Pakistan J. Zool., vol. 50(6), pp 2273-2278, 2018. DOI: http://dx.doi.org/10.17582/journal.pjz/2018.50.6.2273.2278

De novo Assembly and Annotation of the Whole Transcriptome of *Penaeus penicillatus*

Binbin Shan, Yan Liu, Changping Yang, Shengnan Liu and Dianrong Sun*

Key Laboratory of South China Sea Fishery Resources Exploitation and Utilization, Ministry of Agriculture, Guangzhou 510000, P.R. China

ABSTRACT

In this study, we generated the whole transcriptome of *Penaeus penicillatus* from four combined tissues (eyestalk, muscle, intestinal and viscus) using High-seq sequencing technology. A total of 65,703,216 high-quality clean reads were generated to produce 27,850 non-redundant transcripts with a mean length of 918 nt using Trinity and Tgicl software. For homological alignment, 13,083 unigenes had significant hits in Nr database. And then, 10,467 unigenes were annotated into three ontologies: biological processes, cellular components, and molecular functions. Moreover, 18,621 unigenes were mapped 25 different clusters of eukaryotic proteins. In addition, a total of 17,671 unigenes were classified into 311 KEGG pathways. Finally, we predicted the coding sequences of 13,072 unigenes and obtained 9,222 SSRs in the present study. The whole transcriptome is an important foundation for future genomic research on the *P. penicillatus* and can provide comprehensively understanding and further characterizations of transcriptomes of non-model organisms.

Article Information
Received 05 August 2017
Revised 10 September 2017
Accepted 27 September 2017
Available online 11 October 2018

Authors' Contribution
DS designed and conducted the study. BS and YL performed the experiments and analyzed data. SL and CY helped in manuscript preparation. BS finalized the article.

Key words
***Penaeus penicillatus*, Transcriptome, High-Seq sequencing technology, *De novo* assembly, Annotation.**

INTRODUCTION

The red tail prawn (*Penaeus penicillatus*) is mainly distributed from Pakistan to Indonesia in the Indo-West Pacific. It was considered one of the most important commercial marine shrimps in the East and South China Seas in the 1980s and 1990s (Zhang *et al*., 2010; Cao *et al*., 2012). Increased market demand and pressure from over-fishing has led to an expansion of penaeid shrimp culturing (Saoud *et al*., 2003; Navakanitworakul *et al*., 2012). Despite the commercial significance, only limited genetic information is available for many species of penaeid shrimp (Powell *et al*., 2015). What' more, limited information is available for the gene expression of *P. penicillatus*. Large quantities of literatures on *P. penicillatus* have focused on growth, reproduction, population genetic structure and so on (Chen and Jenn, 1991; Cao *et al*., 2012). Understanding the genetic basis underlying the regulation of traits in this species can provide information for selection programmes that can enhance the development of quality broodstock (Santos *et al*., 2014).

Recently, high-throughput RNA-seq have been widely used in the biology research and provide abilities to capture a global gene expression profile for non-model organisms (Dittel *et al*., 2009; Stillman and Tagmount, 2009; Smith *et al*., 2013; Xia *et al*., 2013). In early work, studies based on transcriptome scan have also discovered adaptively important candidate genes and genomic regions in non-model aquatic species including *Astacus astacus* (Theissinger *et al*., 2014), *Eriocheir sinensis* (Li *et al*., 2014), *Nibea albiflora* (Zhan *et al*., 2016) and *Procambarus clarkii* (Du *et al*., 2016). However, relatively no data set is available for transcriptomes of *P. penicillatus*.

In the present study, RNA-seq technology was used to capture a significant portion of whole transcriptomes of *P. penicillatus*. This information was used to generate expression profiles and to estimate the transcript abundance. The functions of genes and pathways were annotated and classified by the Nr, Nt, Swiss-Prot, GO, COG and KEGG databases. This study firstly characterized the transcriptome of *P. penicillatus*, and provided a novel understanding for research on the gene functions, molecular events and signaling pathways related to the regulatory mechanism of *P. penicillatus*, which further enhanced understanding the mechanism of other aquatic organisms.

