BMC Genomics

RESEARCH Open Access

PacBio single-molecule long-read sequencing provides new insights into the complexity of full-length transcripts in oriental river prawn, *macrobrachium nipponense*

Cheng-Yan Mou[1], Qiang Li[1], Zhi-Peng Huang[1], Hong-Yu Ke[1], Han Zhao[1], Zhong-Meng Zhao[1], Yuan-Liang Duan[1], Hua-Dong Li[1], Yu Xiao[2], Zhou-Ming Qian[3], Jun Du[1], Jian Zhou[1*] and Lu Zhang[1*]

Abstract

Background Oriental river prawn (*Macrobrachium nipponense*) is one of the most dominant species in shrimp farming in China, which is a rich source of protein and contributes to a significant impact on the quality of human life. Thus, more complete and accurate annotation of gene models are important for the breeding research of oriental river prawn.

Results A full-length transcriptome of oriental river prawn muscle was obtained using the PacBio Sequel platform. Then, 37.99 Gb of subreads were sequenced, including 584,498 circular consensus sequences, among which 512,216 were full length non-chimeric sequences. After Illumina-based correction of long PacBio reads, 6,599 error-corrected isoforms were identified. Transcriptome structural analysis revealed 2,263 and 2,555 alternative splicing (AS) events and alternative polyadenylation (APA) sites, respectively. In total, 620 novel genes (NGs), 197 putative transcription factors (TFs), and 291 novel long non-coding RNAs (lncRNAs) were identified.

Conclusions In summary, this study offers novel insights into the transcriptome complexity and diversity of this prawn species, and provides valuable information for understanding the genomic structure and improving the draft genome annotation of oriental river prawn.

Keywords Long non-coding RNA, Novel genes, Alternative splicing, Alternative polyadenylation, SMRT sequencing, Oriental river prawn

*Correspondence:
Jian Zhou
zhoujian980@126.com
Lu Zhang
zhanglu425@163.com
[1]Fisheries Institute, Sichuan Academy of Agricultural Sciences, Chengdu, Sichuan 611731, China
[2]Sichuan Academy of Agricultural Sciences, Chengdu, Sichuan 610066, China
[3]Chengdu Eaters Agricultural Group Co., Ltd, Chengdu, Sichuan 610000, China

© The Author(s) 2023. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Introduction

Oriental river prawn (*Macrobrachium nipponense*), which belong to family Palaemonidae, order Decapoda, subphylum Crustacea, are commonly found in the low-salinity and freshwater regions of estuaries in China [1]. It has become one of the most dominant species in shrimp farming in China, with delicious taste and high nutritional value. The annual production capacity of oriental river prawn has increased to 228,765 tons in 2020 [2]. With the rapidly developing sequencing technology, genome and transcriptome information can serve as an effective tool in promoting the breeding process of oriental river prawn. In 2011, 15,806 bp (bp) mitochondrial genome from a single female oriental river prawn were sequenced, which is comprised of 37 genes, including 13 protein-coding genes (PCGs), 22 transfer RNAs (tRNAs) and 2 ribosomal RNAs (rRNAs) [1]. In 2013, transcriptome analysis of oriental river prawn androgenic gland showed that a total of 78,408 isosequences were obtained, among which 57,619 non-redundant transcripts and 40 candidate NGs were found [3]. The latest reference genome assembly (ASM1510439v1) of oriental river prawn was generated by using Illumina and PacBio sequencing, assembling ~4.5 Gb of the genome, with predictions of 44,086 protein-coding genes. This assembly has a higher sequence continuity and accuracy because of using two sequencing methods [4]. Other transcriptome analysis mainly focused on revealing differential gene expression analysis and dynamic spatial gene coexpression networks [5–8]. Although these sequences can serve as useful genetic resources for shrimp breeding, the genome coverage remains incomplete. To date, most of the gene models were predicted in silico, and the information on untranslated regions and alternative isoforms are still lacking [9–11]. Hence, more precise genomic information is essential to improve the functional and structural annotation of the existing oriental river prawn reference genome.

Since the introduction of large-scale sequencing platform, transcriptomic sequencing has received considerable attention in the research of gene expression and regulation [12]. Next-generation sequencing (NGS), including the Illumina platform, has been widely employed for transcriptome and genome analyses in many species because of its multiple advantages, such as accuracy and cost-effective [13–16]. However, the NGS in short amplified fragments makes the reconstruction task more complicated, and increase the difficulty of accurate full-length splice isoform prediction [17, 18]. Recently, the PacBio (PB) Single-Molecule Real-Time (SMRT) sequencing, as a representative of the third generation sequencing (TGS) technology, can directly obtain full-length splice isoforms without assembly, thereby overcoming the limitations of short-read sequences and allowing the identification of rare or novel splice variants [19–22]. At present, PB sequencing has been widely employed in different species, including pearl oyster (*Pinctada fucata martensii*) [23], Cattle (*Bos taurus*) [24], rabbit (*Oryctolagus cuniculus*) [25], Chinese chive maggots (*Bradysia odoriphaga*) [26] and sedges (*Carex breviculmis*) [27]. However, PB sequencing still possesses some disadvantages such as low throughput and high sequencing error rates [28, 29]. Therefore, a combined strategy of SMRT sequencing and Illumina RNA-seq data to complement each other has became increasingly important [30–32]. In this study, we gained a full-length transcriptome from oriental river prawn muscle through PacBio SMRT and Illumina sequencing, and identified NGs, structural variations, AS events, TFs and lncRNAs. These data will improve our understanding of the structural variations and complexity of the transcripts, and provide a strong basis for further genomic research on this prawn species.

Results

Baseline characteristics of the SMRT sequences of oriental river prawn

To further investigate the transcriptome complexity of oriental river prawn, its muscle tissues were collected to extract total RNA, and the SMRT library was constructed for sequencing by using the PB Sequel platform. About 39.47 Gb of raw data consisting 667,816 raw polymerase reads were obtained. In total, 15,481,437 subreads (37.99 Gb) were identified, with the average read length and N50 length of 2,454 and 2,393 bp, respectively. To retrieve more accurate sequencing data, 584,498 circular consensus sequences (CCSs) were identified from subreads that pass at least 2 times through the insert. Of them, 513,236 CCSs belonged to full-length reads, and 512,216 full-length non-chimeric (FLNC) reads with a mean read length of 2,701 bp were identified. Next, all FLNC reads were clustered to remove redundancy and corrected by Arrow software, which finally obtained 21,008 polished consensus sequences. The average length of polished consensus reads was 3,009 bp. To correct the high error rate of PB long reads, ~634.2 million clean reads were generated using the Illumina platform. Next, LoRDEC software was used to correct the PB long reads based on the Illumina short reads. Lastly, 21,008 corrected sequences were obtained, with the average read length and N50 length of 3,007 and 3,396 bp, respectively. The length distributions of all the above sequences are shown in Fig. 1; Table 1.

