs. pyloric stomach. All the readcounts of each sample were first normalized into a standardized readcount. Then the differential expression analysis of each comparison combination was performed using the DESeq R package (1.10.1) mode based on the negative binomial distribution from three biological replicates. The resulting *P* values were adjusted using the Benjamini and Hochberg approach for controlling the false discovery rate. Herein, the $|\log2(\text{Fold Change})| > 1$ and adjusted *P* value (FDR) < 0.05 were used as the threshold for determining differentially expressed genes (DEGs). The log2-transformed FPKM values of the DEGs were used for K-means clustering using Nbclust, an R software package, and the log10 (FPKM +1) values of the DEGs were used for hierarchical clustering using the pheatmap of the R software. The 0.1 threshold of FPKM value is regarded as expression criterion of the immune-related unigene in tested

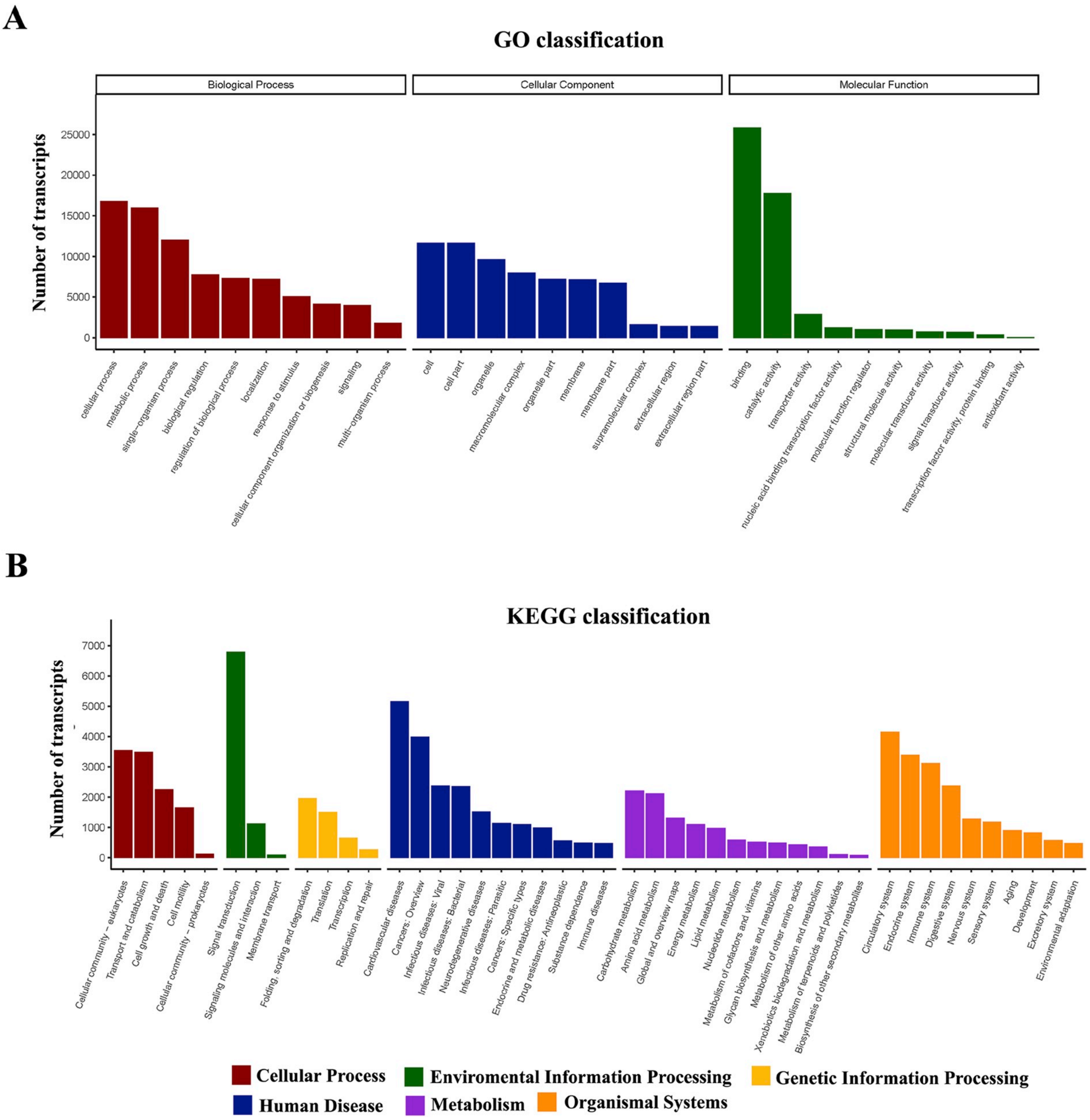

Fig. 2. Function annotation and classification of unigenes in *L. vannamei*. A) GO classification. B) KEGG classification.

tissue (FPKM > 0.1).

2.10. Transcript variations and the real-time PCR assay

The nineteen and the five full-length unigenes were respectively all annotated as putative Toll-like receptor (TLR) and interferon regulatory factor (IRF) genes. Here, the putative IRF members were in turn named as IRF1, IRF2, IRF3, IRF4 and IRF5 according to the length produced by the Iso-Seq analysis. The cDNA sequence and deduced amino acid sequence of the TLRs and IRF1-5 were analyzed using the ORFfinder program of NCBI (https://www.ncbi.nlm.nih.gov/orffinder/). The prediction for the protein structural domain of the TLRs and IRF1-5 was performed using the SMART mode (http://smart.emblheidelbergde/index2.cgi).

Total RNA was extracted using TRIzol reagent and subsequently reverse transcribed using the First-strand cDNA Synthesis Kit (TaKaRa, Tokyo, Japan) according to the manufacture's instructions. The reactions were incubated at 42 °C for 60 min, at 70 °C for 15 min and then held at 4 °C. To validate the expression pattern of DEGs, six immune-related DEGs and five putative IRFs were selected for real-time PCR assay. The six immune-related DEGs were comprised of hemocyanin (Hemy), heat shock cognate 70 (HSP70), cathepsin L (CTSL), C-type lectin (CTLC), integrin and Baculoviral IAP repeat-containing protein 8 (BIRC8) genes. Primer sequences are listed in Supplementary Table 1. Real-time PCR was performed using the Applied Biosystems QuantStudio™ 5 platform (Thermo Fisher Scientific, Waltham, USA) according to the manufacturer's protocol. A total of 0.5 μl of cDNA was used as the template in a 20 μl reaction mixture, along with 10 μl of the SYBR® Green Master Mix (Applied Biosystems, Carlsbad, USA), 0.5 μl of each primer (10 μM) and 8.5 μl ultrapure water under the following conditions: 50 °C for 2 min for UDG (Heated-labile Uracil-DNA Glycosylase) activation and then 95 °C for 2 min, followed by 40 cycles of 95 °C for 15 s, 60 °C for 30 s and 72 °C for 30 s. Each targeted gene was run in triplicate wells with three biological replicates from the five tissues of three *L. vannamei* individuals. The expression levels of targeted genes were calculated using the comparative quantity ($2^{-\triangle\triangle CT}$)

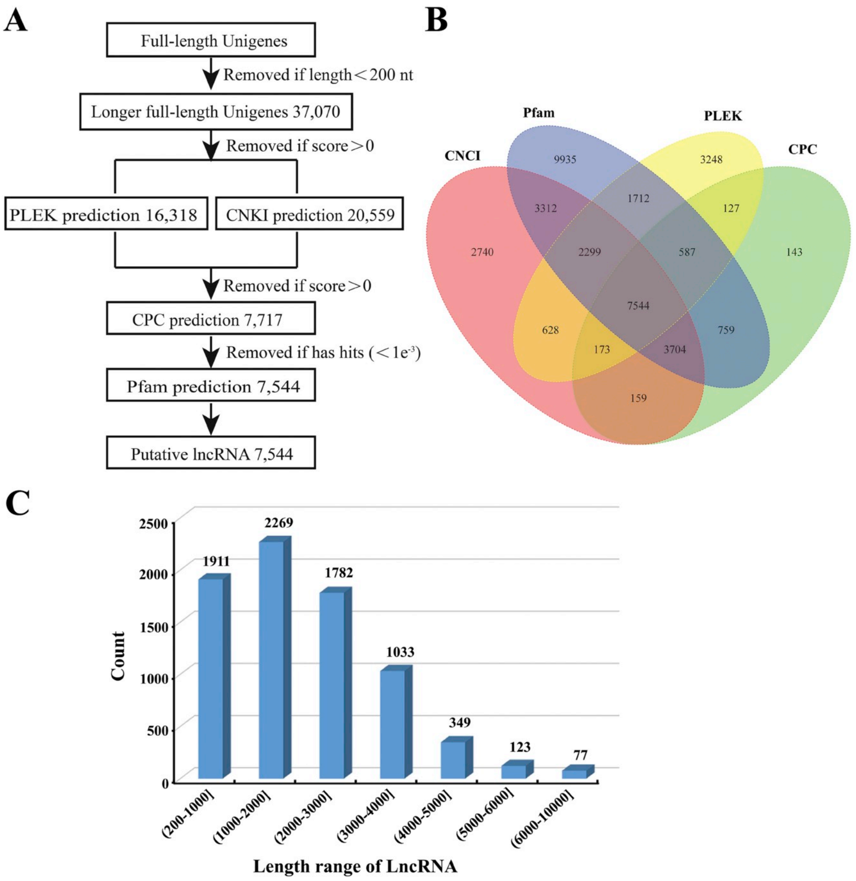

Fig. 3. Identification of long noncoding RNAs (lncRNAs) in *L. vannamei*. A) Pipeline used to identify lncRNAs. The coding potential of unigenes were in turn predicated and filtered by PLEK SVM classifier, CNCI (Coding-Non-Coding Index), CPC (Coding Potential Calculator), and Pfam Scan. B) Venn graph of lncRNAs prediction by four steps, including PLEK, CNCI, CPC, and Pfam. C) Length distribution of identified lncRNAs in *L. vannamei*.