MATERIALS AND METHODS

Tissue and material

P. penicillatus were collected from the coastal water of Beihai (China) on April 4th, 2017 (21°32'N, 109°8'E). Female samples were used for lab experiments to reduce the amount of "noise" in gene expression signals (Kammenga *et al*., 2007). The sample in healthy conditions were acclimated to 6 days in aquarium with fully recirculating aerated seawater (sea water temperature about 28°C

* Corresponding author: drsun73@163.com
0030-9923/2018/0006-2273 $ 9.00/0

and 28 salinity) before dissection (Table I). Then, tissue samples of eyestalk, muscle, sexual, intestinal and viscus of *P. penicillatus* dissected on-site and preserved in RNA-later solution (Life Technologies) before transportation to the laboratory and storage at -80°C.

Total RNA extraction and illumina sequencing

Total RNA was extracted from each tissue using the PureLink RNA Mini Kit (Life Technologies) with a TRIzol step. RNA-seq library preparation and quality control relied on Agilent 2100 Bioanalyzer and ABI StepOnePlus Real-Time PCR System. Subsequently, sequencing was conducted by Illumina HiSeq™ 2000 platform.

De novo *assembly and transcriptome analysis*

In present research, the raw reads in FASTQ format were carried on the quality detection using FastQC and filtered by removing reads with sequencing adaptors, unknown nucleotides (N ratio>5%) and low quality (quality scores< =15). The remaining high-quality reads were used for *de novo* assembly employing Trinity software and then the redundancy sequences were removed using Tgicl software and further were spliced the unigenes (Grabherr *et al*., 2011). The expression level of overall transcripts was normalized to determine FPKM using RSEM and Bowtie2 with default settings (Li and Dewey, 2011; Langmead and Salzberg, 2012). Briefly, a total of 65,703,216 clean reads were screened out from 84,667,798 raw reads, corresponding to 6,570,321,600 nucleotides (Fig. 1A). The Q20/Q30 percentage (percentage of bases whose quality was greater than 20/30 in clean reads), N percentage (percentage of uncertain bases after filtering), and GC percentage were 94.23%/83.95%, 0.00% and 48.28%, respectively. All clean sequences were randomly assembled to produce 36,676 contigs with an N50 of 1,517 nt. The contigs were further trimmed into 27,850 unigenes with an N50 of 1631 nt. The result of transcript abundance showed that 3,102 lowly expression transcripts (FPKM<1) and 22,975 transcripts had significant expression signals in *P. penicillatus*. The summary of assembly analysis is displayed in Table II and the distribution of the unigene length is shown in Figure 1B. In order to further quantitative assessment of the assembly and annotation completeness, we applied

Table I.- MIxS descriptors.

Item	Description
Investigation type	Eukaryote
Project name	Transcriptome for *Penaeus penicillatus*
Collection data	April 4th, 2017
Lat Long	21°32'N, 109°8'E
Geo location name	Beihai, China
Environment	Marine water
Biotic relationship	Free living
Trophic level	Heterotroph
Temp	29°C
Salinity	26 PSU
Estimated size	6.57 Gb
Sequencing meth	IlluminaHiSeq™ 2000
Assembly	Trinity and Tgicl software package
Annot source	Nr/Nt/Swiss-Prot/KEGG/ KOG/ InterPro/GO
BioProject ID of raw reads	PRJNA360178
Accession number of raw reads	SRR5145906
Accession number of transcripts	GFFH00000000