Genome mapping

All the corrected sequences were compared against the oriental river prawn reference genome (ASM1510439v1) via GMAP software. In total, 19,525 reads (92.94%)

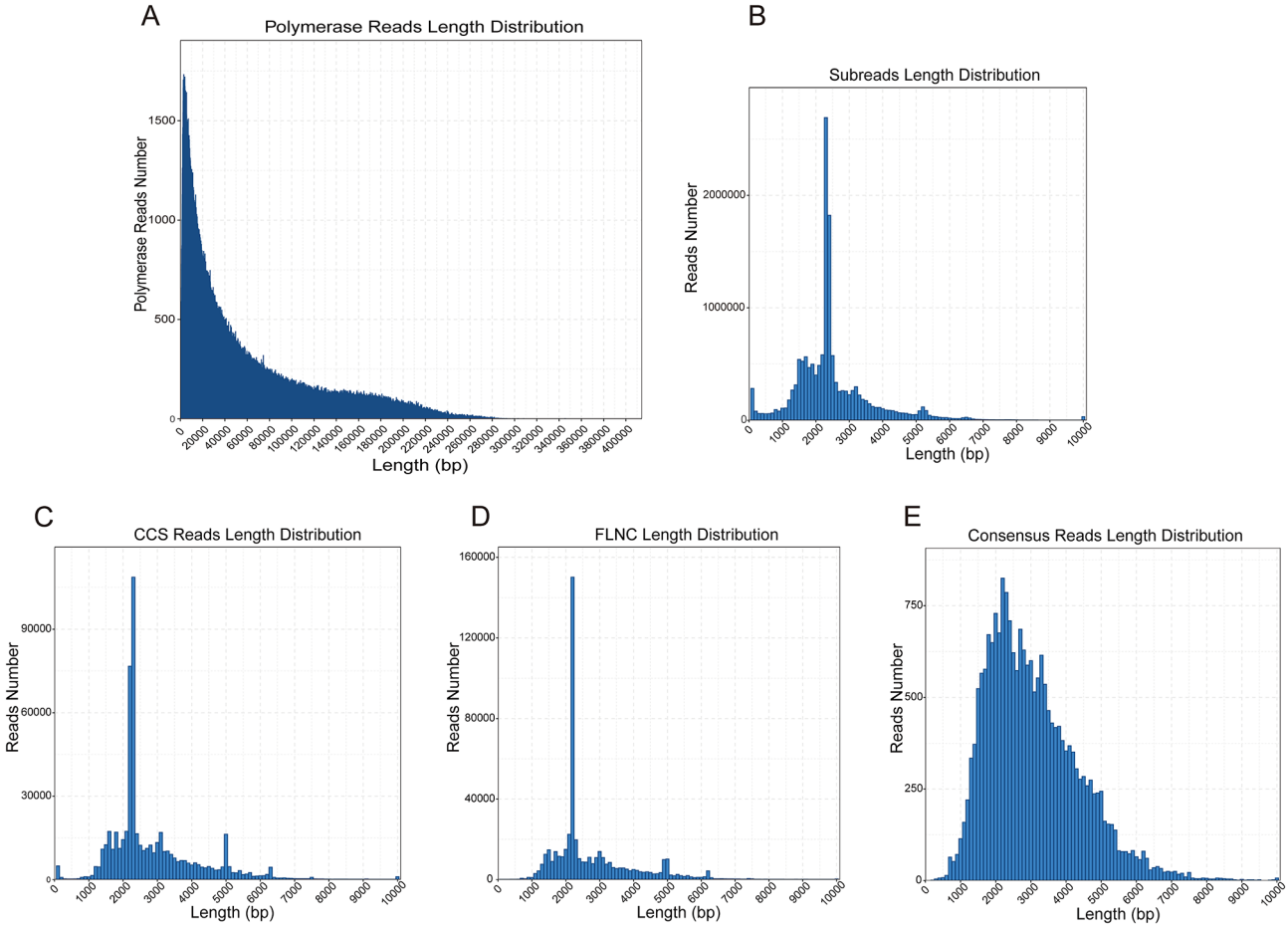

Fig. 1 Length distributions of SMRT sequences. (**A**) length distributions of 667,816 polymerase read. (**B**) length distributions of 15,481,437 subreads sequences. (**C**) length distributions of 584,498 CCS sequences. (**D**) length distributions of 512,216 FLNC sequences. (**E**) length distributions of 21,008 consensus sequences

Table 1 Summary of PB SMRT sequencing reads

	Polymerase reads	Subreads	CCS	FLNC	Corrected consensus
Number	667,816	15,481,437	484,498	512,216	21,008
Mean length (bp)	59,109	2,454	2,796	2,701	3,009
N50	117,297	2,393	2,942	2,776	3,396

were mapped to the reference genome. As shown in Fig. 2A, these reads were divided into 4 groups: mapped to plus (+), mapped to minus (−), multiple mapped and unmapped. These four groups comprised of 11,399 reads (54.26%) mapped to the positive strand, 8,051 reads (38.32%) mapped to the opposite strand, 75 reads (0.36%) with multiple alignments and 1,483 reads (7.06%) without any mapping to the reference genome, respectively. A saturation level was observed in the curve of the corrected isoform numbers (Fig. 2B), and 75% high-quality reads with identity and coverage values of >98% were identified (Fig. 2C).

After correction, the transcript sequences were mapped against the reference genome. The Genome Mapping and Alignment Program (GMAP) output file and genome annotation file (http://gigadb.org/dataset/100843) were employed for the analysis of transcript and gene isoforms. Reads that were mapped to different exons in known gene regions were defined as new isoforms, and isoforms that spanned more than one gene were excluded from downstream splice analysis. Subsequently, 6,599 isoforms were generated, which could be assigned to 3 groups: (i) 296 isoforms of known genes; (ii) 5,537 novel isoforms from known genes; and (iii) 766 isoforms from NGs (Fig. 2D). Additionally, 620 NGs (no annotation in reference genome) were also identified (Table S1).

Functional annotation of NGs

To enhance functional annotation, 620 NGs were annotated by NCBI-Nt, NCBI-Nr, Pfam, KOG, GO, KEGG and Swissprot databases. There were 35 genes overlapped across the 7 databases, and 365 genes were detected in

