Marine Genomics xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Marine Genomics

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

De novo assembly and annotation of the whole transcriptome of *Sepiella maindroni*

Kuo Tian [a], Fangrui Lou [a,b], Tianxiang Gao [a], Yongdong Zhou [a], Zhenqing Miao [a], Zhiqiang Han [a,*]

[a] Fishery College, Zhejiang Ocean University, Zhoushan 316022, China
[b] Fishery College, Ocean University of China, Qingdao 266003, China

ARTICLE INFO

Article history:
Received 10 April 2017
Received in revised form 17 May 2017
Accepted 16 June 2017
Available online xxxx

Keywords:
Sepiella maindroni
Transcriptome
RNA-seq technology
De novo assembly
Gene annotation

ABSTRACT

As an important cephalopods species, *Sepiella maindroni* fishery has suffered a severe decline due to over-fishing since the 1980s. Stock enhancement has been applied to the resource recovery of this species, but a sexual precocity appeared in breeding process. In order to understand the regulatory mechanism of this phenomenon, we generated the whole transcriptome of *S. maindroni* based on the total RNA of tissue samples (eyestalk, peduncle, tentacle, gill, muscle and ovary) using Illumina RNA-seq technology. *De novo* assembly was performed using Trinity software and a total of 31,979,244 high-quality clean reads were randomly assembled to produce 74,245 contigs. All contigs were further assembled and clustered into 58,224 unigenes. Among the predictable unigenes, a total of 14,346 unigenes were annotated based on protein databases. We assessment the annotation completeness using BUSCO software package and the result showed that 91.5% protein-coding genes were found in our assembled transcripts. At last, we predicted the structure of all unigenes using TransDecoder software and MIcroSAtellite identification tool, respectively. Result showed that a total of 26,037 nucleotide sequences of coding regions (direction of the sequences is 5′ → 3′) were confirmed to the protein database and a total of 13,471 simple sequence repeats were identified. Our goal was to provide an important foundation for future genomic research on the cephalopod and further evaluate the effectiveness of stock enhancement.

1. Introduction

As an important cephalopods species, *Sepiella maindroni* was once one of four major cephalopod species fished in East China Sea and Yellow Sea (Zheng et al., 2001; Zheng et al., 2003). However, fishery of *S. maindroni* has suffered a severe decline due to over-fishing since the 1980s. Stock enhancement has been applied to the resource recovery of *S. maindroni* in recent years and obtained a better result (Wu et al., 2006). However, sexual precocity appeared in breeding process and numerous literatures on this phenomenon were focused on physiological ecology of *S. maindroni* (Zheng et al., 2010; Ping, 2015). It's very important to detect the regulatory mechanism of sexual precocity at genetic level. Nevertheless, limited information is available to the gene expression of *S. maindroni*, which is particular importance for the study of the characteristics of sexual precocity.

High-throughput RNA-seq has widely used in the biology research and provides huge potential to capture a global gene expression profile for marine organisms (Dittel and Epifanio, 2009; Smith et al., 2013; Xia et al., 2013; Xu et al., 2013; Stillman and Tagmount, 2009). A large number of studies in non-model organisms on the basic of RNA-seq technology have discovered the primary regulatory mechanism at the transcription level (Theissinger et al., 2016; Wen et al., 2016; Yang et al., 2015). However, limited genetic data set is available for *S. maindroni*. In the present study, RNA-seq technology was used to capture a significant portion of whole transcriptomes of *S. maindroni*. This study presented a comprehensive analysis and a general view of the potential molecular mechanisms based on construction of an annotated whole transcriptome library by *de novo* assembly of raw reads generated from high-throughput technology. The present study aimed to pave the way to development of a more complete understanding of the complex gene related to the regulatory mechanism of *S. maindroni*. This study will also give provide basic data for understanding the regulatory mechanism of sexual precocity in this species.

2. Data description

2.1. Sampling and sequencing

One female of *S. maindroni* were collected from the aquaculture farm of Zhoushan (China) on November 12th, 2016. MIxS descriptors are presented in Table 1. Tissue samples of eyestalk, peduncle, tentacle, gill, muscle and ovary were rapidly sampled, snap-frozen in liquid nitrogen and stored at −80 °C prior to the RNA extraction.

* Corresponding author.
E-mail address: d6339124@163.com (Z. Han).

http://dx.doi.org/10.1016/j.margen.2017.06.004
1874-7787/

Table 1
MIxS descriptors.