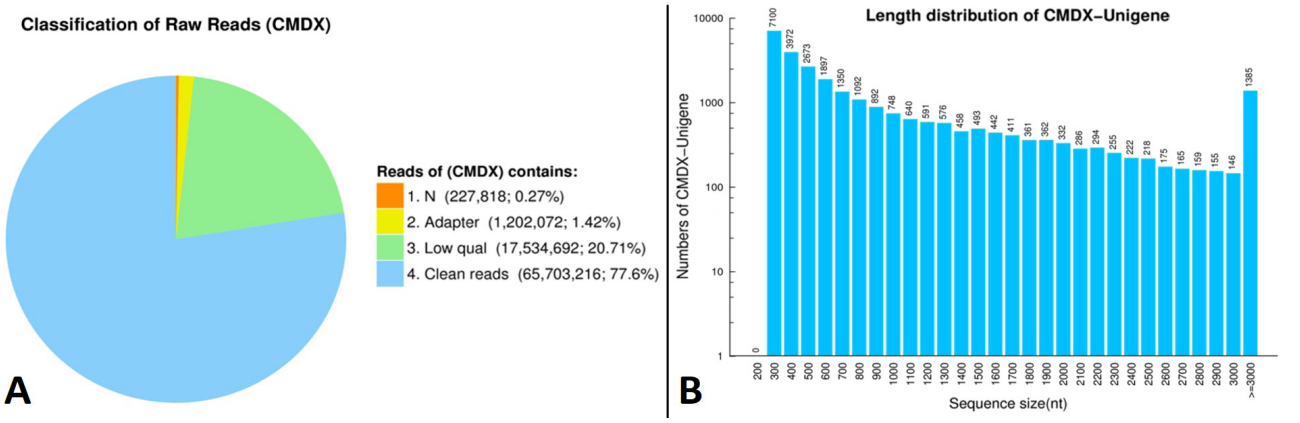

Fig. 1. A, classification of raw reads; B, length distribution of CMDX-Unigene.

the BUSCO v1.22 (Simão *et al*., 2015) with default setting and downloaded from Ensembl Metazoa as the reference. The result showed that 81.5% protein-coding genes were found in our assembled transcripts, the proportions of mismatching genes were calculated and only 11.5%. The BUSCO analysis shows that this transcriptome assembly will provide a good basis for further transcriptomic research of the *P. penicillatus*.

Table II.- Statistics for the assembly of *P. penicillatus* whole transcriptome.

	Total No.	Total length (nt)	Mean length (nt)	N50	N90	GC (%)
Contig	36,676	29,768,415	811	1,517	307	43.09
Unigene	27,850	25,571,120	918	1,631	358	43.30