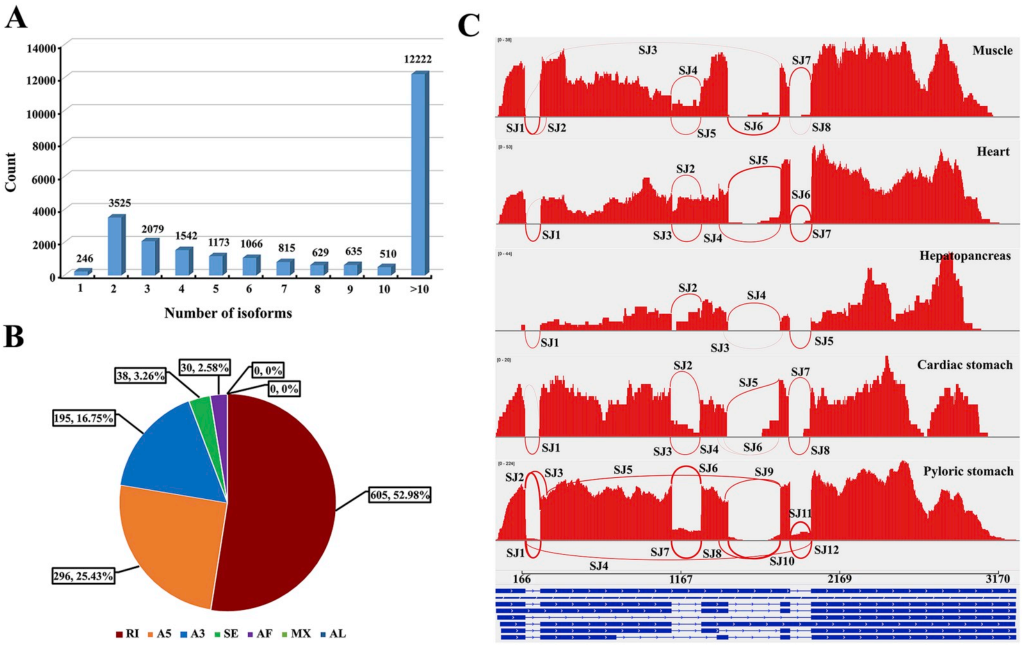

Fig. 4. Alternative splicing (AS) analysis of *L. vannamei* unigenes generated by the PacBio Iso-Seq platform. A) Distribution of isoform numbers for UniTransModels. B) Pie chart displaying types and numbers of different AS events detected in *L. vannamei*. C) Sashimi plot showing an example of the same UniTransModel gene (PB.4435_0_path3) generating seven different transcript isoforms detected with our pipeline using Iso-Seq. There are respectively eight, seven, five, eight and twelve splicing junction sites (SJ) in muscle, heart, hepatopancreas, cardiac stomach and pyloric stomach. The thickness of curve in red crossing two splicing junction sites (SJ) represents the coverage degree of Illumina RNA-Seq reads. For each isoforms, blocks in blue represent exons and lines represent introns.

method [33] after normalization to EF1α (GenBank accession NO. GU136229). The mean and standard deviation (M ± SD) was calculated from the technological and biological replicates. The statistical significance was measured using the independent samples *t*-test by SPSS 17.0, with $P < 0.05$ indicating significance.

3. Results

3.1. The full-length sequences of L. vannamei using PacBio sequencing

The full-length transcriptome of *L. vannamei* was generated using the PacBio Sequel platform on the pooled RNA of five main tissues, including the cardiac stomach, heart, hepatopancreas, muscle, and

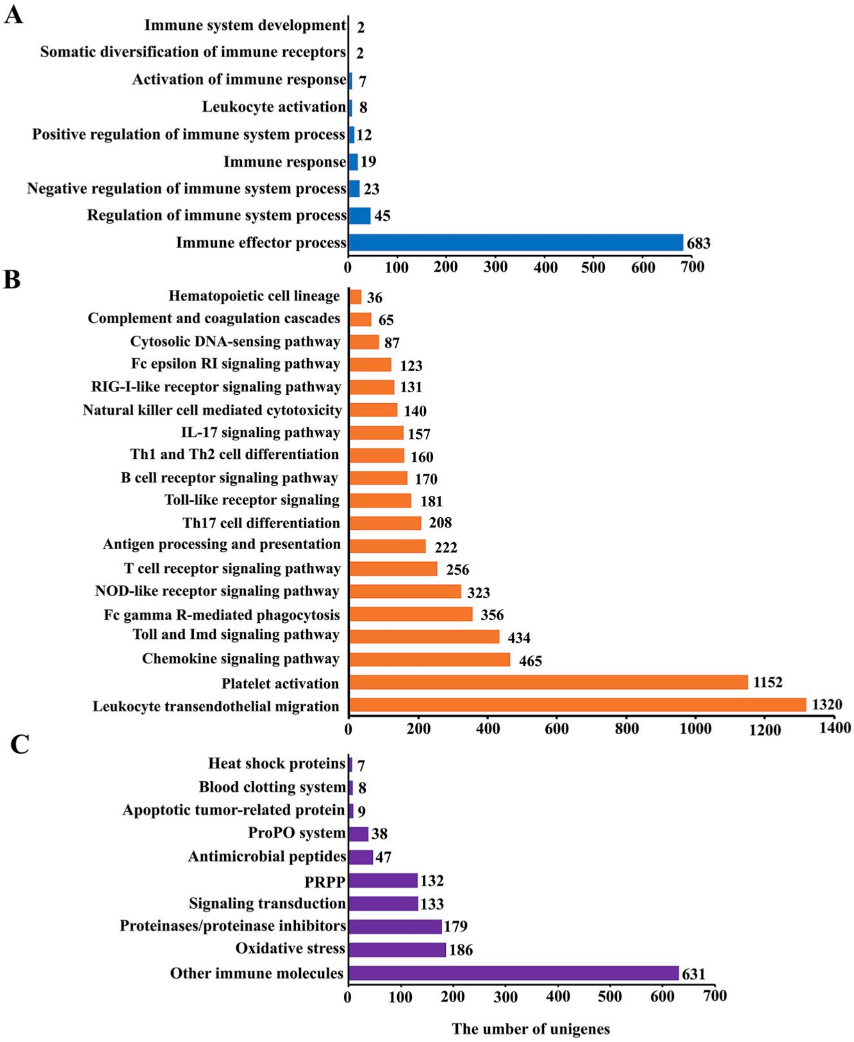

Fig. 5. Categories of 5279 nonredundant unigenes involved in the innate immune system. A) Immune system process (GO annotation). B) Immune system pathway (KEGG annotation). C) Other immune molecules in shrimp immunity.

pyloric stomach. As a result, a total of 21.41 G subreads bases were generated by three SMRT cells from the PacBio library; therefore, the 10,608,982 polymerase reads were produced with mean passes of 11 circles (Supplementary Table 2). After self-correction among subreads, a total of 595,166 ROIs were successfully extracted. By applying the standard Iso-Seq classification and clustering protocol, all the ROIs were further classified into 70,123 nonfull-length (nFL) sequences and 498,084 full-length nonchimeric (FLNC) reads with mean length of 2635 bp. Based on the algorithm of the ICE Quiver and Arrow polishing, we produced 257,047 polished full-length consensus transcripts with a mean length of 2545 bp, including 44,534 high-quality (HQ) and 212,513 low-quality (LQ) sequences (Supplementary Table 3). After correction using short reads produced by Illumina sequencing and subsequently removing redundancy via the CD-Hit program, the consensus transcripts were finally clustered into a total of 72,648 unigenes (isoform level) for subsequent analysis. We found that 89.69% of the unigenes had a main range of 0.5–6 k in length distribution (Fig. 1). Statistics of Iso-Seq in *L. vannamei* by PacBio Sequel platform are listed in Table 1.

3.2. Efficient gene annotation of L. vannamei

To obtain a comprehensive functional annotation of the *L. vannamei* transcriptome, we annotated 72,648 unigenes with seven databases, including NR, NT, Pfam, SwissProt, KOG, GO, and KEGG. A total of 85.57% of the unigenes (62,164 of 72,648) were successfully annotated with significant hits (E-value $< 1E^{-10}$) in these well-curated databases; 19,488 unigenes were collectively annotated among the seven databases. The statistics of the unigenes' annotation are listed in Table 2. The remaining numbers of unannotated unigenes (10,484 unigenes) might represent novel *L. vannamei* species-specific genes.

Among the annotated 54 classified GO terms, cellular process was identified as the most common annotation in the biological process, metabolic process and single-organism process were the next most abundant GO terms. In addition, we also found that a low proportion of unigenes were annotated into immune system process (793 unigenes), reproduction (449 unigenes) and reproduction process (436 unigenes) terms. In the molecular function and cellular component categories, binding and cell part annotations were identified as the most abundant terms. The GO classification of the unigenes in *L. vannamei* are shown in Fig. 2A and Supplementary Table 4.

In the KEGG classification, human diseases, organismal systems, and cellular processes were the top three categories with higher proportions (Fig. 2B and Supplementary Table 5). Briefly, a total of 20,178 (36.03%) nonredudant unigenes were involved in human disease related pathways, in which 2352 unigenes were predicted to infectious disease: viral and 2380 unigenes were predicted to infectious disease: bacterial. A total of 32.71% of the annotated unigenes were classified as belonging to organismal systems related pathways, in which the circulatory system (4158 unigenes), endocrine system (3389 unigenes)

Table 3
The immune-related unigenes involved in Toll and Imd signaling pathway in *L.vannamei*.