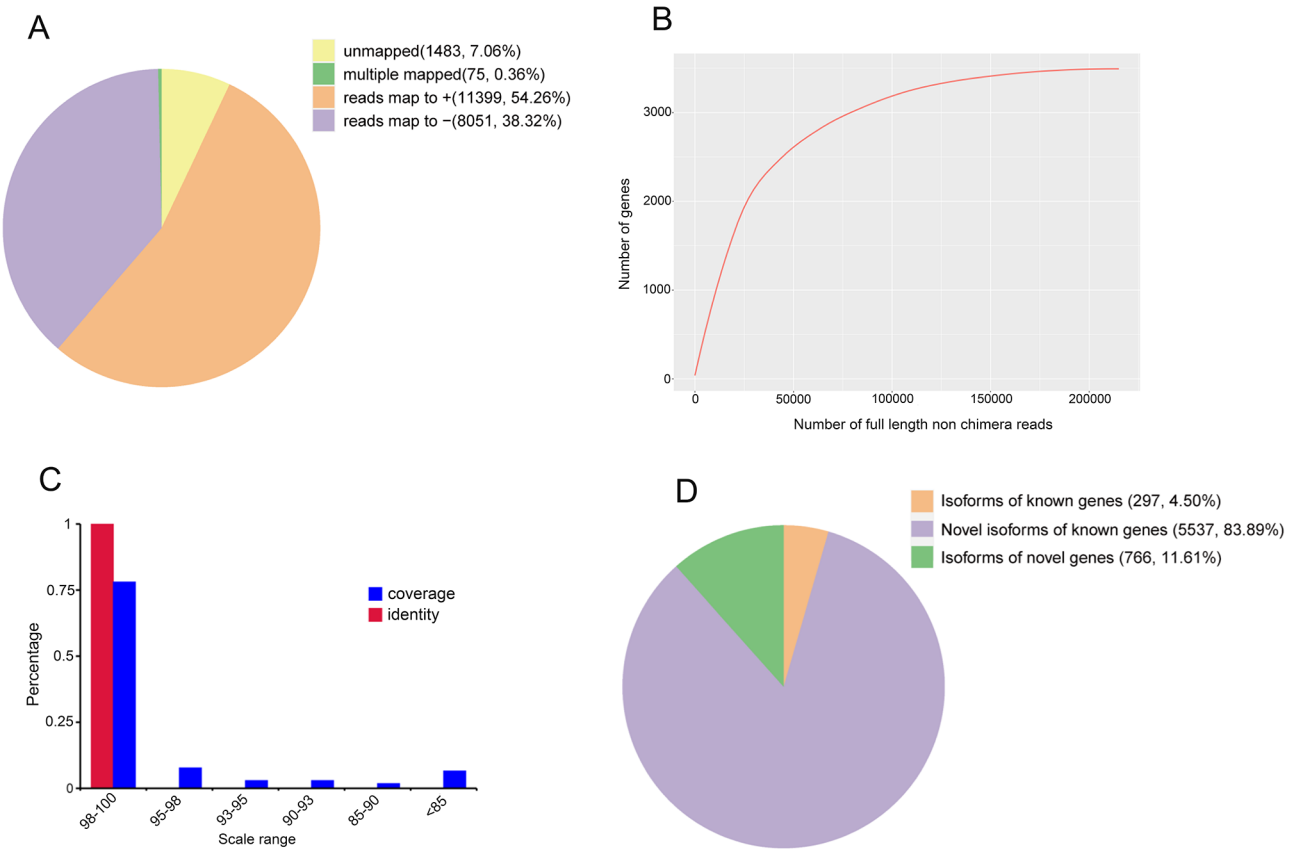

Fig. 2 GMAP analysis of SMRT sequences. (**A**) GMAP mapping of the corrected sequences. (**B**) Saturation curve of the corrected sequences. (**C**) Sequence identity and coverage. (**D**) Classification of the identified transcript isoforms

at least 1 database (Fig. 3A and Table S2). The NGs were compared against the Nr database to identify homologous genes. It was found that the top 5 NGs have homologues in *Hyalella azteca* (64), *Macrobrachium nipponense* (28), *Limulus polyphemus* (20), *Daphnia magna* (12), and *Pediculus humanus corporis* (10) (Fig. 3B). Moreover, GO analysis revealed that "cellular process", "metabolic process", and "single-organism process" were significantly enriched in the "biological process", "Cell", "Cell part", and "membrane" were significantly enriched in the "cellular components", and "binding" and "catalytic activity" were significantly enriched in the "molecular functions"(Fig. 3C). KOG analysis demonstrated that the NGs were clustered into 23 functional groups, and the "General function prediction only", "Signal transduction mechanisms", and "Energy production and conversion" ranked as the three most common categories (Fig. 3D). KEGG analysis indicated that the NGs were mapped to 110 KEGG pathways (Fig. 3E).

Determination of alternative splicing (AS) and alternative polyadenylation (APA)

In this study, SUPPA software was applied to determine AS events [33]. Seven types of AS events were identified, including alternative first exon (AF), alternative 3' splice site (A3), alternative 5' splice site (A5), alternative last exon (AL), mutually exclusive exon (MX), retained intron (RI) and skipped exon (SE). A total of 2,263 AS events were found from 3,605 gens (Table S3). Four kinds of events, SE (698), AF (441), A5 (385) and A3 (358) were relatively more common than other three AS events (Fig. 4A). PB sequencing also allows the determination of APA sites. A total of 2,555, poly(A) sites were detected in 886 genes, among which 80 and 540 genes had 5 and 2 poly(A) sites, respectively (Fig. 4B and Table S4). The mean number of poly(A) sites in each gene was 2.88.

Identification of TFs and lncRNAs

TFs play vital roles in regulating animal growth and development. The animal TFDB 2.0 database was employed to identify and classify TFs [34]. In total, 199 putative TFs were identified from 29 families, among which 16 novel TFs were identified. The numbers of enriched TF families were as follows: zf-C2H2 (58), ZBTB (31), CSD (12), TF_bZIP (12), bHLH (11), Homeobox (8), HMG (8), MYB (6) and ARID (6) (Fig. 4C and Table S5).

According to the prediction results of CNCI, CPC, Pfam, and PLEK tools, 2,312 transcripts were regarded

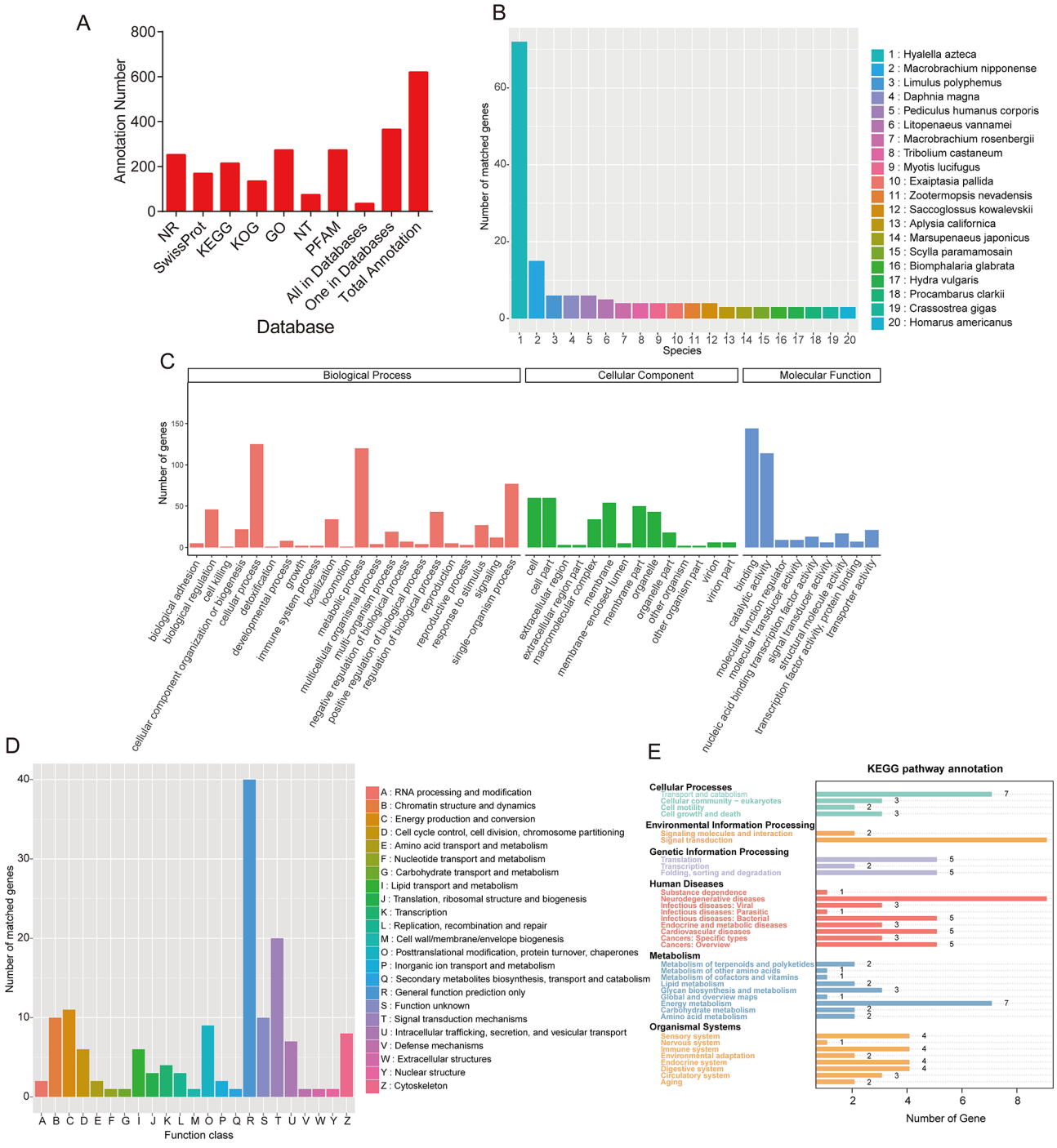

Fig. 3 Functional annotation of NGs. (**A**) Functional annotation of NGs across seven databases. (**B**) Nr homologous species distribution of NGs. (**C**) Distribution of GO terms for all annotated transcripts. (**D**) KOG enrichment of NGs. (**E**) KEGG pathway enrichment of NGs.