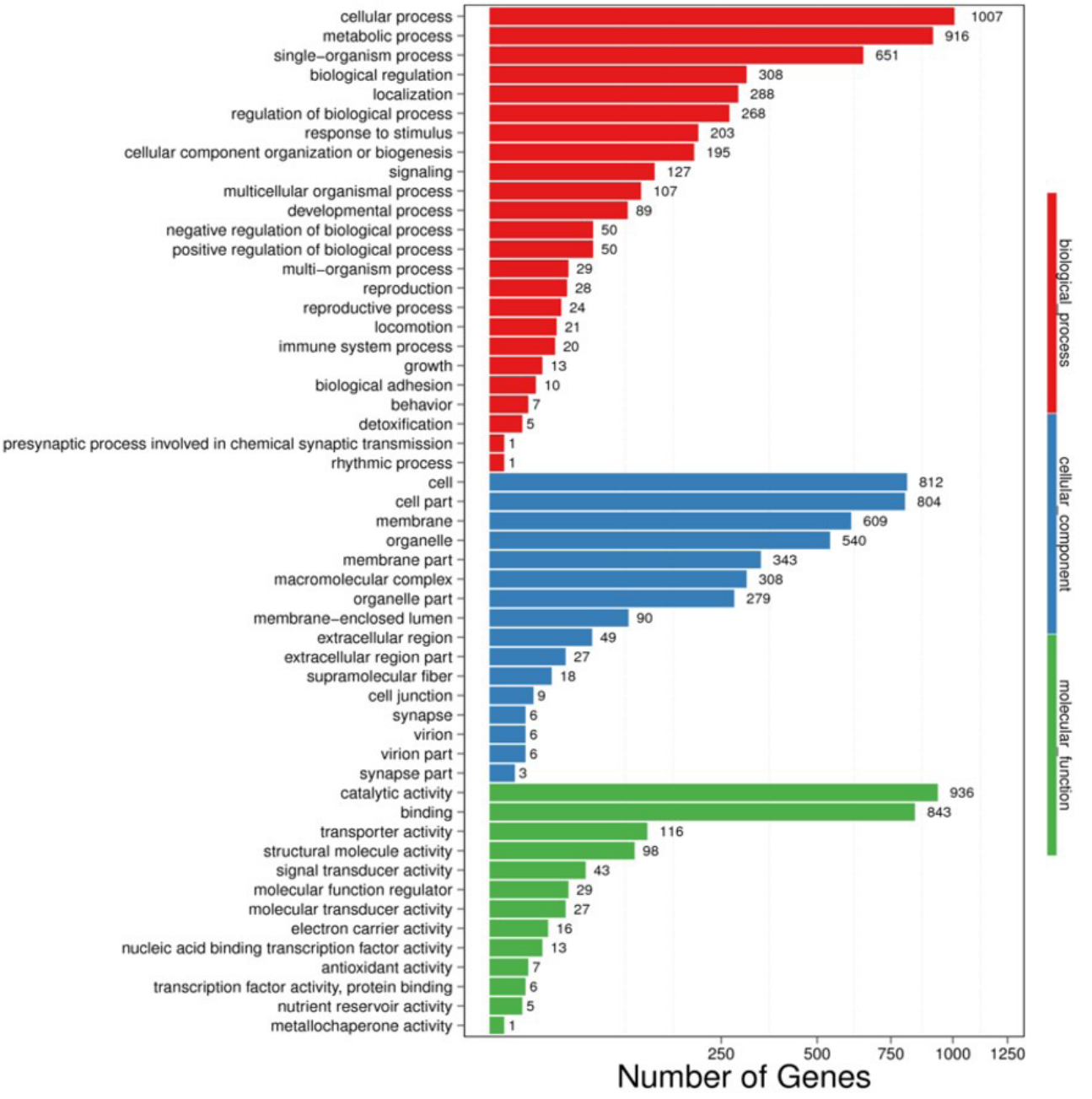

Fig. 2. Gene ontology classification for the transcriptome from *P. penicillatus*.

Functional annotation and classifications and predict the coding sequences

To further analysis the functional classifications of unigenes according on gene annotation, we acquired 15,474 homology searches by comparing all unigenes against the Nr, Nt, Swiss-Prot, KEGG, KOG, InterPro and GO database using the Blastx alignment (E-value < 1e-5). Of all annotated unigenes, 13,083, 6,873, 11,136, 10,887, 10,380, 9,594 and 2,080 unigenes had significant matches with sequences in the Nr, Nt, Swiss-Prot, KEGG, KOG, InterPro and GO databases, respectively.

From the annotated information of six databases, 13,083 unigenes had significant hits in Nr database and distributed among 568 other species (Supplementary Fig. S1). In brief, about 1,254 (9.58%), 695 (5.31%), 585 (4.47%), and 436 (3.33%), unigenes were matched with genes from *Zootermopsis nevadensis*, *Limulus polyphemus*, *Daphnia pulex*, and *Lingula anatina,* respectively. The remaining unigenes (10,097, 77.31%) were hits in the other species.

As an internationally standardized gene function classification system, GO classification can provide a control vocabulary with dynamically updating and exactly define the characteristics of genes and their products (Lena *et al*., 2015; Chen *et al*., 2015). On the basis of the Nr annotation, GO functional classifications of the unigenes were performed. In total, 10,467 unigenes with BLAST matches had GO categories. Among these functional groups, the biological process assignments made up the majority (4418, 42.21%) followed by cellular component (3909, 37.35%) and molecular function (2,140, 20.44%). The associated functions of transcripts covered a broad range of GO categories. In the biological process category, cellular process (1,007, 22.79%), metabolic process (916, 20.73%) and single-organism process (651, 14.74%) were prominently overrepresented. In the cellular component category, cell (812, 20.77%) and cell part (804, 20.57%) took up most of the categories (Fig. 2).

Against clusters of orthologous groups for eukaryotic complete genomes (KOG) database, 18,621 transcripts with BLAST matches (E-value 10-5) were assigned to 25 classifications of four main classes (Information storage and processing, Metabolism, Cellular processes and signaling, and Poorly characterized) (Supplementary Fig. S2). Among these clusters, general function prediction (2656, 14.26%) represented the largest group, followed by signal transduction mechanisms (2206, 11.85%), Function unknown (1560, 8.38%), posttranslational modification (1358, 7.29%) and transcription (1313, 7.05%).

