Aquaculture 464 (2016) 545–553

Contents lists available at ScienceDirect

Aquaculture

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

De novo assembly and transcriptome characterization of major growth-related genes in various tissues of *Penaeus monodon*

Cuong Nguyen [a,*,1], Thu Giang Nguyen [b,1], Lam Van Nguyen [a], Huy Quang Pham [a], Trieu Hai Nguyen [a], Hoa Thi Pham [a], Hoa Thi Nguyen [a], Thu Thi Ha [a], Tung Huy Dau [a], Hien Thi Vu [a], Duy Dinh Nguyen [a], Nhung Tuyet Thi Nguyen [a], Ninh Huu Nguyen [c], Dong Van Quyen [a], Ha Hoang Chu [a], Khang Duy Dinh [a,*]

[a] Institute of Biotechnology, Vietnam Academy of Science and Technology, 18 Hoang Quoc Viet, Cau Giay, Hanoi, Vietnam
[b] Science Technology and Environmental Department, Ministry of Agriculture and Rural Development, 1st floor, A9 Building, 2 Ngoc Ha, Ba Dinh, Hanoi, Vietnam
[c] Research Institute for Aquaculture No 1, Dinh Bang, Tu Son, Bac Ninh, Vietnam

ARTICLE INFO

Article history:
Received 2 November 2015
Received in revised form 7 July 2016
Accepted 1 August 2016
Available online 2 August 2016

Keywords:
De novo assembly
Transcriptome analysis
Growth-related genes
Next-generation sequencing
Penaeus monodon

ABSTRACT

Black tiger shrimp (*Penaeus monodon*) is an economically important species for aquaculture in Vietnam; however, very little information is available regarding its transcriptome. In this study, we sequenced four transcriptome libraries from mRNA of heart, muscle, hepatopancreas, and eyestalk tissues of *P. monodon* individuals using the Illumina MiSeq® platform, yielding 78,575,346 raw paired-end reads. After quality control, a total of 62,327,726 paired-end reads were *de novo* assembled into 69,089 unigenes with an average size of 447.89 bp, N50 of 481 bp, and transcriptome size of 30.94 Mb. Comparisons of the assembled unigenes against five public protein databases identified 165 unigenes related to growth and development, and expression analysis revealed that 14,229 unigenes were differentially expressed among tissues. In addition, we detected 12,768 simple sequence repeats (SSRs) in 11,173 unigenes. This work provides candidate genes for trait mapping, marker-assisted breeding, genetic studies and functional genome annotation of *P. monodon*.
Statement of relevance: This paper is relevant for the Shrimp Industry.

1. Introduction

Penaeus monodon (Crustacea: Decapoda: Dendrobranchiata), commonly known as black tiger shrimp, is locally cultured in Vietnam and plays a major role in the Vietnam aquaculture industry. According to a Vietnamese government report, domestic production of *P. monodon* reached 260,000 tons in 2014.[2] Therefore, development of more comprehensive genome-wide or transcriptome-wide datasets would provide insights into biological processes in this species and help to improve its performance in aquaculture.

Because of the great economic value of *P. monodon* and the importance of biological information regarding this species, previous studies have mostly focused on developing genetic linkage maps based on microsatellite, single-nucleotide polymorphism (SNP) and amplified fragment length polymorphism (AFLP) markers (Preechaphol et al., 2010; You et al., 2010); Sanger-based sequencing of expressed sequence tags (ESTs) (Karoonuthaisiri et al., 2009; Lehnert et al., 1999; Preechaphol et al., 2007); characterization of single functional genes (Brady et al., 2013; de la Vega et al., 2007; James et al., 2010); and sequencing of the mitochondrial genome (Wilson et al., 2000).

The relatively low cost and high accuracy of next-generation sequencing (NGS) technology enables rapid and economical sequencing. Analysis of the transcriptome by RNA-seq is a powerful and accurate tool for quantifying gene expression levels and is widely regarded as superior to previous transcriptomic methods such as EST sequencing, serial analysis of gene expression (SAGE), massively parallel signature sequencing (MPSS), and microarrays (Mortazavi et al., 2008; Wang et al., 2009). Therefore, as an attractive alternative to whole-genome analysis, especially in non-model organisms, transcriptome analysis can be used to detect and quantify new splice isoforms and low-abundance transcripts, determine transcriptional boundaries of genes at single-nucleotide resolution, identify expressed SNPs, investigate the functional factors of the genome, and expose the expression mechanisms of tissue-specific genes (Vera et al., 2008). Over the past 5 years, RNA-seq has been used to make significant progress in identification of new genes and simple sequence repeat (SSR) and SNP markers, as well as analysis of differentially expressed genes, in marine crustaceans such as *Litopenaeus vannamei* (Guo et al., 2013; Sookruksawong et al., 2013), *Fenneropenaeus chinensis* (Li et al., 2013), *Eriocheir sinensis* (Erchao Li, 2014; Hui et al., 2014; Liu et al., 2015), and *Macrobrachium nipponense* (Ma et al., 2012; Sun et al., 2015). The resulting data are

* Corresponding authors.
E-mail addresses: cuongnguyen@ibt.ac.vn (C. Nguyen), khangvspt@ibt.ac.vn (K.D. Dinh).
[1] These authors contributed equally to this work.
[2] http://www.fistenet.gov.vn/thong-tin-huu-ich/thong-tin-thong-ke/thong-ke-1/tinh-hinh-san-xuat-thuy-san-nam-2014.

http://dx.doi.org/10.1016/j.aquaculture.2016.08.003
0044-8486/

Table 1
Primers used for qPCR.