Gene Name and Description	NO. unigenes
NFKB1,Relish, nuclear factor NF-kappa-B p105 subunit	20
TAK1,transforming growth factor beta-activated kinase 1; MAP3K7,mitogen-activated protein kinase kinase kinase 7	6
Effete,UBE2D, ubiquitin-conjugating enzyme E2 D	10
Ankyrin, ANK ankyrin	180
FBXW1_11, BTRC, beta-TRCP F-box and WD-40 domain protein 1/1	9
Uev1A, UBE2V ubiquitin-conjugating enzyme E2 variant	5
DUOX, THOX dual oxidase	8
Toll	19
IAP2, BIRC2_3 baculoviral IAP repeat-containing protein 2/3	42
TAB2 TAK1-binding protein 2	4
JNK c-Jun N-terminal kinase	7
P38 p38 MAP kinase	14
MODSP modular serine protease	1
SPZ protein spaetzle	4
IRAK4 interleukin-1 receptor-associated kinase 4	5
REL c-Rel proto-oncogene protein	10
MAP3K4, MEKK4 mitogen-activated protein kinase kinase kinase 4	2
FAF1 FAS-associated factor 1	4
MAP2K7, MKK7 mitogen-activated protein kinase kinase 7	2
IMD immune deficiency	1
ATF2, CREBP1 cyclic AMP-dependent transcription factor ATF-2	17
CASP8 caspase 8	3
JUN transcription factor AP-1	1
FOSLN fos-like antigen	14
NFKBIA NF-kappa-B inhibitor alpha	36
IRAK1 interleukin-1 receptor-associated kinase 1	4
MYD88 myeloid differentiation primary response protein	1
UBE2N, BLU, UBC13 ubiquitin-conjugating enzyme E2 N	8
MAP2K3, MKK3 mitogen-activated protein kinase kinase 3	4

and immune system (3117 unigenes) were the top three pathways with the most abundant unigenes. In addition, a total of 19.80% of the annotated unigenes might participate in cellular processes pathways. In environmental information processing pathways, 84.62% of the unigenes were predicted to be involved in signal transduction. Meanwhile, the lower percentage of the annotated unigenes (7.86%) was associated with genetic information processing.

3.3. LncRNA identification from full-length transcriptomes of L. vannamei

We identified lncRNAs from these Iso-Seq data sets using the customized filtering pipeline (Fig. 3A). Mainly, after the removal of coding potential transcripts (score > 0), 16,318 and 20,559 unigenes were retained from 37,070 full-length unigenes with the length of over 200 nt by PLEK and CNKI prediction, respectively. After filtering of the CPC prediction, there were 7717 unigenes for further procedure. Finally, a total of 7544 unigenes were identified as putative lncRNAs from *L. vannamei* (Fig. 3B and Supplementary Table 6). From the statistics of length distribution, almost all of these lncRNAs (92.72%) were less than 4000 nt long, and the length range from 1000 nt to 2000 nt was the most abundant distribution of lncRNAs (Fig. 3C).

3.4. Alternative splicing analysis from full-length transcriptomes of L. vannamei

As L. *vannamei* had no reference genome, we used Congent tool to further divide the full-length transcripts into clustering families and reconstruct each family into one transcript model or several full-length unique transcript models (named as UniTransModels) based on K-mer clustering and De Bruijin graph methods. As a result, 44,354 full-length transcripts were divided into transcript families, and therefore a total of 24,442 UniTranModels were yielded. Among these, the 98.99% of UniTransModels had more than one isoform, and it is interesting to note that the half of UniTransMoldels had more than 10 isoforms (Fig. 4A).

Full-length transcripts produced by the PacBio sequel platform permit us to explore the complexity of the alternative splice (AS) at transcriptome scale. Herein, we described the specific types of AS events using the UniTransModels as the sequence reference. As a result, a total of 2768 UniTransModels–based AS events in *L. vannamei* were detected, among which, a total of 1164 nonredundant full-length transcripts (only 449 UniTransModels) participated in AS events. Briefly, retained introns (RI) were identified as the most abundant AS events, accounting for 52.98% of all events. The type of alternative 5′ splice sites (A5) accounting for 25.43% and 195 AS events (16.75%) with alternative 3′ splice sites (A3) were followed. Meanwhile, we also found that there were no AS events for the types of mutually exclusive exons (MX) and alternative last exons (AL). The overview of AS events in *L. vannamei* is shown in a pie chart (Fig. 4B). Subsequently, we emphasized different splicing junction sites of the same UniTransModel (Fig. 4C). This suggests a potentially substantial role for alternative splicing in regulating gene expression in *L. vannamei*.

3.5. Identification and function analysis of immune-related molecules in L. vannamei

In advance, immune-related molecules are stored in tissue and/or organs of shrimp before activation by cell wall components of pathogens, such as peptidoglycan (PG), lipopolysaccharides (IPS) and β-glucans (BGs). Pattern recognition proteins (PRPs) or pattern recognition receptors (PRRs) recognize and bind these components and then activate various immune response. Herein, GO and KEGG functional annotation and classification of unigenes allowed us to obtain a comprehensive immune characteristics from *L. vannamei.* As a result, a total of 793 nonredundant unigenes involved in nine immune processes were extracted from GO annotation. Among these, immune effector process was the main category, followed by regulation of immune system process and negative regulation of immune system process (Fig. 5A). KEGG pathway also revealed that other 3117 nonredundant immune-related unigenes were predicted. The 19 immune system related pathways were exacted, such as Toll and Imd signaling pathway, NOD-like receptor signaling pathway and IL-17 signaling pathway (Fig. 5B). To discover more immune molecules, we further searched the immune molecules participating in the innate immune system of shrimp summarized in the literature review [1], including antimicrobial peptides, serine proteinases and inhibitors, phenoloxidases, oxidative enzymes, clottable proteins, pattern recognition proteins, lectins, Toll receptors, and other humoral factors. These immunity molecules were mainly discovered by high throughput genomic/proteomic approaches, for example, the expressed sequence tag (EST); however, the PacBio sequel platform permitted us to discover various full-length transcripts of the immune-related molecules. Furthermore, other immune molecules from recently reported references, for example, L-type lectin [10] and interferon regulatory factor (IRF) [34] were also considered. Therefore, a total of 1370 additional immune-related molecules were identified, and the detailed function categories are listed in Fig. 5C. In total, 5279 nonredundant immune molecules were identified from the full-length transcripts of *L. vannamei*. The details of the full-length immune-related uningenes are listed in Supplementary Table 7.

3.6. Toll and Imd signaling pathway and putative transcript variants

We found that the sequence variation of immune molecules increased the numbers of unigenes; for example, we detected 29 nonredundant immune molecules predicted from 441 unigenes in the Toll and Imd signaling pathway (Table 3). Toll-like receptors (TLRs) are an one important class of host pattern recognition receptors (PRRs), and are now considered to be the primary sensor of pathogens in all metazoans. We found that 19 unigenes were annotated into putative TLR

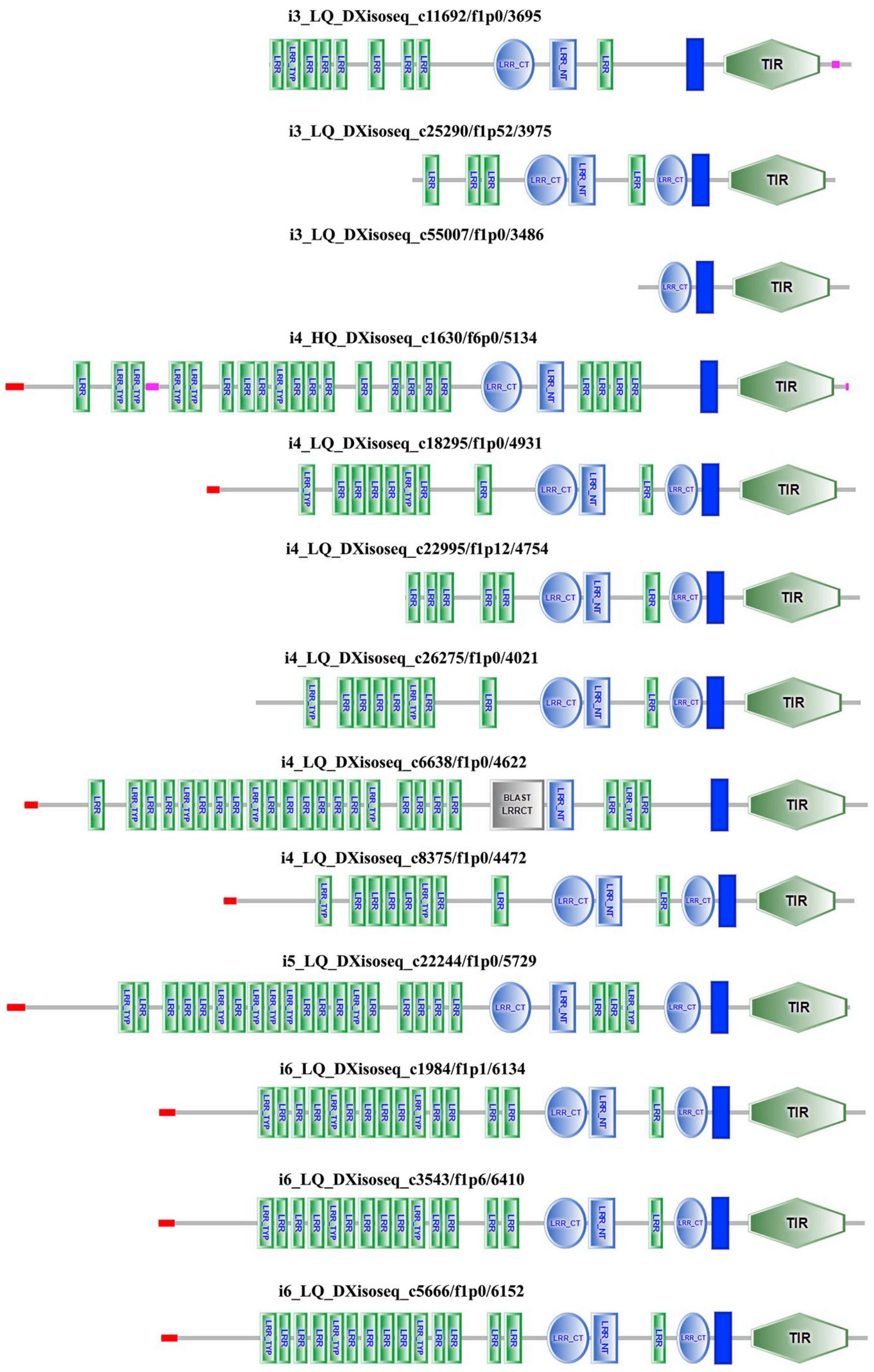
i3_LQ_DXisoseq_c11692/f1p0/3695
i3_LQ_DXisoseq_c25290/f1p52/3975
i3_LQ_DXisoseq_c55007/f1p0/3486
i4_HQ_DXisoseq_c1630/f6p0/5134
i4_LQ_DXisoseq_c18295/f1p0/4931
i4_LQ_DXisoseq_c22995/f1p12/4754
i4_LQ_DXisoseq_c26275/f1p0/4021
i4_LQ_DXisoseq_c6638/f1p0/4622
i4_LQ_DXisoseq_c8375/f1p0/4472
i5_LQ_DXisoseq_c22244/f1p0/5729
i6_LQ_DXisoseq_c1984/f1p1/6134
i6_LQ_DXisoseq_c3543/f1p6/6410
i6_LQ_DXisoseq_c5666/f1p0/6152
LRR
LRR_TYP
LRR_CT
LRR_NT
BLAST LRRCT
TIR