The research of biochemical metabolic pathway was based on KEGG pathway annotation. Based on a comparison against the KEGG database using Blastx with an E-value threshold of 1e-5, 17,671 unigenes have significant matches and were assigned to 6 main categories including 311 KEGG pathways. Among the 6 main categories, human diseases category is the largest represented class (4,276, 24.20%), the top hit pathway is the signal transduction pathway in environmental information processing category with 1750 unigenes in (9.9% of KEGG annotated unigenes). These results could provide valuable resources for investigating specific processes, functions and pathways of *P. penicillatus* and the top 30 statistically significant KEGG classifications (Supplementary Table I).

Prediction of the coding sequences (CDS) and transcription factors (TFs)

The CDS in unigene was predicted by TransDecoder software (Lee *et al*., 2015). Functional annotation of the unigenes was performed using the Swiss-Prot database by using Blastx with an E-value cutoff of 1e-5. When unigenes did not align to any of the above databases, Hmmscan (HMMER) software (version 4.1) was used for annotating though functional domain prediction (Grabherr *et al*., 2011). The result showed that 13,072 CDSs (Direction of the sequences was 5'-3') were confirmed based on the maximum value in the alignment results and these coding sequences will be translated into protein sequences according to the standard codes.

TFs are considered as key proteins and play crucial roles in regulating gene expression by binding to specific DNA sequences. A total of 3,196 putative TF genes from 60 families, were identified in the transcriptome of *P. penicillatus*. The proportion of each major TF family in *P. penicillatus* was shown in Supplementary Figure S3. Among these families, the first three largest families were zf-C2H2 (1032), bHLH (530) and Homeobo (306).

SSR markers identification

In the present study, we detected SSRs among>1000 bp putative unigene sequences using the MISA software (Thiel *et al*., 2003). In total, 9,222 potential simple sequence repeats (SSRs) were identified in 6,337 unigenes. Of these 6,337 unigenes, 1,925 sequences contained more than one SSR. Six types of SSRs were detected in the *P. penicillatus*, and the most abundant repeat type was di-nucleotide repeat motif (4,034, 43.74%), followed by mono-nucleotide repeat motif (2,219, 24.06%), tri-nucleotide (2,194, 23.79%) (Supplementary Table II).

CONCLUSION

To the best of our knowledge, this study is the first systematic research of the whole transcriptome in *P.*

penicillatus. Our results presented here provide a key resource for future investigations into gene functions, molecular events, and signaling pathways related to the regulatory mechanism of *P. penicillatus* and we can obtain a comprehensively understanding for the physiological and genomic level of *P. penicillatus* by characterizing the transcriptome. Finally, the present study adds a substantial contribution to the sequence data available for *P. penicillatus*, providing valuable resources for further studies.

Nucleotide sequence accession numbers

The raw reads of the present study were uploaded to the SRA databases of NCBI under BioProject PRJNA392833, with accession number SRR5798783. The assembled and annotated transcript has been deposited at DDBJ/EMBL/GenBank under the accession GFRT00000000 (https://www.ncbi.nlm.nih.gov/nuccore/GFRT00000000). The version described in this paper is the first version, GFRT01000000.

ACKNOWLEDGMENTS

This work was supported by Sino-Vietnam Fishery Stock Enhancement and Conservation in Beibu Gulf.

Supplementary material

There is supplementary material associated with this article. Access the material online at: http://dx.doi.org/10.17582/journal.pjz/2018.50.6.2273.2278

Supplementary material includes: Table I, the top 30 statistically significant KEGG classifications in the transcriptome of *P. penicillatus*; Table II, summary of the EST-SSRS in *P. penicillatus* unigenes; Figure S1, species distribution of Blast top-hits against the NCBI Nr database; Figure S2, Cluster of orthologous groups (KOG) classification of putative proteins; Figure S3, the proportion of *P. penicillatus* transcripts belonging to different major TF families.