as putative non-coding RNAs. The 291 transcripts obtained from all four prediction tools were deemed as lncRNAs, and 203 (69.76%) of them were novel lncRNAs (Fig. 5A and Table S6). These lncRNAs were divided into 4 groups: antisense lncRNA (n=58, 19.93%), sense intronic lncRNA (n=19, 6.53%), sense overlapping lncRNA (n=26, 8.93%), and lincRNA (n=188, 64.60%) (Fig. 5B). Length distribution analysis showed that the lengths of lncRNAs ranged from 0.24 to 5.75 kb, and the average length was 2.34 kb (Fig. 5C). Additionally, The lncRNAs predicted have fewer exons when compared to the mRNAs and 228 (78.35%) single-exon lncRNAs were identified (Fig. 5D).

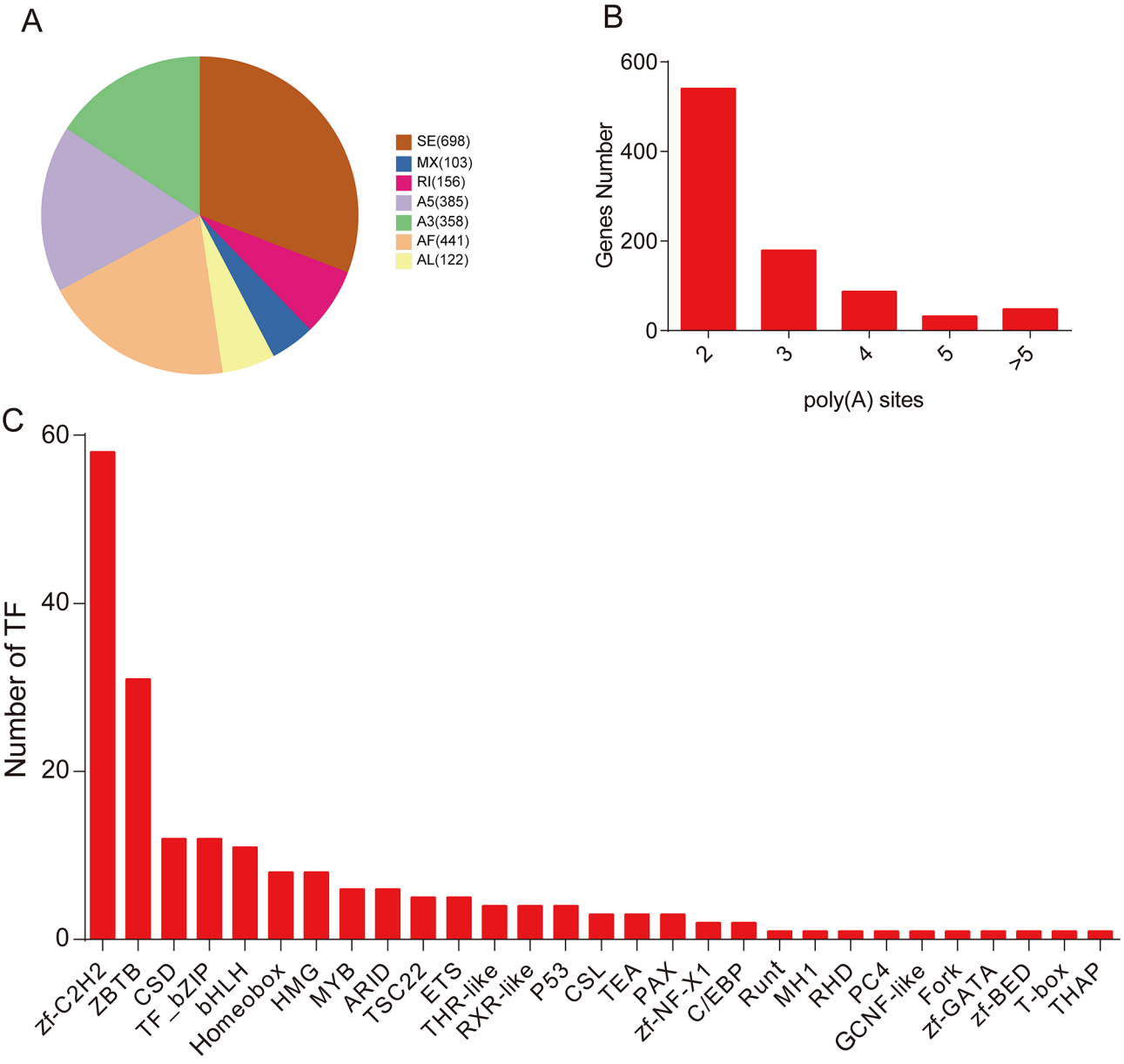

Fig. 4 Identification of AS, APA events and transcription factors according to the SMRT sequences. (**A**) Number and category of the identified AS events. (**B**) The number of poly(A) sites in each gene. (**C**) Number and type of the identified transcription factors

Discussion

The NGS technologies have been widely used to construct genome and transcriptome with significant advantages including accurate, cost-effective and high throughput [13–16, 35]. Therefore, the data about oriental river prawn transcriptome based on gene expression profiling and genome were mainly produced by NGS sequencing [36–41]. However, the fusion transcripts, full-length mRNAs, AS events and APA sites of oriental river prawn have not been well characterized due to the lack of full-length transcripts. PB SMRT sequencing can be used to directly obtain full-length transcripts without further assembly, thus overcoming the above-mentioned limitations [42–45].

The transcriptome analysis of oriental river prawn mainly involves gonads and hepatopancreas tissues in the previous study. For instance, a total of 78,408 isosequences were obtained in de novo transcriptome assembly data of oriental river prawn androgenic gland tissues, among which contain 57,619 non-redundant transcripts and 40 candidate NGs [3]. Besides, by transcriptome analysis of oriental river prawn ovarian, a total of 63,336 unigenes were assembled, and among 9 key DEGs may be related to sexual precocity [5]. Hepatopancreas is the largest functional organ of shrimp so its transcriptome