Fig. 6. Schematic representations of the domain topology of putative TLR genes annotated from full-length unigenes according to SMART analysis.

genes of *L. vannamei*, which belonged to five numbers of TLR family, including TLR1 (5 unigenes), TLR2 (7 unigenes), TLR3 (5 unigenes), TLR4 (1 unigene) and TLR7 (1 unigene). The function domain analysis indicated that these putative TLR genes adopted typical domain organization characteristics of the Toll family gene, including leucine rich repeat (LRR), leucine rich repeat C-terminal domain, transmembrane domain or conserved TIR domains. Furthermore, thirteen putative TLR genes (not all from 19 unigenes) collectively contained conserved TIR domains (Fig. 6). The significant differences both in number of extracellular LRRs and sequence identities of intracellular TIR domains may suggest that these putative TLR genes may respond to various extracellular stresses and use different intercellular signaling pathways. We also screened five full-length transcripts as putative IRF members of IRF family, named IRF1-5, according to the length of the Iso-Seq. Details of the TLR and IRF genes, putative unigene variants and conserved domain information are listed in Table 4.

Herein, the full length of the IRF1 transcript is 1370 bp, comprising a 134 bp 5′-untanslated region (5′-UTR), a 612 bp 3′-UTR and a 624 bp ORF encoding a 207 amino-acid region. The full length of the IRF2 transcript is 1432 bp, comprising a 397 5′-untranslated region (5′-UTR), a 816 bp 3′-UTR and a 219 bp ORF encoding a 72 amino-acid region. The full length of the IRF3 transcript is 1749 bp, comprising a 192 5′-untranslated region (5′-UTR), a 468 bp 3′-UTR and a 1089 bp ORF encoding a 362 amino-acid region. The full length of the IRF4 transcript is 2086 bp, comprising a 100 5′-untanslated region (5′-UTR), a 1362 bp 3′-UTR and a 624 bp ORF encoding a 207 amino-acid region. The full length of the IRF5 transcript is 2245 bp, comprising a 768 5′-untanslated region (5′-UTR), a 470 bp 3′-UTR and a 1008 bp ORF

Table 4
Details of TLR and IRF genes, putative unigene variants and conserved domain information.

Candidate genes	Symbol	Accession number	Species	Source	Length (aa)	Putative transcript variants from Iso-seq sequencing analysis	Length of transcripts (bp)	Length of ORF (aa)	Conserved domain
Toll-like receptor 1	TLR1	ABK58729.1	*Litopenaeus vannamei*	CDS	926	i4_LQ_DXisoseq_c18295/f1p0/4931	4938	926	YES
						i4_LQ_DXisoseq_c19089/f1p37/4988	5018	98	
						i4_LQ_DXisoseq_c26275/f1p0/4021	4025	862	YES
						i4_LQ_DXisoseq_c4857/f1p1/4349	4470	417	
						i4_LQ_DXisoseq_c8375/f1p0/4472	4471	900	YES
Toll-like receptor 2	TLR2	AEK86516.1	*Litopenaeus vannamei*	CDS	1008	i3_LQ_DXisoseq_c25290/f1p52/3975	4094	603	YES
						i3_LQ_DXisoseq_c55007/f1p0/3486	3499	301	YES
						i4_LQ_DXisoseq_c22995/f1p12/4754	4748	648	YES
						i6_LQ_DXisoseq_c1984/f1p1/6134	6407	1008	YES
						i6_LQ_DXisoseq_c3543/f1p6/6410	6431	1008	YES
						i6_LQ_DXisoseq_c5666/f1p0/6152	6160	1008	YES
						i6_LQ_DXisoseq_c5855/f1p13/6032	6318	772	
Toll-like receptor 3	TLR3	AEK86517.1	*Litopenaeus vannamei*	CDS	1244	i2_LQ_DXisoseq_c86490/f1p12/2148	2147	459	
						i3_LQ_DXisoseq_c11692/f1p0/3695	3777	830	YES
						i4_HQ_DXisoseq_c1630/f6p0/5134	5135	1227	YES
						i5_LQ_DXisoseq_c28416/f1p0/5108	5180	1052	
		AOF79112.1	*Marsupenaeus japonicus*	CDS	811	i5_LQ_DXisoseq_c22244/f1p0/5729	5737	1402	YES
Toll-like receptor 4	TLR4	AOF79113.1	*Marsupenaeus japonicus*	CDS	1352	i4_LQ_DXisoseq_c6638/f1p0/4622	4621	1349	YES
Toll-like receptor 7	TLR7	AOF79116.1	*Marsupenaeus japonicus*	CDS	880	i1_LQ_DXisoseq_c23267/f1p6/1749	1749	582	
interferon regulatory factor	IRF	AKG54423.1	*Litopenaeus vannamei*	CDS	362	i1_LQ_DXisoseq_c110860/f1p0/1369	1370	207	
						i1_LQ_DXisoseq_c160130/f1p0/1450	1432	72	
						i1_HQ_DXisoseq_c31822/f6p0/1749	1749	362	YES
						i2_LQ_DXisoseq_c178909/f1p0/2122	2086	207	YES
						i2_LQ_DXisoseq_c63793/f1p0/2244	2245	335	

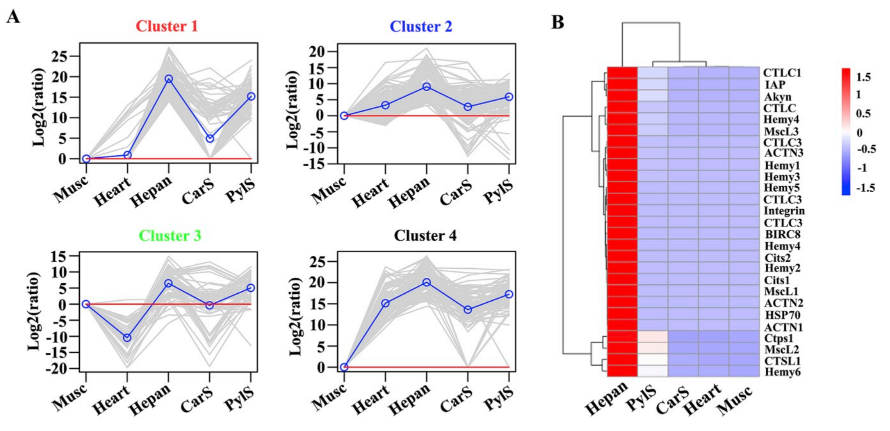

Fig. 7. Differentially expressed genes (DEGs) analysis of immune-related unigenes. A) K-means clustering of 480 DEGs from four comparison combinations union showing the tissues expression profiles. Muscle (Musc) was regarded as reference tissue. Line in red is reference of relative expression ratio. Above of red line represent relatively up-regulated transcripts, and under of red line represents relatively down-regulated transcripts. The X-axis represents five different tissues, and the y-axis represents the relative expression level (Log2 (ratio) transformed data). B) Hierarchical clustering of 27 DEGs from intersections of four comparison combinations displaying expression profile in five tissues. log10 (FPKM +1) values were used. DEGs were identified from three biological replicates. Hepan: hepatopancreas; Musc: muscle; Heart: heart; CarS: cardiac stomach and PylS: pyloric stomach.

encoding a 335 amino-acid region. The information of the IRF 1–5 sequences was listed in Supplementary File 1. The ORF sequences of the putative IRF 1–5 transcripts exhibited 67.4%, 79.2%, 95.3%, 89.1% and 93.2% similarity, respectively, to that of the IRF of *L. vannamei* [34]. The SMART analysis indicated that only IRF3 and IRF4 had a highly conserved DNA-binding domain in the amino-terminal regions that is homologous to that of the IRF of *L. vannamei*.