Statement of conflict of interest

Authors have declared no conflict of interest.

REFERENCE

Cao, Y.Y., Li, Z.B., Zhang, G.L., Chen, X.J., Chen, L.N. and Li, Q.H., 2012. Isolation and characterization of ten microsatellite markers of *Fenneropenaeus penicillatus*. *Conserv. Genet. Resour.,* **4**: 261-263. https://doi.org/10.1007/s12686-011-9520-6

Chen, K., Li, E.C., Li, T.Y., Xu, C., Wang, X.D., Lin, H.Z., Qin, J.Q. and Chen, L.Q., 2015. Transcriptome and molecular pathway analysis of the hepatopancreas in the Pacifc white shrimp *Litopenaeus vannamei* under chronic low-salinity stress. *PLoS One,* **10**: e0131503. https://doi.org/10.1371/journal.pone.0131503

Chen, H.Y. and Jenn, J.S., 1991. Combined effects of dietary phosphatidylcholine and cholesterol on the growth, survival and body lipid composition of marine shrimp, *Penaeus penicillatus*. *Aquaculture,* **96**: 167-178. https://doi.org/10.1016/0044-8486(91)90147-Y

Dittel, A.I. and Epifanio, C.E., 2009. Invasion biology of the Chinese mitten crab *Eriochier sinensis*: A brief review. *J. exp. Mar. Biol. Ecol.,* **374**: 79-92. https://doi.org/10.1016/j.jembe.2009.04.012

Du, Z.Q., Jin, Y.H. and Ren, D.M., 2016. In-depth comparative transcriptome analysis of intestines of red swamp crayfish, *Procambarus clarkii*, infected with WSSV. *Sci. Rep. U.K.,* **6**: 1-12. https://doi.org/10.1038/srep26780

Garber, M., Grabherr, M.G., Guttman, M. and Trapnell, C., 2011. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods,* **8**: 469-477. https://doi.org/10.1038/nmeth.1613

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Raychowdhury, R., Zeng, Q.D., Chen, Z.H., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N. and Regev, A., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.,* **29**: 644-652. https://doi.org/10.1038/nbt.1883

Kammenga, J.E., Herman, M.A., Ouborg N.J., Johnson, L. and Breitling, R., 2007. Microarray challenges in ecology. *Trends Ecol. Evol.,* **22**: 273-279. https://doi.org/10.1016/j.tree.2007.01.013

Li, E.C., Wang, S.L., Li, C., Wang, X.D., Chen, K. and Chen, L.Q., 2014. Transcriptome sequencing revealed the genes and pathways involved in salinity stress of Chinese mitten crab, *Eriocheir sinensis*. *Physiol. Genom.,* **46**: 177-190. https://doi.org/10.1152/physiolgenomics.00191.2013

Langmead, B. and Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods,* **9**: 357-359. https://doi.org/10.1038/nmeth.1923

Li, B. and Dewey, C.N., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics,* **12**: 323. https://doi.org/10.1186/1471-2105-12-323

Lena, D.P., Domeniconi, G., Margara, L. and M, G.,

2015. GOTA: GO term annotation of biomedical literature. *BMC Bioinformatics,* **16**: 346. https://doi.org/10.1186/s12859-015-0777-8

Lee, B.Y., Kim, H.S., Choi, B.S., Hwang, D.S., Choi, A.Y., Han, J.H, Won, E.J., Choi, I.Y., Lee, S.H., Om, A.S., Park, H.G. and Lee, J.S., 2015. RNA-seq based whole transcriptome analysis of the cyclopoid copepod Paracyclopina nana focusing on xenobiotics metabolism. *Comp. Biochem. Physiol. D.,* **15**: 12-19. https://doi.org/10.1016/j.cbd.2015.04.002

Navakanitworakul, R., Deachamag, P., Wonglapsuwan, M. and Chotigeat, W., 2012. The roles of ribosomal protein S3a in ovarian development of *Fenneropenaeus merguiensis* (De Man). *Aquaculture,* **338**: 208-215. https://doi.org/10.1016/j.aquaculture.2012.01.024