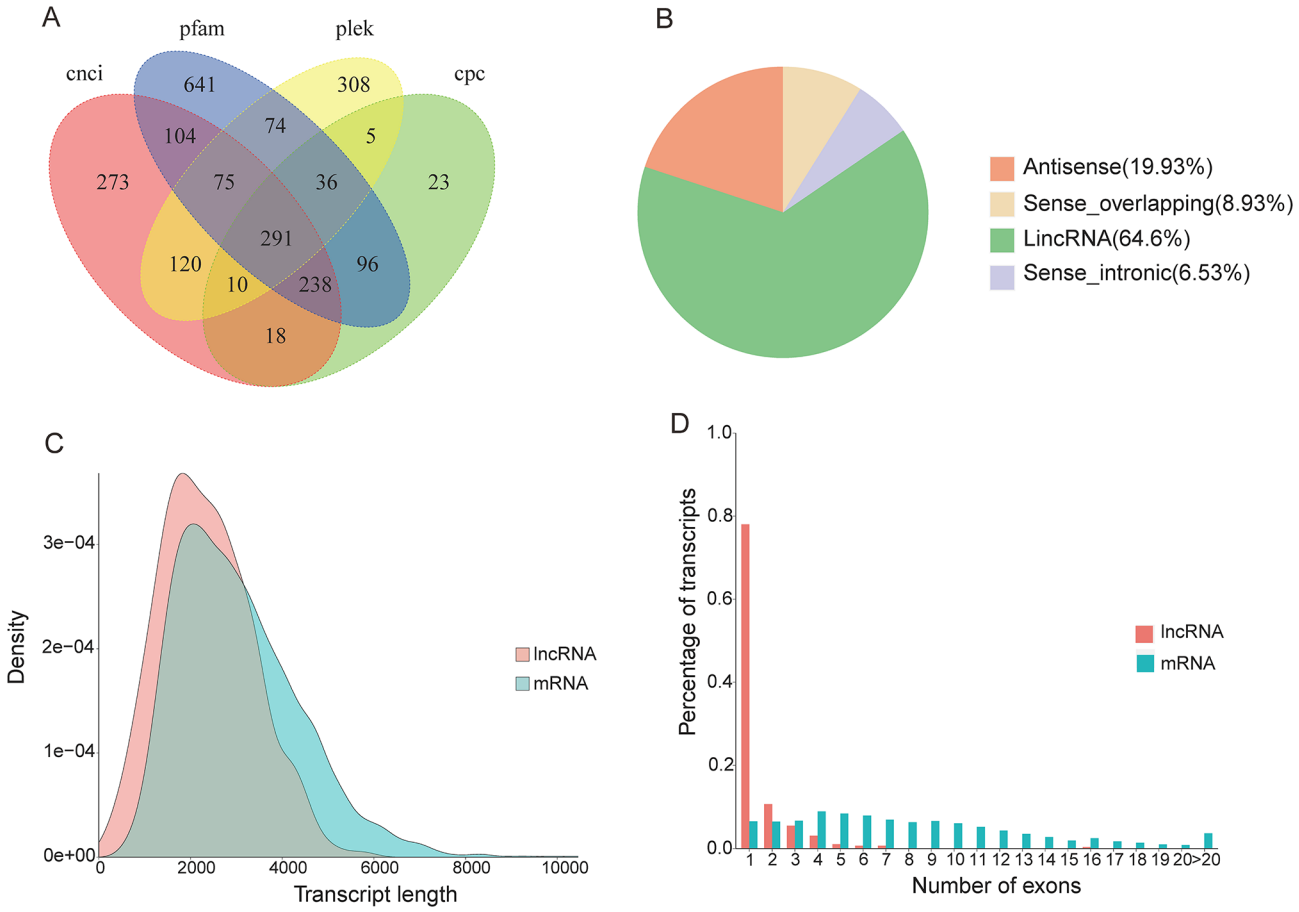

Fig. 5 Identification of lncRNA according to the SMRT sequences. (**A**) Venn diagram of lncRNA estimated by CPC, CNCI, Pfam, and PLEK tools. (**B**) Classification of the lncRNA types. (**C**) Length and density distributions of the annotated lncRNA and mRNA. (**D**) Comparison of the exon numbers of the annotated lncRNA and mRNA.

studies are the most common in oriental river prawn. These transcriptome sequencing produced large number of unigenes by using Illumina platform, and revealed differential gene expression profile and related signaling pathways enrichment rule under a variety of treatment conditions [46–48].This study is the first transcriptome analysis of oriental river prawn muscle using a hybrid sequencing approach. In total, 37.99 Gb subread data were retrieved and 584,498 CCS sequences were generated after correction. By detecting the sequences, 512,216 FLNC sequences were identified with an average length of 2,701 bp. After eliminating duplicate sequences, 21,008 consensus sequences were acquired. Moreover, the paired-end reads were retrieved using the Illumina platform, and were then employed to correct the consensus isoform sequences after quality filtering. Lastly, the combination of SMRT with Illumina data generated a total of 21,008 corrected consensus reads. After mapping the consensus reads against the oriental river prawn reference genome, the mapping rate was 92.94% (>70%), indicating the quality of the sequencing data is good [49]. In the previous work, multiple isoforms of anti-lipopolysaccharide factors (ALFs) were identified from the ridgetail prawn *Exopalaemon carinicauda* and showed different function in modulating the in vivo bacterial and viral propagation [50]. Base on proteomics informed by transcriptomics, eleven different black tiger shrimp *Penaeus monodon* hemocyanin (PmoHc) γ isoforms and one PmoHc β isoform were successfully identified in black tiger shrimp *P. monodon* and showed specific expression patterns in shrimp different stages of development. The average identity of amino acid sequence ranged from 24 to 97% between putative PmoHc gene isoforms [51]. In this study, 6,599 high-quality isoforms were obtained based on the PB full-length sequences, among which 5,537 and 766 were classified as novel isoforms from known genes and isoforms from NGs, respectively. These isoforms may effectively enrich the diversity of proteins in oriental river prawn.

Previous studies have shown that eukaryotic transcriptome is highly complex due to posttranscriptional processing (e.g., AS and APA) of precursor mRNAs [18, 52]. Here, AS and APA events were identified from oriental river prawn by using PB sequences. AS has contributed

greatly to enrich the functional and structural polymorphisms of genes and proteins [53–55]. In freshwater giant prawn (*Macrobrachium rosenbergii*), two crustacean hyperglycemic hormone (chh and chh- l) isoforms were identified and demonstrated to come from a Chh gene transcribed in an AS manner. The chh transcript contains exons I, II, and IV, whereas the chh-l transcript contains all 4 exons [56]. In addition, two Cactus (MnCactus-a and MnCactus-b) and four Taiman (MnTai-A, MnTai-B, MnTai-C, and MnTai-D) isoforms were characterized from oriental river prawns and proved to produce by AS [57, 58]. Cactus-a encodes a protein of 377 amino acids (aa) and Cactus-b encodes a protein of 471 aa [57]. The full-length cDNA of MnTai-A contains all exons (20) and encoded a protein of 1665 aa. The second to last (-exon2) and the third to last (-exon3) exons can be AS, and the deprivation of -exon2 or -exon3 produces MnTai-B or MnTai-C, respectively, whereas both exons are absent in MnTai-D. All these four isoforms were ubiquitous in a variety of tissues [58]. In this study, 2,263 AS events were found among 3,605 genes, which may provide more new knowledge about the complexity and diversity of isoforms of transcripts and corresponding proteins. For example, The full-length cDNA of twitchin-like contains a total of 9 exons, the variable splicing occurred on the fifth exon, eventually resulting in a splice variant containing 8 exons. PB sequencing is more effective than NGS for analyzing poly(A) sites [18, 59, 60]. In this study, a draft genome map of APA was constructed, which consisted of 2,555 poly(A) sites in 886 genes. These data may underestimate the exact number of APA genes due to the downregulated expression of proximal poly(A) sites.

LncRNA has been characterized in many species, which plays important parts in developmental and pathological processes [61, 62]. However, the lncRNAs identified by NGS are inaccurate due to a lack of poly(A) tails [63]. In our study, 291 lncRNAs were predicted according to SMRT sequencing data and 203 of these were identified as NGs, which could serve as lncRNA candidates for future functional characterization. In many species, NGs detected by full-length transcript sequencing effectively supplemented the reference genome data, such as Cattle (*Bos taurus*) [24], Gnetales (*Gnetum*) [42], and Perennial ryegrass (*Lolium perenne*) [64]. In this study, 620 NGs were detected when these transcripts were mapped with the oriental river prawn reference genome, which provided more comprehensive supplement data for genome sequences and gene functions in oriental river prawn.

Taken altogether, the current study represents an example of PB SMRT sequencing insights into the transcriptome complexity and diversity of oriental river prawn, which characterized full-length transcript and refined the annotation of the reference genome. These findings are beneficial for molecular breeding of oriental river prawn.