3.7. Differential expression analysis of immune-related molecules

To identify key immune molecules, the differently expressed genes (DEGs) with high expression and tissue-specific expression in the hepatopancreas (Hepan) were screened from the following four comparison combinations: Hepan vs. muscle (Musc), Hepan vs. heart (Heart), Hepan vs. cardiac stomach (CarS) and Hepan vs. pyloric stomach (PylS).

A total of 480 DEGs were screened from four comparison combinations union from 5279 immune-related molecules (Supplementary Table 8). Using the K-means clustering algorithm, the 480 DEGs were classified into the following four subclusters (Cluster 1–4): I) 96 transcripts (account for 20%) with lower expression in Musc and Heart, but the highest expression level in Hepan, followed by PhyS and CarS; II) nearly half of the DEGs union (228 transcripts, account for 47.5%) with gentle rising trend in five tissues; III) 68 transcripts (account for 14.17%) with the lowest expression level in Heart than in the other four tissues; IV) 88 transcripts (account for 18.33%) with a sharp rising trend and a relatively higher expression level in the other four tissues except for Musc. The tissue expression patterns of the 480 DEGs union in five tissues is shown in Fig. 7A. Expression levels of all these unigenes were highest in Hepan in clusters 1–4. Meanwhile, we also found that the expression levels of all these unigenes in PylS were relatively higher than that in CarS. Furthermore, the DEGs with a higher expression (fold change > 13, readcount > 10000 and FDR < 0.05) and a specific expression (readcount > 10000 and FDR < 0.05) in Hepan are emphatically shown in Table 5.

Subsequently, the twenty-seven DEGs were further extracted from the intersections of four comparison combinations, which suggests that these transcripts were potentially important in the innate immune system of *L. vannamei* (Fig. 7B). The hierarchical clustering indicated that 8 transcripts were differentially expressed in Hepan and Pyls, and the other 19 transcripts were only expressed in Hepan. Among these DEGs, some transcripts were involved in innate immune system pathways by KEGG category; for example, HSP70 and CTSL involved in antigen processing and presentation pathway, actinin alpha (ACTN) and Integrin involved in the leukocyte transendothelial migration pathway, and BIRC8 involved in the NOD-like receptor signaling pathway.

3.8. Verification of important DEGs related to immune-related molecules using real-time PCR

To validate the expression patterns in the five tissues from transcriptome data, six immune-related DEGs were selected for real-time PCR assay, including Hemy, HSP70, CTSL, C-type lectin (CTLC), integrin and BIRC8 genes. As a result, the trends of the real-time PCR results were significantly correlated with the Iso-Seq results (Fig. 8A). Furthermore, the transcriptional expression levels of the putative IRF 1–5 transcripts in five tissues were also analyzed. The real-time PCR assay provided expression patterns in the five tissues similar to that of the Iso-Seq levels (Fig. 8B). The expression of IRF1 and IRF2 in Hepan is significantly higher than that in Musc ($P < 0.05$), and the expression of IRF3 and IRF4 was shown at the highest levels in Hepan, followed by PylS and CarS, and weakly in Musc and Heart. However, the expression level of IRF5 was relatively higher in Hepan and PylS than that in other tissues. Consequently, the results confirmed the identical trend patterns of these DEGs and putative IRFs in both assays, suggesting accuracy of the transcriptome data by combined PacBio Iso-Seq sequencing and Illumina short-read sequencing.

4. Discussion

PacBio third generation sequencing has recently become available, which could capture full-length transcripts without assembly and overcomes the difficulty posed by the Illumina short-read data. This technology has been successfully applied in a few studies in plants, animals, even human beings, and provides further information of transcriptome, including alternative splicing, alternative polyadenylation, long noncoding RNAs, and the identification of novel genes. Here, the full-length transcript sets of *L. vannamei* generated by PacBio Iso-Seq provided an isoform-level reference transcriptome, enabling a comprehensive insight into the innate immune system of shrimp. In the present study, we first produced high confidence and full-length transcriptome data from five independent tissues (heart, muscle, hepatopancreas, cardiac stomach and pyloric stomach), which maximizes the transcript diversity using the PacBio Sequl sequencing approach. As expected, a large amount of transcriptome data was generated, including 72,648 unigenes with a mean length of 2,545 bp, compared with previous transcriptome reports yielding only N50 less than 1000 bp [14,35]. A total of 62,164 (85.57%) out of 72,648 unigenes were successfully annotated as known homologous genes using seven well-curated databases (Table 2) and 7544 unigenes were identified as putative lncRNAs (Fig. 3). This implies that the remaining 10,484 unigenes generated in the present study were not annotated according to the existing databases. There could be several reasons for this finding, for example the absence of reference genomic information and the unigenes without hits probably belonged to untranslated regions. It is also possible that these unigenes were putative novel genes for *L.*

Table 5
The differentially expressed genes (DEGs) in immune-related unigenes with significantly higher expression and tissue-specific expression in the hepatopancreas.

Gene ID	Gene Symbol (annotation)	Mrd. Hepenc	Mrd. PvlS	Mrd CarS	Mrd. Heart	Mrd. Musc	Log2FC	FDR
Higher expression with fold change > 13, readcount > 10000 and FDR < 0.05.								
i1_LQ_DXisoseq_c256578/f1p0/1751	chitinase	26487.24	0.23				16.81	0.007
i2_LQ_DXisoseq_c62242/f1p17/2145	hemocyanin	13418.64	0.46				14.83	0.000
i1_LQ_DXisoseq_c49958/f1p0/1446	beta-actin, ACTB_G1	32631.14	1.94				14.04	0.006
i2_LQ_DXisoseq_c161937/f1p0/2054	hemocyanin	20982.24	1.84				13.48	0.004
i2_LQ_DXisoseq_c103035/f1p0/2063	heat shock cognate 70, HSPA1_8	35100.53		0.62			15.80	0.038
i2_LQ_DXisoseq_c28692/f1p3/2160	receptor-type tyrosine-protein phosphatase, CD148	13698.69		0.54			14.62	0.028
i1_HQ_DXisoseq_c1923/f42p0/1692	retinoid-inducible serine carboxypeptidase-like, SCPEP1	20339.06		1.97			13.33	0.036
i2_LQ_DXisoseq_c4618/f1p19/2127	hemocyanin	14463.67		1.54			13.20	0.018
i6_HQ_DXisoseq_c695/f4p0/6528	cyclin B	51199.35			0.42		16.90	0.021
i3_HQ_DXisoseq_c17081/f2p0/3373	thrombospondin II, THBS2S	32484.93			0.33		16.58	0.004
i2_LQ_DXisoseq_c149845/f1p1/2828	thrombospondin protein, THBS2S	15560.13			0.33		15.52	0.000
i0_LQ_DXisoseq_c22035/f1p0/616	C-type lectin	19503.85			0.42		15.50	0.003
i3_LQ_DXisoseq_c23485/f1p1/3153	von Willebrand factor A domain-containing protein 7, PTPRS	35210.05			1.00		15.11	0.011
i0_HQ_DXisoseq_c39989/f35p0/544	C-type lectin-like, CLEC17A	51294.07			2.43		14.36	0.048
i2_LQ_DXisoseq_c144456/f1p0/2242	hemocyanin	35682.64			1.84		14.25	0.013
i1_LQ_DXisoseq_c19530/f12p0/1610	legumain, LGMN	40835.40			2.52		13.98	0.024
i1_HQ_DXisoseq_c20395/f2p0/1237	actin beta/gamma 1, ACTB_G1	34092.01			2.66		13.65	0.023
i0_HQ_DXisoseq_c39514/f15p0/635	C-type lectin	10184.10			0.84		13.57	0.011
i0_LQ_DXisoseq_c17962/f20p0/819	zinc proteinase Mpc1	25074.12			2.34		13.39	0.008
i1_LQ_DXisoseq_c25485/f1p22/1606	cathepsin C, CTSC	22062.71			8.47		11.35	0.011
i1_LQ_DXisoseq_c329287/f1p0/1003	chitinase	33862.34			3.36		13.30	0.041
i1_LQ_DXisoseq_c329724/f1p0/1029	cathepsin L, CTSL	33438.19			4.02		13.02	0.018
i2_HQ_DXisoseq_c196573/f2p1/2116	hemocyanin	184236.22				1.26	17.16	0.008
i1_LQ_DXisoseq_c94694/f1p0/1131	cathepsin L, CTSL	48233.52				2.86	14.04	0.009
i1_LQ_DXisoseq_c41725/f1p1/1554	legumain, LGMN	75161.55				6.19	13.57	0.008
i1_HQ_DXisoseq_c16898/f17p0/1251	cathepsin D-like protein, CTSD	63494.26				5.79	13.42	0.012
i1_LQ_DXisoseq_c43871/f5p0/1292	cathepsin L, CTSL	54036.72				5.05	13.39	0.005
i0_LQ_DXisoseq_c4878/f1p0/559	destabilase I	94120.29				8.86	13.38	0.015
i0_LQ_DXisoseq_c17950/f19p2/545	C-type lectin	10327.21				1.06	13.25	0.026
i0_LQ_DXisoseq_c326/f16p0/824	C-type lectin 4, CLEC17A	21011.55				2.42	13.09	0.012
i1_LQ_DXisoseq_c34869/f1p0/1055	cathepsin L, CTSL	18892.94				2.23	13.05	0.012
Specific expression only in hepatopancreas with readcount > 10000 and FDR < 0.05.								
i2_LQ_DXisoseq_c135312/f1p23/2145	hemocyanin	48727.76	0				Inf	0.039
i1_LQ_DXisoseq_c53015/f1p0/1689	hemocyanin	16550.11	0				Inf	0.000
i2_LQ_DXisoseq_c114265/f1p0/2108	hemocyanin	63857.13		0			Inf	0.049
i1_LQ_DXisoseq_c25485/f1p22/1606	cathepsin C, CTSC	24679.17		0			Inf	0.013
i1_LQ_DXisoseq_c49958/f1p0/1446	beta-actin, ACTB_G1	24338.62		0			Inf	0.013
i2_LQ_DXisoseq_c161937/f1p0/2054	hemocyanin	15711.96		0			Inf	0.009
i2_LQ_DXisoseq_c74701/f1p0/2033	hemocyanin	54957.21			0		Inf	0.022
i1_LQ_DXisoseq_c201946/f1p3/1785	hemocyanin	52568.77			0		Inf	0.022
i2_LQ_DXisoseq_c5879/f1p654/2951	hemocyanin	34441.98			0		Inf	0.018
i3_LQ_DXisoseq_c19616/f1p0/3619	thrombospondin, THBS2S	24760.32			0		Inf	0.001
i3_LQ_DXisoseq_c8884/f1p0/3829	thrombospondin II, THBS2S	23113.47			0		Inf	0.001
i1_LQ_DXisoseq_c256578/f1p0/1751	chitinase	17814.71			0		Inf	0.004
i5_LQ_DXisoseq_c10654/f1p0/5217	thrombospondin	16221.61			0		Inf	0.000
i1_HQ_DXisoseq_c11168/f8p0/1448	cathepsin D-like, CTSD	11362.65			0		Inf	0.000
i1_LQ_DXisoseq_c103364/f1p0/1310	hemocyanin	61252.80				0	Inf	0.023
i1_LQ_DXisoseq_c19530/f12p0/1610	legumain, LGMN	27095.46				0	Inf	0.001
i2_LQ_DXisoseq_c144456/f1p0/2242	hemocyanin	23427.52				0	Inf	0.000
i1_LQ_DXisoseq_c33882/f1p0/1120	actin, cytoplasmic 2 isoform X3, ACTB_G1	13255.89				0	Inf	0.001
i1_LQ_DXisoseq_c28363/f4p0/1636	cathepsin C, CTSC	12151.26				0	Inf	0.003