Powell, D., Knibb, W., Remilton, C. and Elizur, A., 2015. De-novo transcriptome analysis of the banana shrimp (*Fenneropenaeus merguiensis)* and identification of genes associated with reproduction and development. *Mar. Genom.,* **22**: 71-78. https://doi.org/10.1016/j.margen.2015.04.006

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M., 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics,* **31**: 3210-3212. https://doi.org/10.1093/bioinformatics/btv351

Saoud, I.P., Davis, D.A. and Rouse, D.B., 2003. Suitability studies of inland well waters for *Litopenaeus vannamei* culture. *Aquaculture,* **217**: 373-383. https://doi.org/10.1016/S0044-8486(02)00418-0

Santos, C.A., Blanck, D.V. and de Freitas, P.D., 2014. RNA-seq as a powerful tool for penaeid shrimp genetic progress. *Front. Genet.,* **5**: 298. https://doi.org/10.3389/fgene.2014.00298

Smith, S., Bernatchez, L. and Beheregaray, L.B., 2013. RNA-seq analysis reveals extensive transcriptional plasticity to temperature stress in a freshwater fish species. *BMC Genom.,* **14**: 375-384. https://doi.org/10.1186/1471-2164-14-375

Stillman, J.H. and Tagmount, A., 2009. Seasonal and latitudinal acclimatization of cardiac transcriptome responses to thermal stress in porcelain crabs, *Petrolisthes cinctipes*. *Mol. Ecol.,* **18**: 4206-4226. https://doi.org/10.1111/j.1365-294X.2009.04354.x

Theissinger, K., Falckenhayn, C., Blande, D., Toljamoc, A., Gutekunstb, J., Makkonenc, J., Jussilac, J., Lykob, F., Schrimpfa, A., Schulza, R. and Kokkoc, H., 2016. De novo assembly and annotation of the freshwater crayfish *Astacus astacus* transcriptome. *Mar. Genom.,* **28**: 7-10. https://doi.org/10.1016/j.margen.2016.02.006

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., Sridhar Rao, B., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J. and Natale D.A., 2003. The COG database: An updated version includes eukaryotes. *BMC Bioinformatics*, **4**: 41. https://doi.org/10.1186/1471-2105-4-41

Thiel, T., Michalek, W., Varshney, R. and Graner, A., 2003. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor. appl. Genet.,* **106**: 411-422. https://doi.org/10.1007/s00122-002-1031-0

Wang, C.Z., Lin, G.R., Yan, T., Zheng, Z.P., Chen, B. and Sun, F.L., 2014. The cellular community in the intestine of the shrimp *Penaeus penicillatus* and its culture environments. *Fish. Sci.,* **80**: 1001-1007. https://doi.org/10.1007/s12562-014-0765-3

Xia, J.H., Liu, P., Liu, F., Lin, G., Sun, F., Tu, R. and Yue, G.H., 2013. Analysis of stress-responsive transcriptome in the intestine of Asian seabass (*Lates calcarifer*) using RNA-seq. *DNA Res.,* **20**: 449-460. https://doi.org/10.1093/dnares/dst022

Xu, J., Ji, P., Wang, B., Zhao, L., Wang, J., Zhao, Z.X., Zhang, Y., Li, J.T., Xu, P. and Sun, X.W., 2013. Transcriptome sequencing and analysis of wild Amur Ide (*Leuciscus waleckii*) inhabiting an extreme alkaline-saline lake reveals insights into stress adaptation. *PLoS One,* **8**: e59703. https://doi.org/10.1371/journal.pone.0059703

Zhan, W., Chen, R.Y., Laghari, M.Y., Xu, D.D., Mao, G.M., Shi, H.L. and Lou, B., 2016. Characterization of *Nibea albiflora* Transcriptome: Sequencing, De Novo assembly, annotation and comparative genomics. *Pakistan J. Zool.,* **48**: 427-434.

Zhang, G.L., Li, Z.B., Wang, Z.L., Lin, X.Y. and Wu, N., 2010. Study status and perspective of *Fenneropenaeus penicillatus*. *Mod. Fish. Inform.,* **2**: 7-10.