Conclusion

In summary, we identified 2,263 AS events, 2,555 APA sites, 620 NGs, 291 novel lncRNAs, and 197 TFs based on the full-length transcriptome analysis of oriental river prawn, which provided a strong molecular basis for exploring the transcriptome diversity of oriental river prawn. In addition, these data can be useful for elucidating the transcriptomic profile, understanding the genomic structure, and improving the draft genome annotation of oriental river prawn.

Materials and methods

Sample collection and RNA preparation

Specimens of 1-year-old adult oriental river prawn were collected from a wild population in Minjiang river, Sichuan, China. 5 individuals with body weights of 11.01–13.45 g were selected for sequencing. Fresh muscle tissues of 5 individuals were collected and immediately frozen in liquid nitrogen before carrying out RNA extraction. Subsequently, total RNA from muscle tissues were extracted by using TRIzol reagent (Takara, Japan) according to the manufacturer's instructions. The quality and quantity were assessed by agarose gel electrophoresis and Agilent Bioanalyzer 2100 System (Agilent, USA), respectively. The qualified RNA specimens were subjected to cDNA library construction and sequencing. We hereby declare that the study is reported in accordance with ARRIVE guidelines.

SMRT library preparation and PB sequencing

First, the qualified RNA samples were equally pooled together. Then, full-length cDNA synthesis was conducted using the SMARTer PCR cDNA Synthesis Kit (Clontech, USA). Next, the BluePippinTM Size Selection System (Sage Science, USA) was applied for cDNA size fractionation and length selection. Subsequently, the PB library was prepared using the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences, USA). Lastly, the PB Sequel platform was used for SMRT sequencing.

Illumina cDNA library construction and NGS analysis

Total RNA was extracted from independent biological replicates and prepared for double-stranded cDNA library construction. The first and second cDNA strands successively were synthesized using a NEBNext® Ultra™ RNA Library Prep Kit (NEB, USA). Next, an Illumina NovaSeq 6000 platform was used to sequence the qualified libraries to generate raw paired-end reads with 150-bp read length. Quality filtering was conducted with NGS QC Toolkit v2.3.3 [65], such as trimming the first five bases at the 5'-end and removing reads containing the low-quality bases ($QA \leq 30$) > 20% or ambiguous bases > 1%. Finally, the obtained Illumina clean reads

were assembled independently using Stringtie v2.1.1 and Hisat2 v2.1.0 for correcting PB long reads [66, 67].

Quality filtering and error correction

SMRTlink v8.0 software was employed to process the PB raw data based on the following parameters: minPasses=1, minLength=50, maxLength=15000. CCSs were generated from the subread.bam files (parameters: min_length 200, max_drop_fraction 0.8, no_polish TRUE, min_zscore −9999, min_passes 1, min_predicted_accuracy 0.8, max_length 18,000), and then these CCS. bam file were output. By searching for the poly(A) tail and the 5' and 3' adapters, the CCSs were classified into full-length and non-full-length reads. Full-length reads without chimeras were defined as FLNC reads. Then, these FLNC reads were clustered to clear redundancy by Iterative Clustering for Error Correction (ICE) and then corrected to obtain high-quality (post-correction accuracy above 99%) polished consensus reads by SMRT-Link built-in Arrow software (https://github.com/PacificBiosciences/pbbioconda). The LoRDEC v0.7 software [68] was employed to correct mismatches and nucleotide indels in consensus reads (parameters: -k 23; -s 3). Construction of a high-quality PB corrected consensus read dataset without redundant isoforms was then performed.

Mapping to the reference genome and structural analysis

GMAP v2017-06-20 [49] was used to align the corrected isoforms against the oriental river prawn reference genome (ASM1510439v1) based on the following parameters: –no-chimeras, –expand-offsets 1 - B 5 -f samse -n 1. The genome annotation file (http://gigadb.org/dataset/100843) was employed for the determination of genes and transcripts. Genome-guided transcriptome assembly was then carried out. Structure analysis of the transcripts was conducted using the TAPIS pipeline v1.2.1 [18]. AS events were identified and classified by SUPPA v2.3 [33]. TAPIS was used to analyze the APA events. The animal TFDB 2.0 database was employed for predicting transcription factors (TFs) [34].

Because of the limitation of library construction, only polyA tails-containing lncRNAs were obtained. The coding potential was calculated using the CNCI [69], PLEK [70], CPC [71], and Pfam database [72]. The transcripts (>200 bp) with at least 2 exons were chosen as lncRNA candidates. To ensure the accuracy of the results, only the lncRNAs identified simultaneously from the four tools were retained for further analysis.

Identification and functional annotation of novel genes

NGs were defined as those (compared to GigaDB genebuild) that did not match any annotation in the oriental river prawn reference genome (ASM1510439v1). The identified NGs were annotated by 7 databases, including NCBI-Nr (NCBI non-redundant protein sequences), NCBI-Nt (NCBI non-redundant nucleotide sequences), KOG/COG [73], Pfam [72], SwissProt [74], GO [75], and KEGG [76]. NCBI-BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) was employed for Nt analysis; Hmmscan (https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan) for Pfam analysis; Diamond v0.8.36 [77] for Nr, KOG/COG, KEGG, and Swiss-Prot analyses; and the E-value was set as "1e-5".

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-023-09442-x.

Supplementary Material 1
Supplementary Material 2
Supplementary Material 3
Supplementary Material 4
Supplementary Material 5
Supplementary Material 6
Supplementary Material 7
Supplementary Material 8
Supplementary Material 9
Supplementary Material 10
Supplementary Material 11
Supplementary Material 12
Supplementary Material 13

Acknowledgements

The authors would like to express their gratitude to EditSprings (https://www.editsprings.cn) for the expert linguistic services provided, and the policy of Talent Introduction and Training of Sichuan Academy of Agricultural Sciences for providing financial support.

Authors' contributions

Writing original draft, review and editing, Cheng-Yan Mou; Funding acquisition, Jian Zhou and Jun Du; Project administration, Zhou-Ming Qian; conceptualization and date curation, Lu Zhang and Qiang Li; Investigation and methodology, Zhi-Peng Huang; Resources, Hong-Yu Ke; Software and supervision, Yuan-Liang Duan; Validation, Zhong-Meng Zhao; Formal analysis, Yu Xiao and Hua-Dong Li; visualization, Han Zhao.

Funding

This work was supported by Sichuan Science and Technology Planning Project (2021YFYZ0015), Investigation on Fishery Resources and Environment in Key Waters of Northwest China and Agriculture Research System of China (CARS-46), "1 + 9" open competition mechanism to select the best candidates and scientific and technological project of Sichuan Academy of Agricultural Sciences (1 + 9KJGG004).

Data Availability

The raw bam files and Illumina RNA-Seq data have been deposited in the Sequence Read Archives (SRA) of the National Center for Biotechnology Information (NCBI) under the accession number PRJNA935961 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA935961) and PRJNA902553 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA902553), respectively.