vannamei.

Due to the lack of predictable adaptive immunity, shrimp rely on an innate immune system to defend themselves against invading microbes by recognizing and clearing them through humoral and cellular immune responses. For the L. *vannamei* innate immune system, much of the interest has been focused on the identification of the immune molecules, providing the genetic materials for deep insight into the immune mechanism in shrimp. For this purpose, the plenty of studies using homologous cloning method have been reported [8,36] and several transcriptome studies also recently have been reported [13,37]; however, these studies were limited by either number or length of the generated sequence information, necessitating further cloning efforts to obtain full-length cDNA sequences for the investigation of potential roles in the innate immune system. In the present study, full-length sequences for a significantly high proportion of the immune genes involved in innate immune pathway were determined by combining PacBio long reads sequencing and second-generation short reads sequencing. A total of 5279 unigenes with full-length transcripts involved in the innate immunity processes were collected by combining efficient function annotation and literature progresses (Supplementary Table 7). In our immunity related unigenes, the abundance of sequence variations generates different isoform clusters and increases the number of unigenes, which possibly resulted from the library construction by pooled RNAs of five different tissues and three sample individuals. This may have resulted in the lower error rate in the PacBio Iso-seq, for example mismatches, insertions and deletions [38]. The diversity and variation of the full-length transcripts increased the complexity of the biological processes, which have been widely reported in previous

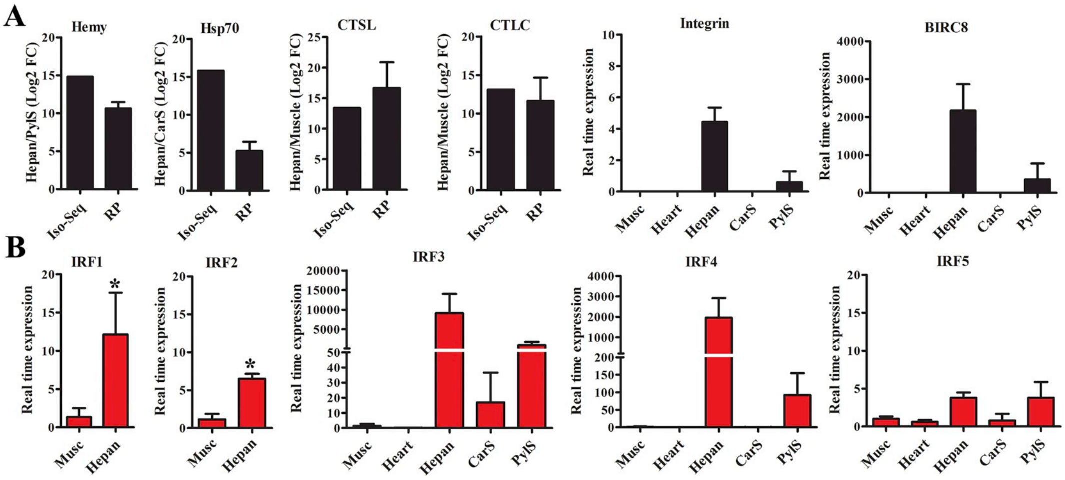

Fig. 8. Validation of differentially expressed genes (DEGs) and interferon regulatory factors (IRFs) by real-time PCR. A) The relative gene expression levels of six selected DEGs. B) The relative expression levels of IRF 1–5. RP: Real-time PCR; Hepan: hepatopancreas; Musc: muscle; Heart: heart; CarS: cardiac stomach and PylS: pyloric stomach. Bars with asterisk symbol indicate statistical differences ($P < 0.05$).

studies involving *Salvia miltiorrhiza* [39], *Panax ginseng* [40], and *Coffea arabica* [41]. Therefore, our data will be an invaluable resource for identifying candidate immunity molecules, and for comprehensively understanding functional pathways involved in the innate immune pathways in shrimp.

Using short-read RNA-Seq strategies, extensive alternative prediction is impractical and a high variability of isoforms expression quantification is impossible in shrimp without a true genome reference. However, the PacBio Iso-Seq strategy provides the convenience of finding more numbers of AS events of genes in many species, including reference-free species [42]. In our study, we identified 1164 non-redundant full-length transcripts (449 UniTransModels) participating in AS events by long read reference genome-free reconstruction of full-length transcriptome data in *L. vannamei*. We also revealed that there were different AS events in different tissues, which probably correspond to tissue-specific functions. Further studies are required to investigate the biological functions of these tissues-specific patterns.

We compared our results with published studies with viral and bacterial infection in shrimp. The previous study indicated that the largest number of unigenes is assigned to the pathway of signal transduction under *Vibrio parahaemolyticus* infection in *Exopalaemon carinicauda* [43]. Meanwhile, the abundant distribution of unigenes is predicted to human disease related pathways of KEGG classification from the pooled database of the hepatopancreas transcriptome (control group and infection group) in *M. japonicas* [44]; even in plant, a few full-length unigenes are annotated in human disease category of KEGG classification [42]. Above analysis may suggest that these human disease related genes are in advance stored in normal immune tissues and/or organs of shrimp. More importantly, many immune-related unigenes revealed by the present study were found to be differentially expressed in the hepatopancreas under experimental viral and bacterial infection, such as C-type lectin, hemocyanin, cathepsin C, chitinase, integrin, Toll-like receptors and so on. Therefore, the immune-related full-length unigenes will contribute to a deep insight into the innate immune system in shrimp.

The comparative analysis combined the PacBio Iso-seq transcripts with Illumina short-read data, which allowed us to evaluate the differential genes expression analysis. The hepatopancreas is the main tissues or organ involving in humoral immunity and cellular immunity in shrimp, and plays an important role in the immune system [3,44,45]. The epithelial cells of the hepatopancreas are major resources of immune molecules, such as lectins, hemocyanin, ferritin, antibacterial and antiviral proteins, proteolytic enzymes and nitric oxide [46]. We identified 480 DEGs with high and specific expression pattern in the hepatopancreas compared with that in the other four tested tissues of *L. vannamei* from 5279 immune related transcripts (Fig. 7A and Supplementary Table 8). Furthermore, we emphasized 31 highly expressed unigenes with fold change > 13, readcount > 10000 and 19 specific-expression unigenes only in the hepatopancreas with readcount > 10000 (Table 5), which were all well annotated by known databases according to sequence homologous alignment. These DEGs were involved in an immune-related pathway of KEGG classifications, such as endocytosis (ko04144), apoptosis (ko04210), lysosome (ko04142), focal adhesion (ko04510), phagosome (ko04145), PI3K-Akt signaling pathway (ko04151),sphingolipid signaling pathway (ko04071), hippo signaling pathway (ko04390), ECM-receptor interaction (ko04512), Rap1 signaling pathway (ko04015), platelet activation (ko04611), leukocyte transendothelial migration (ko04670), and antigen processing and presentation (ko04612). Among these DEGs, the tissue preferred expression pattern of some genes in the hepatopancreas has been also reported in the previous studies. For example, lectins were isolated and mainly distributed in the hepatopancreas in most shrimp [47,48]. A novel L-type lectin had an especially high expression in the hepatopancreas of *L. vannamei* and *M. rosenbergii*, respectively [10,49]. Meanwhile, the function of some genes involved in the immune system was widely investigated, such as hemocyanin [50] and heat shock cognate 70 (HSP70) [51].