Annotation	Sequence_ID	Primer	Sequence
Beta-actin	comp122440_c1_seq1	Forward	5′-GCTTGCTGATCCACATCTGCT -3′
		Reverse	5′- ACTACCATCGGCAACGAGA -3′
Myosin-heavy chain	comp123891_c2_seq2	Forward	5′- GTGTCCTACAACCTGACTGG -3′
		Reverse	5′- TCCCTTTCCTTTGCCACCAC-3′
Ecdysteroid	comp93369_c1_seq1	Forward	5′- CTCCACACCTTAGACAGACC -3′
		Reverse	5′- GGAGGTCTACGTGCTAAAGG -3′
Cathepsin L	comp123524_c1_seq2	Forward	5′- AGCAGTGGTCGTCATGGTAC -3′
		Reverse	5′- GGACGGTAAGTGTCGCTTCG-3′
Alpha-amylase	comp200235_c1_seq1	Forward	5′- CGTCCGACAGATCACAGTTTGG-3′
		Reverse	5′- GGATCTGAAGGAGACGCTGC-3′

crucial for improving the molecular information available for each species and constructing optimal strategies to improve productivity in culture. Many molecular biology studies of *P. monodon* that had been published but utilizing low-throughput sequencing (Sanger sequencing) or only characterization of single functional genes. Currently, there was only one study of Baranski et al. (2014) in *P. monodon* that focused on development of a genetic linkage map which used in the genotyping study by using next generation sequencing data, but did not incorporate expression analysis or identification of growth-related genes. Therefore, the demand of other studies in biological information regarding *P. monodon* in high throughput sequencing data will be important.

In this study, we used the Illumina MiSeq® platform to *de novo* assemble a *P. monodon* transcriptome dataset using mRNA samples from heart, muscle, hepatopancreas and eyestalk tissues. The assembled unigenes were annotated against public protein databases followed by NR-NCBI, Swiss-Prot, GO, COG, and KEGG classification and investigated to identify growth-related genes. To investigate expression profiles, we identified genes that were differentially expressed in the four tissues. In addition, we identified SSR markers. This resource and the associated findings will contribute to studies of functional genomics and biogeography of *P. monodon.*

2. Methodology

2.1. Sample preparation, RNA isolation, cDNA library construction, and sequencing

Samples were collected in Nghe An province, Vietnam. The sex of each specimen was determined based on the presence of petasma in males and thelycum in females, using existing identification keys (Wakida-Kusunoki et al., 2013). Screening all samples with primers for Infectious Hypodermal and Hematopoietic Necrosis Virus (IHHNV), Taura, Monodon Baculovirus (MBV), White spot syndrome virus (WSSV), and Yellow Head virus was necessary to establish a reference database for *P. monodon*. Total RNA from muscle, eyestalk, hepatopancreas, and heart tissues (0.5–0.7 g per sample) was extracted using the Trizol reagent (Invitrogen Dynal AS, Oslo, Norway) as instructed by the manufacturer.

Total RNA concentration, quality, and integrity were determined using a NanoDrop Lite spectrophotometer and checked by real-time PCR. Purification of mRNA was conducted using the Dynabeads® mRNA DIRECT™ Micro Kit (Invitrogen Dynal AS, Oslo, Norway) according to the manufacturer's protocols and quantified using a NanoDrop spectrophotometer (Thermo Fisher Scientific Inc., Waltham MA, USA) and an Agilent 2100 Bioanalyzer with an Agilent RNA Pico Chip Kit (Agilent Technologies, Inc., Santa Clara, California, USA). The purified mRNA was submitted to Biomedic Inc. (Hanoi, Vietnam) for NGS transcriptome sequencing on the Illumina MiSeq® platform (Illumina, San Diego, CA, USA).

2.2. Data pre-processing and de novo transcriptome assembly

Raw paired-end reads were assessed and subjected to quality control using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and Trimmomatic (Bolger et al., 2014) (parameters: ILLUMINACLIP:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:70) to obtain a set of clean paired-end reads. After pre-processing, FastQC was used again to report features of the pre-processing libraries and verify the effectiveness of trimming.

Pre-processing reads from heart, muscle, hepatopancreas, and eyestalk tissues were *de novo* assembled using Trinity version trinityrnaseq_r20140717 (Haas et al., 2013) with default parameters (*k-mer length* = 25).

In order to filter out transcriptional artifacts, misassembled transcripts, and poorly supported transcripts, the original trimmed paired-end reads were mapped to assembled transcripts using RSEM 1.2.19 and the Trinity script *align_and_estimate_abundance.pl* (http://trinityrnaseq.github.io/). Following this script, RNA-Seq by Expectation Maximization (RSEM) read abundance values were calculated on a per-isoform and per-gene basis. In addition, the percent distribution of each transcript component was calculated for each gene. From these results, the original assembly file produced by Trinity was filtered using the script *filter_fasta_by_rsem_values.pl* to remove transcripts that represented <1% of the RSEM-based expression level of their parent genes or transcripts with Fragments Per Kilobase of exon Per Million fragments mapped (FPKM) values below 1. To address the issue of highly similar/redundant contigs or transcripts, custom Perl scripts were

Table 2
Quality filtering of Illumina MiSeq data before *de novo* assembly.

Tissues	Parameters	Before treatment	After treatment	Remaining percentage
Heart	Untreated paired-end reads	22,531,716	18,113,880	80.39%
	Length of reads	35–200	70–200	
Skeletal	Untreated paired-end reads	12,312,819	8,533,944	69.31%
	Length of reads	35–251	70–251	
Hepatic pancreas	Untreated paired-end reads	20,512,979	17,964,211	87.57%
	Length of reads	35–151	70–151	
Eyestalk	Untreated paired-end reads	23,217,832	17,715,691	76.30%
	Length of reads	35–151	70–151	
High-quality paired-end reads (total)	62,327,726 (79.32%)			