Declarations

Ethics approval and consent to participate
This study is in compliance with all ethical regulations and was approved by the Animal Care and Use Committee of the Fishery Institute of the Sichuan Academy of Agricultural Sciences (20170226001 A).
All animal experiments were performed according to protocols that were approved by the ARRIVE guideline.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 26 September 2022 / Accepted: 11 June 2023
Published online: 20 June 2023

References

1. Ma K, Feng J, Lin J, Li J. The complete mitochondrial genome of Macrobrachium nipponense. Gene. 2011;487(2):160–5.
2. Wang D, Wu FX, Song DD, Gao HQ. Bureau of Fisheries, Ministry of Agriculture of the People's Republic of China. China fishery statistical yearbook. Bejing: China agriculture press;; 2021.
3. Jin S, Fu H, Zhou Q, Sun S, Jiang S, Xiong Y, Gong Y, Qiao H, Zhang W. Transcriptome analysis of androgenic gland for discovery of novel genes from the oriental river prawn, Macrobrachium nipponense, using Illumina Hiseq 2000. PLoS ONE. 2013;8(10):e76840.
4. Jin S, Bian C, Jiang S, Han K, Xiong Y, Zhang W, Shi C, Qiao H, Gao Z, Li R et al. A chromosome-level genome assembly of the oriental river prawn, Macrobrachium nipponense. Gigascience 2021, 10(1).
5. Jiang H, Li X, Sun Y, Hou F, Zhang Y, Li F, Gu Z, Liu X. Insights into sexual precocity of female Oriental River Prawn Macrobrachium nipponense through Transcriptome Analysis. PLoS ONE. 2016;11(6):e0157173.
6. Yuan H, Zhang W, Jin S, Jiang S, Xiong Y, Chen T, Gong Y, Qiao H, Fu H. Transcriptome analysis provides novel insights into the immune mechanisms of Macrobrachium nipponense during molting. Fish Shellfish Immunol. 2022;131:454–69.
7. Xue C, Xu K, Jin Y, Bian C, Sun S. Transcriptome analysis to Study the Molecular Response in the Gill and Hepatopancreas tissues of Macrobrachium nipponense to Salinity Acclimation. Front Physiol. 2022;13:926885.
8. Sun S, Wu Y, Jakovlic I, Fu H, Ge X, Qiao H, Zhang W, Jin S. Identification of neuropeptides from eyestalk transcriptome profiling analysis of female oriental river prawn (Macrobrachium nipponense) under hypoxia and reoxygenation conditions. Comp Biochem Physiol B Biochem Mol Biol. 2020;241:110392.
9. Yu B. Role of in silico tools in gene discovery. Mol Biotechnol. 2009;41(3):296–306.
10. Yu B. In silico gene discovery. Methods Mol Med. 2008;141:1–22.
11. Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res. 2014;42(22):13534–44.
12. Levy SE, Myers RM. Advancements in Next-Generation sequencing. Annu Rev Genomics Hum Genet. 2016;17:95–115.
13. Lan P, Li W, Schmidt W. Complementary proteome and transcriptome profiling in phosphate-deficient Arabidopsis roots reveals multiple levels of gene regulation. Mol Cell Proteomics. 2012;11(11):1156–66.
14. Oono Y, Kawahara Y, Yazawa T, Kanamori H, Kuramata M, Yamagata H, Hosokawa S, Minami H, Ishikawa S, Wu J, et al. Diversity in the complexity of phosphate starvation transcriptomes among rice cultivars based on RNA-Seq profiles. Plant Mol Biol. 2013;83(6):523–37.
15. Du H, Yu Y, Ma Y, Gao Q, Cao Y, Chen Z, Ma B, Qi M, Li Y, Zhao X, et al. Sequencing and de novo assembly of a near complete indica rice genome. Nat Commun. 2017;8:15324.
16. Carruthers M, Yurchenko AA, Augley JJ, Adams CE, Herzyk P, Elmer KR. De novo transcriptome assembly, annotation and comparison of four ecological and evolutionary model salmonid fish species. BMC Genomics. 2018;19(1):32.
17. Slatko BE, Gardner AF, Ausubel FM. Overview of next-generation sequencing Technologies. Curr Protoc Mol Biol. 2018;122(1):e59.
18. Abdel-Ghany SE, Hamilton M, Jacobi JL, Ngam P, Devitt N, Schilkey F, Ben-Hur A, Reddy AS. A survey of the sorghum transcriptome using single-molecule long reads. Nat Commun. 2016;7:11706.
19. Zuo C, Blow M, Sreedasyam A, Kuo RC, Ramamoorthy GK, Torres-Jerez I, Li G, Wang M, Dilworth D, Barry K, et al. Revealing the transcriptomic complexity of switchgrass by PacBio long-read sequencing. Biotechnol Biofuels. 2018;11:170.
20. Jia D, Wang Y, Liu Y, Hu J, Guo Y, Gao L, Ma R. SMRT sequencing of full-length transcriptome of flea beetle Agasicles hygrophila (Selman and Vogt). Sci Rep. 2018;8(1):2197.
21. Jiao WB, Schneeberger K. The impact of third generation genomic technologies on plant genome assembly. Curr Opin Plant Biol. 2017;36:64–70.
22. Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, et al. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019;37(10):1155–62.
23. Zhang H, Xu H, Liu H, Pan X, Xu M, Zhang G, He M. PacBio single molecule long-read sequencing provides insight into the complexity and diversity of the Pinctada fucata martensii transcriptome. BMC Genomics. 2020;21(1):481.
24. Chang T, An B, Liang M, Duan X, Du L, Cai W, Zhu B, Gao X, Chen Y, Xu L, et al. PacBio single-molecule Long-Read sequencing provides New Light on the complexity of full-length transcripts in cattle. Front Genet. 2021;12:664974.
25. Chen SY, Deng F, Jia X, Li C, Lai SJ. A transcriptome atlas of rabbit revealed by PacBio single-molecule long-read sequencing. Sci Rep. 2017;7(1):7648.
26. Chen H, Lin L, Xie M, Zhong Y, Zhang G, Su W. Survey of the Bradysia odoriphaga Transcriptome using PacBio single-molecule Long-Read sequencing. Genes (Basel) 2019, 10(6).
27. Teng K, Teng W, Wen H, Yue Y, Guo W, Wu J, Fan X. PacBio single-molecule long-read sequencing shed new light on the complexity of the Carex breviculmis transcriptome. BMC Genomics. 2019;20(1):789.
28. Wang B, Tseng E, Regulski M, Clark TA, Hon T, Jiao Y, Lu Z, Olson A, Stein JC, Ware D. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nat Commun. 2016;7:11708.
29. Wang XM, Chen SY, Shi X, Liu DN, Zhao P, Lu YZ, Cheng YB, Liu ZS, Nie XJ, Song WN, et al. Hybrid sequencing reveals insight into heat sensing and signaling of bread wheat. Plant J. 2019;98(6):1015–32.
30. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczesniak MW, Gaffney DJ, Elo LL, Zhang X, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13.
31. Sharon D, Tilgner H, Grubert F, Snyder M. A single-molecule long-read survey of the human transcriptome. Nat Biotechnol. 2013;31(11):1009–14.
32. Xu Z, Peters RJ, Weirather J, Luo H, Liao B, Zhang X, Zhu Y, Ji A, Zhang B, Hu S, et al. 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. 2015;82(6):951–61.
33. Alamancos GP, Pages A, Trincado JL, Bellora N, Eyras E. Leveraging transcript quantification for fast computation of alternative splicing profiles. RNA. 2015;21(9):1521–31.
34. Zhang HM, Liu T, Liu CJ, Song S, Zhang X, Liu W, Jia H, Xue Y, Guo AY. AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors. Nucleic Acids Res. 2015;43(Database issue):D76–81.
35. Li Z, Xu C, Li K, Yan S, Qu X, Zhang J. Phosphate starvation of maize inhibits lateral root formation and alters gene expression in the lateral root primordium zone. BMC Plant Biol. 2012;12:89.
36. Zhang Y, Jiang S, Qiao H, Xiong Y, Fu H, Zhang W, Gong Y, Jin S, Wu Y. Transcriptome analysis of five ovarian stages reveals gonad maturation in female Macrobrachium nipponense. BMC Genomics. 2021;22(1):510.
37. Zhu P, Wang H, Zeng Q. Comparative transcriptome reveals the response of oriental river prawn (Macrobrachium nipponense) to sulfide toxicity at molecular level. Aquat Toxicol. 2021;230:105700.
38. Hu Y, Fu Y, Jin S, Fu H, Qiao H, Zhang W, Jiang S, Gong Y, Xiong Y, Wu Y, et al. Comparative transcriptome analysis of lethality in response to RNA interference of the oriental river prawn (Macrobrachium nipponense). Comp Biochem Physiol Part D Genomics Proteomics. 2021;38:100802.
39. Jin S, Fu Y, Hu Y, Fu H, Jiang S, Xiong Y, Qiao H, Zhang W, Gong Y, Wu Y. Transcriptome profiling analysis of the Testis after Eyestalk ablation for selection of the candidate genes involved in the male sexual development in Macrobrachium nipponense. Front Genet. 2021;12:675928.
40. Jin S, Hu Y, Fu H, Sun S, Jiang S, Xiong Y, Qiao H, Zhang W, Gong Y, Wu Y. Analysis of testis metabolome and transcriptome from the oriental river prawn (Macrobrachium nipponense) in response to different temperatures