The express pattern of some immune-related unigenes and different transcript variants were selected for real-time PCR, which is regarded as gold criteria of gene expression with high sensitivity and accuracy. Hemocyanin played multiple roles in innate immune defense, immune modulation, hematopoiesis, signal transduction and microbicidal activities [52,53]. Previous study have indicated that Hemocyanin was identified as the highest abundant gene in hepatopancreas of vibrio parahaemolyticus AHPND challenged shrimp [54]. In the present study, various hemocyanin transcripts with full-length sequences had differentially higher and specific expression in the hepatopancreas (Fig. 7B and Table 5), which might suggest abundant functions in innate immunity in *L. vannamei*. C-type lectins are a family of calcium-dependent carbohydrate-binding proteins, which are highly conserved in vertebrates but show considerable diversity among invertebrates [55]. In crustaceans, the hepatopancreas is regarded as an important organ involved in innate immunity. High expression in the hepatopancreas of CTLC in *L. vannamei* was consistent with previous study in other shrimp species, such as *M. rosenbergii* [48]. In vertebrate, the interferon (IFN) response is the hallmark of antiviral immunity, which is characterized by the induction of IFNs and the subsequent establishment of the cellular antiviral state. Importantly, interferon regulatory factor (IRF) family is a group of transcriptional factors with a highly conserved DNA-binding domain in the amino-terminal regions, which play critical roles in the activation of the IFN. The first invertebrate IRF from *L. vannamei* with a length of 1416 bp was identified in a previous study which indicated that an IRF-like gene shared the similar antiviral mechanism via the IRF-Vago-JAK/STAT regulatory axis in invertebrates with that of the IRF−IFN−JAK/STAT axis of vertebrates [34]. We also found that five additional different transcripts were annotated as the same IRF gene of *L. vannamei* [34]. Among these, the expression levels of IRF1, IRF3 and IRF4 with significantly higher expression in the hepatopancreas were validated by Iso-Seq and real-time PCR; however, that of IRF2 and IRF5 were relatively low (Fig. 8B), which may suggest the different roles in the IFN responses need to be further investigated.

In summary, we combined PacBio Iso-seq with Illumina short-read sequencing methods to conduct a comprehensive transcriptome analysis in *L. vannamei*. This enabled the generation of full-length transcripts as well as related analysis, that is, efficient gene annotation, alternative splicing, and long noncoding RNAs. More importantly, transcript variants and expression profiles survey of the immune related molecules of *L. vannamei* contributed to a comprehensive insight into the immune system. Therefore, our study provide a valuable resource of a comprehensive full-length transcripts set for genomic reference, which is interesting and worthy of further in-depth studies in shrimp.

Acknowledge

This work was supported by the Funds of Technology Transfer from Guangdong Academy of Sciences and Zhongshan [2016G1FC0008], the Guangzhou Science Technology and Innovation Commission [201804020080], the National Natural Science Fund of China [31802279], the GDAS Special Project of Science and Technology Development [2017GDASCX-0107 and 2018GDASCX-0107], and the Planning funds of Scientific and Technological of Guangdong Province [2016B070701016].

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fsi.2019.01.023.

References

[1] A. Tassanakajon, K. Somboonwiwat, P. Supungul, S. Tang, Discovery of immune molecules and their crucial functions in shrimp immunity, Fish Shellfish Immunol. 34 (2013) 954–967.

[2] C.Y. Xie, J.R. Kong, C.S. Zhao, Y.C. Xiao, T. Peng, Y. Liu, W.N. Wang, Molecular characterization and function of a PTEN gene from *Litopenaeus vannamei* after Vibrio alginolyticus challenge, Dev. Comp. Immunol. 59 (2016) 77–88.

[3] P. Jiravanichpaisal, B.L. Lee, K. Soderhall, Cell-mediated immunity in arthropods: hematopoiesis, coagulation, melanization and opsonization, Immunobiology 211 (2006) 213–236.

[4] A.F. Rowley, A. Powell, Invertebrate immune systems specific, quasi-specific, or nonspecific? J. Immunol. 179 (2007) 7209–7214.

[5] E. Bachere, Y. Gueguen, M. Gonzalez, J. de Lorgeril, J. Garnier, B. Romestand, Insights into the anti-microbial defense of marine invertebrates: the penaeid shrimps and the oyster Crassostrea gigas, Immunol. Rev. 198 (2004) 149–168.

[6] X. Zhang, Y. Zhang, C. Scheuring, H.B. Zhang, P. Huan, B. Wang, C. Liu, F. Li, B. Liu, J. Xiang, Construction and characterization of a bacterial artificial chromosome (BAC) library of Pacific white shrimp, *Litopenaeus vannamei*, Mar. Biotechnol. 12 (2010) 141–149.

[7] Y. Yu, X. Zhang, J. Yuan, F. Li, X. Chen, Y. Zhao, L. Huang, H. Zheng, J. Xiang, Genome survey and high-density genetic map construction provide genomic and genetic resources for the Pacific White Shrimp *Litopenaeus vannamei*, Sci. Rep. 5 (2015) 15612.

[8] L. Shi, S. Chan, C. Li, S. Zhang, Identification and characterization of a laccase from Litopenaeus vannamei involved in anti-bacterial host defense, Fish Shellfish Immunol. 66 (2017) 1–10.

[9] P.H. Wang, J.P. Liang, Z.H. Gu, D.H. Wan, S.P. Weng, X.Q. Yu, J.G. He, Molecular cloning, characterization and expression analysis of two novel Tolls (LvToll2 and LvToll3) and three putative Spätzle -like Toll ligands (LvSpz1-3) from *Litopenaeus vannamei*, Dev. Comp. Immunol. 36 (2012) 359–371.

[10] Y. Tian, T. Chen, W. Huang, P. Luo, D. Huo, L. Yun, C. Hu, Y. Cai, A new L-type lectin (LvLTLC1) from the shrimp *Litopenaeus vannamei* facilitates the clearance of Vibrio harveyi, Fish Shellfish Immunol. 73 (2018) 185–191.

[11] X.D. Huang, Z.X. Yin, J.X. Liao, P.H. Wang, L.S. Yang, H.S. Ai, Z.H. Gu, X.T. Jia, S.P. Weng, X.Q. Yu, J.G. He, Identification and functional study of a shrimp Relish homologue, Fish Shellfish Immunol. 27 (2009) 230–238.

[12] Q. Liang, J. Zheng, H. Zuo, C. Li, S. Niu, L. Yang, M. Yan, S.P. Weng, J. He, X. Xu, Identification and characterization of an interleukin-16-like gene from pacific white shrimp *Litopenaeus vannamei*, Dev. Comp. Immunol. 74 (2017) 49–59.

[13] N. Ghaffari, A. Sanchez-Flores, R. Doan, K.D. Garcia-Orozco, P.L. Chen, A. Ochoa-Leyva, A.A. Lopez-Zavala, J.S. Carrasco, C. Hong, L.G. Brieba, E. Rudino-Pinera, P.D. Blood, J.E. Sawyer, C.D. Johnson, S.V. Dindot, R.R. Sotelo-Mundo, M.F. Criscitiello, Novel transcriptome assembly and improved annotation of the whiteleg shrimp (*Litopenaeus vannamei*), a dominant crustacean in global seafood mariculture, Sci. Rep. 4 (2014) 7081.

[14] H. Guo, C.X. Ye, A.L. Wang, J.A. Xian, S.A. Liao, Y.T. Miao, S.P. Zhang, Trascriptome analysis of the Pacific white shrimp *Litopenaeus vannamei* exposed to nitrite by RNA-seq, Fish Shellfish Immunol. 35 (2013) 2008–2016.

[15] D. Zeng, X. Chen, D. Xie, Y. Zhao, C. Yang, Y. Li, N. Ma, M. Peng, Q. Yang, Z. Liao, H. Wang, X. Chen, Transcriptome analysis of Pacific white shrimp (*Litopenaeus vannamei*) hepatopancreas in response to Taura syndrome Virus (TSV) experimental infection, PLoS One 8 (2013) e57515.

[16] K.F. Au, V. Sebastiano, P.T. Afshar, J.D. Durruthy, L. Lee, B.A. Williams, H. van Bakel, E.E. Schadt, R.A. Reijo-Pera, J.G. Underwood, W.H. Wong, Characterization of the human ESC transcriptome by hybrid sequencing, Proc. Natl. Acad. Sci. U. S. A. 110 (2013) E4821–E4830.

[17] S.Y. Chen, F. Deng, X. Jia, C. Li, S.J. Lai, A transcriptome atlas of rabbit revealed by PacBio single-molecule long-read sequencing, Sci. Rep. 7 (2017) 7648.

[18] S.J. Zhang, C. Wang, S. Yan, A. Fu, X. Luan, Y. Li, Q. Sunny Shen, X. Zhong, J.Y. Chen, X. Wang, B. Chin-Ming Tan, A. He, C.Y. Li, Isoform evolution in primates through independent combination of alternative RNA processing events, Mol. Biol. Evol. 34 (2017) 2453–2468.

[19] R.K. Patel, M. Jain, NGS QC Toolkit: a toolkit for quality control of next generation sequencing data, PLoS One 7 (2012) e30619.

[20] I. Naguibneva, M. Ameyar-Zazoua, A. Polesskaya, S. Ait-Si-Ali, R. Groisman, M. Souidi, S. Cuvellier, A. Harel-Bellan, The microRNA miR-181 targets the homeobox protein Hox-A11 during mammalian myoblast differentiation, Nat. Cell Biol. 8 (2006) 278–284.