and illumination times. Comp Biochem Physiol Part D Genomics Proteomics. 2020;34:100662.
41. Xu L, Fu Y, Fu H, Zhang W, Qiao H, Jiang S, Xiong Y, Jin S, Gong Y, Wang Y, et al. Transcriptome analysis of hepatopancreas from different living states oriental river prawn (Macrobrachium nipponense) in response to hypoxia. Comp Biochem Physiol Part D Genomics Proteomics. 2021;40:100902.
42. Deng N, Hou C, Ma F, Liu C, Tian Y. Single-molecule Long-Read sequencing reveals the diversity of full-length transcripts in Leaves of Gnetum (Gnetales). Int J Mol Sci 2019, 20(24).
43. Ardui S, Ameur A, Vermeesch JR, Hestand MS. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res. 2018;46(5):2159–68.
44. Nakano K, Shiroma A, Shimoji M, Tamotsu H, Ashimine N, Ohki S, Shinzato M, Minami M, Nakanishi T, Teruya K, et al. Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. Hum Cell. 2017;30(3):149–61.
45. Oikonomopoulos S, Bayega A, Fahiminiya S, Djambazian H, Berube P, Ragoussis J. Methodologies for transcript profiling using Long-Read Technologies. Front Genet 2020, 11.
46. Xu Z, Li T, Li E, Chen K, Ding Z, Qin JG, Chen L, Ye J. Comparative transcriptome analysis reveals molecular strategies of oriental river prawn Macrobrachium nipponense in response to acute and chronic nitrite stress. Fish Shellfish Immunol. 2016;48:254–65.
47. Yu J, Sun J, Zhao S, Wang H, Zeng Q. Transcriptome analysis of oriental river Prawn(Macrobrachium nipponense)Hepatopancreas in response to ammonia exposure. Fish Shellfish Immunol. 2019;93:223–31.
48. Yi C, Lv X, Chen D, Sun B, Guo L, Wang S, Ru Y, Wang H, Zeng Q. Transcriptome analysis of the Macrobrachium nipponense hepatopancreas provides insights into immunoregulation under Aeromonas veronii infection. Ecotoxicol Environ Saf. 2021;208:111503.
49. Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005;21(9):1859–75.
50. Lv X, Li S, Zhang C, Xiang J, Li F. Multiple Isoforms of Anti-Lipopolysaccharide factors and their antimicrobial functions in the Ridgetail Prawn Exopalaemon carinicauda. Mar Drugs 2018, 16(5).
51. Mendoza-Porras O, Kamath S, Harris JO, Colgrave ML, Huerlimann R, Lopata AL, Wade NM. Resolving hemocyanin isoform complexity in haemolymph of black tiger shrimp Penaeus monodon - implications in aquaculture, medicine and food safety. J Proteom. 2020;218:103689.
52. Marquez Y, Brown JW, Simpson C, Barta A, Kalyna M. Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. Genome Res. 2012;22(6):1184–95.
53. Han H, Braunschweig U, Gonatopoulos-Pournatzis T, Weatheritt RJ, Hirsch CL, Ha KCH, Radovani E, Nabeel-Shah S, Sterne-Weiler T, Wang J, et al. Multilayered control of Alternative Splicing Regulatory networks by transcription factors. Mol Cell. 2017;65(3):539–553e537.
54. Niklas KJ, Bondos SE, Dunker AK, Newman SA. Rethinking gene regulatory networks in light of alternative splicing, intrinsically disordered protein domains, and post-translational modifications. Front Cell Dev Biol. 2015;3:8.
55. Ule J, Blencowe BJ. Alternative Splicing Regulatory Networks: functions, mechanisms, and evolution. Mol Cell. 2019;76(2):329–45.
56. Chen SH, Lin CY, Kuo CM. Cloning of two crustacean hyperglycemic hormone isoforms in freshwater giant prawn (Macrobrachium rosenbergii): evidence of alternative splicing. Mar Biotechnol (N Y). 2004;6(1):83–94.
57. Huang Y, Si Q, Du J, Ren Q. Yorkie negatively regulates the expression of antimicrobial proteins by inducing Cactus transcription in Prawns Macrobrachium nipponense. Front Immunol. 2022;13:828271.
58. Huang X, Ma F, Zhang R, Dai X, Ren Q. Taiman negatively regulates the expression of antimicrobial peptides by promoting the transcription of cactus in Macrobrachium nipponense. Fish Shellfish Immunol. 2020;105:152–63.
59. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008;320(5881):1344–9.
60. Wang T, Wang H, Cai D, Gao Y, Zhang H, Wang Y, Lin C, Ma L, Gu L. Comprehensive profiling of rhizome-associated alternative splicing and alternative polyadenylation in moso bamboo (Phyllostachys edulis). Plant J. 2017;91(4):684–99.
61. Zhu J, Fu H, Wu Y, Zheng X. Function of lncRNAs and approaches to lncRNA-protein interactions. Sci China Life Sci. 2013;56(10):876–85.
62. Jarroux J, Morillon A, Pinskaya M. History, Discovery, and classification of lncRNAs. Adv Exp Med Biol. 2017;1008:1–46.
63. Yang L, Duff MO, Graveley BR, Carmichael GG, Chen LL. Genomewide characterization of non-polyadenylated RNAs. Genome Biol. 2011;12(2):R16.
64. Xie L, Teng K, Tan P, Chao Y, Li Y, Guo W, Han L. PacBio single-molecule long-read sequencing shed new light on the transcripts and splice isoforms of the perennial ryegrass. Mol Genet Genomics. 2020;295(2):475–89.
65. Patel RK, Jain M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS ONE. 2012;7(2):e30619.
66. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–60.
67. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33(3):290–5.
68. Salmela L, Rivals E. LoRDEC: accurate and efficient long read error correction. Bioinformatics. 2014;30(24):3506–14.
69. Sun L, Luo H, Bu D, Zhao G, Yu K, Zhang C, Liu Y, Chen R, Zhao Y. Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts. Nucleic Acids Res. 2013;41(17):e166.
70. Li A, Zhang J, Zhou Z. PLEK: a tool for predicting long non-coding RNAs and messenger RNAs based on an improved k-mer scheme. BMC Bioinformatics. 2014;15:311.
71. Kong L, Zhang Y, Ye ZQ, Liu XQ, Zhao SQ, Wei L, Gao G. CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. Nucleic Acids Res 2007, 35(Web Server issue):W345–349.
72. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, et al. The pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016;44(D1):D279–285.
73. Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. 2000;28(1):33–6.
74. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. Nucleic Acids Res. 1999;27(1):49–54.
75. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25(1):25–9.
76. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucleic Acids Res. 2004;32(Database issue):D277–280.
77. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12(1):59–60.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.