[21] L. Salmela, E. Rivals, LoRDEC: accurate and efficient long read error correction, Bioinformatics 30 (2014) 3506–3514.

[22] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT, Accelerated for clustering the next-generation sequencing data, Bioinformatics 28 (2012) 3150–3152.

[23] C.P. Hong, S.J. Lee, J.Y. Park, P. Plaha, Y.S. Park, Y.K. Lee, J.E. Choi, K.Y. Kim, J.H. Lee, J. Lee, H. Jin, S.R. Choi, Y.P. Lim, Construction of a BAC library of Korean ginseng and initial analysis of BAC-end sequences, Molecular genetics and genomics, MGG (Mol. Gen. Genet.) 271 (2004) 709–716.

[24] J. Ye, L. Fang, H. Zheng, Y. Zhang, J. Chen, Z. Zhang, J. Wang, S. Li, R. Li, L. Bolund, J. Wang, WEGO: a web tool for plotting GO annotations, Nucleic Acids Res. 34 (2006) W293–W297.

[25] A. Li, J. Zhang, Z. Zhou, PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme, BMC Bioinf. 15 (2014) 311.

[26] L. Sun, H. Luo, D. Bu, G. Zhao, K. Yu, C. Zhang, Y. Liu, R. Chen, Y. Zhao, Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts, Nucleic Acids Res. 41 (2013) e166.

[27] L. Kong, Y. Zhang, Z.Q. Ye, X.Q. Liu, S.Q. Zhao, L. Wei, G. Gao, CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine, Nucleic Acids Res. 35 (2007) W345–W349.

[28] R.D. Finn, P. Coggill, R.Y. Eberhardt, S.R. Eddy, The Pfam Protein Families Database: towards a More Sustainable Future vol. 44, (2016), pp. D279–D285.

[29] T.D. Wu, J. Reeder, M. Lawrence, G. Becker, M.J. Brauer, GMAP and GSNAP for genomic sequence alignment: enhancements to speed, accuracy, and functionality, Methods Mol. Biol. 1418 (2016) 283–334.

[30] G. Ning, X. Cheng, P. Luo, F. Liang, Z. Wang, G. Yu, X. Li, D. Wang, M. Bao, Hybrid sequencing and map finding (HySeMaFi): optional strategies for extensively deciphering gene splicing and expression in organisms without reference genome, Sci. Rep. 7 (2017) 43793.

[31] G.P. Alamancos, A. Pages, J.L. Trincado, N. Bellora, E. Eyras, Leveraging transcript quantification for fast computation of alternative splicing profiles, RNA 21 (2015) 1521–1531.

[32] B. Li, C.N. Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, BMC Bioinf. 12 (2011) 323.

[33] K.J. Livak, T.D. Schmittgen, Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ Method, Methods 25 (2001) 402–408.

[34] C. Li, H. Li, Y. Chen, Y. Chen, S. Wang, S.P. Weng, X. Xu, J. He, Activation of Vago by interferon regulatory factor (IRF) suggests an interferon system-like antiviral mechanism in shrimp, Sci. Rep. 5 (2015) 15078.

[35] C. Li, S. Weng, Y. Chen, X. Yu, L. Lu, H. Zhang, J. He, X. Xu, Analysis of *Litopenaeus vannamei* transcriptome using the next-generation DNA sequencing technique, PLoS One 7 (2012) e47442.

[36] J. Feng, L. Zhao, M. Jin, T. Li, L. Wu, Y. Chen, Q. Ren, Toll receptor response to white spot syndrome virus challenge in giant freshwater prawns (*Macrobrachium rosenbergii*), Fish Shellfish Immunol. 57 (2016) 148–159.

[37] Z. Qin, V.S. Babu, Q. Wan, M. Zhou, R. Liang, A. Muhammad, L. Zhao, J. Li, J. Lan, L. Lin, Transcriptome analysis of Pacific white shrimp (*Litopenaeus vannamei*) challenged by Vibrio parahaemolyticus reveals unique immune-related genes, Fish Shellfish Immunol. 77 (2018) 164–174.

[38] N.V. Hoang, A. Furtado, P.J. Mason, A. Marquardt, L. Kasirajan, P.P. Thirugnanasambandam, F.C. Botha, R.J. Henry, A survey of the complex

transcriptome from the highly polyploid sugarcane genome using full-length isoform sequencing and de novo assembly from short read sequencing, BMC Genomics 18 (2017) 395.

[39] Z. Xu, R.J. Peters, J. Weirather, H. Luo, B. Liao, X. Zhang, Y. Zhu, A. Ji, B. Zhang, S. Hu, K.F. Au, J. Song, S. Chen, Full-length transcriptome sequences and splice variants obtained by a combination of sequencing platforms applied to different root tissues of *Salvia miltiorrhiza* and tanshinone biosynthesis, Plant J. : Cell Mol. Biol. 82 (2015) 951–961.

[40] I.H. Jo, J. Lee, C.E. Hong, D.J. Lee, W. Bae, S.G. Park, Y.J. Ahn, Y.C. Kim, J.U. Kim, J.W. Lee, D.Y. Hyun, S.K. Rhee, C.P. Hong, Isoform sequencing provides a more comprehensive view of the *Panax ginseng* transcriptome, Genes 8 (9) (2017) pii: E228.

[41] B. Cheng, A. Furtado, R.J. Henry, Long-read sequencing of the coffee bean transcriptome reveals the diversity of full-length transcripts, GigaScience 6 (2017) 1–13.

[42] J. Li, Y. Harata-Lee, M.D. Denton, Long read reference genome-free reconstruction of a full-length transcriptome from *Astragalus membranaceus* reveals transcript variants involved in bioactive compound biosynthesis, Cell Discov. 3 (2017) 17031.

[43] Q. Ge, J. Li, J. Wang, J. Li, H. Ge, Q. Zhai, Transcriptome analysis of the hepatopancreas in Exopalaemon carinicauda infected with an AHPND-causing strain of Vibrio parahaemolyticus, Fish Shellfish Immunol. 67 (2017) 620–633.

[44] S. Zhong, Y. Mao, J. Wang, M. Liu, M. Zhang, Y. Su, Transcriptome analysis of Kuruma shrimp (*Marsupenaeus japonicus*) hepatopancreas in response to white spot syndrome virus (WSSV) under experimental infection, Fish Shellfish Immunol. 70 (2017) 710–719.

[45] H. Yang, X. Gao, X. Li, H. Zhang, N. Chen, Y. Zhang, X. Liu, X. Zhang, Comparative transcriptome analysis of red swamp crayfish (*Procambarus clarkia*) hepatopancreas in response to WSSV and Aeromonas hydrophila infection, Fish Shellfish Immunol. 83 (2018) 397–405.

[46] T. Roszer, The invertebrate midintestinal gland ("hepatopancreas") is an evolutionary forerunner in the integration of immunity and metabolism, Cell Tissue Res. 358 (2014) 685–695.

[47] P.S. Gross, T.C. Bartlett, C.L. Browdy, R.W. Chapman, G.W. Warr, Immune gene discovery by expressed sequence tag analysis of hemocytes and hepatopancreas in the Pacific White Shrimp, *Litopenaeus vannamei*, and the Atlantic White Shrimp, L. setiferus, Dev. Comp. Immunol. 25 (2001) 565–577.

[48] H. Zhu, J. Du, K.M. Hui, P. Liu, J. Chen, Y. Xiu, W. Yao, T. Wu, Q. Meng, W. Gu, Q. Ren, W. Wang, Diversity of lectins in *Macrobrachium rosenbergii* and their expression patterns under spiroplasma MR-1008 stimulation, Fish Shellfish Immunol. 35 (2013) 300–309.

[49] X. Huang, K. Han, T. Li, W. Wang, Q. Ren, Novel L-type lectin from fresh water prawn, Macrobrachium rosenbergii participates in antibacterial and antiviral immune responses, Fish Shellfish Immunol. 77 (2018) 304–311.

[50] P. Yang, D. Yao, J.J. Aweya, F. Wang, P. Ning, S. Li, H. Ma, Y. Zhang, c-Jun regulates the promoter of small subunit hemocyanin gene of *Litopenaeus vannamei*, Fish Shellfish Immunol. 84 (2018) 639–647.

[51] W. Junprung, P. Supungul, A. Tassanakajon, *Litopenaeus vannamei* heat shock protein 70 (LvHSP70) enhances resistance to a strain of Vibrio parahaemolyticus, which can cause acute hepatopancreatic necrosis disease (AHPND), by activating shrimp immunity, Dev. Comp. Immunol. 90 (2019) 138–146.

[52] C.J. Coates, J. Nairn, Diverse immune functions of hemocyanins, Dev. Comp. Immunol. 45 (2014) 43–55.

[53] C.J. Coates, H. Decker, Immunological properties of oxygen-transport proteins: hemoglobin, hemocyanin and hemerythrin, Cell. Mol. Life Sci. : CMLS 74 (2017) 293–317.

[54] P. Boonchuen, P. Jaree, A. Tassanakajon, K. Somboonwiwat, Hemocyanin of *Litopenaeus vannamei* agglutinates Vibrio parahaemolyticus AHPND (VPAHPND) and neutralizes its toxin, Dev. Comp. Immunol. 84 (2018) 371–381.

[55] M.J. Robinson, D. Sancho, E.C. Slack, S. LeibundGut-Landmann, C. Reis e Sousa, Myeloid C-type lectins in innate immunity, Nat. Immunol. 7 (2006) 1258–1265.