Item	Description
Investigation_type	Eukaryote
Project_name	PRJNA379668
Lat_lon	29°52′N, 122°30′E
Geo_loc_name	Zhoushan, China
Collection_data	2016-11-12
Environment	Marine water
Biotic_relationship	Free living
Trophic_level	Heterotroph
Sequencing_meth	IlluminaHiSeq™ 2000
Assembly	Trinity software package
Annot_source	NR/Swiss-Prot/GO/COG/EggNOG4.5/ KOG/Pfam/KEGG
Estimated_size	4.7 billion base-pairs
Biome	ENVO:01000048
Feature	ENVO:02000049
Material	ENVO:00002150
Temp	18 °C
Salinity	28 PSU
Accession number of raw reads	SRR5358819
Accession number of transcripts	GFLT00000000

2.2. RNA extraction and Illumina sequencing

The total RNA of individual tissue was isolated and pooled in equal amounts using a standard Trizol Reagent Kit according to the manufacturer's protocol.

A total amount of 3 μg RNA per sample was used as input material for the RNA sample preparations. Sequencing libraries with cDNA fragments of 150–200 bp in length were generated and index codes were added to attribute sequences to each sample. RNA-seq library quality was assessed on the Agilent Bioanalyzer 2100 System. The clustering of adaptor-ligated cDNA and flow cell (every flow cell have 8 lane) was performed according to the manufacturer's instructions. Subsequently, the cDNA library was sequenced using Illumina HiSeq 2000 platform and paired-end reads were generated.

2.3. Transcriptome de novo assembly

The raw reads in FASTQ format with sequencing adaptors, unknown nucleotides (N ratio > 10%) and low quality (quality scores ≤ 5) were removed according to FastQC software. The Q30 percentage (percentage of bases whose quality was >30 in clean reads), N percentage (percentage of uncertain bases after filtering), and GC percentage were 94.52%, 0.00% and 41.49%, respectively. Then, the remaining high-quality reads were *de novo* assembled using Trinity software package (version 2.0.6) with min_kmer_cov set to 2 by default and all other parameters set default. All the redundancy sequences were removed using Tgicl software package and further spliced the longest unigenes. For the *S. maindroni*, a total of 31,979,244 high-quality clean reads were randomly assembled to produce 74,245 contigs with an N50 of 1575 nt, corresponding to 68,245,420 nucleotides. The contigs were further assembled and clustered into 58,224 unigenes with an N50 of 1269 nt, corresponding to 45,553,051 nucleotides. We compared the clean reads with the unigenes and obtained 22,591,518 mapped reads.

2.4. Gene ontology and KEGG pathway analysis

After transcriptome *de novo* assembly for the clean reads, all unigenes were applied to the gene function annotation and the putative functions were analyzed based on protein databases. Among the predictable unigenes, a total of 14,346 unigenes were annotated using the Blast alignment (*E*-value <0.00001). Of all annotated unigenes, 13,729, 7,852, 4,831, 4,562, 12,695, 9,348, 10,700 and 6,852 unigenes had significant matches with sequences in the NR, Swiss-Prot, GO, COG, EggNOG4.5, KOG, Pfam and KEGG databases, respectively.

We acquired the homology searches by comparing all unigenes against the NR protein database and 13,729 unigenes had positive hits (*E*-value <0.00001) in the database (Fig. 1). Among these unigenes, most unigenes were related to those genes of the *Lottia gigantea* (3,174, 23.14%), *Crassostrea gigas* (2,996, 21.84%) and *Aplysia californica* (1,843, 13.43%). Subsequently, 682 (4.97%), 511 (3.72%), 461 (3.36%), 313 (2.28%), 259 (1.89%), 220 (1.60%) and 201 (1.47%) unigenes were matched with genes from *Capitella teleta*, *Saccoglossus kowalevskii*, *Branchiostoma floridae*, *Strongylocentrotus purpuratus*, *Stegodyphus mimosarum*, *Acyrthosiphon pisum* and *Hydra vulgaris*, respectively. The remaining unigenes (3,059, 22.30%) were hits in the other species.

In addition, the gene ontology annotation and cluster of all unigenes were analyzed based on gene ontology and orthologous classifications (Young et al., 2010). In the present study, a GO, KOG, EggNOG and COG analysis was carried out using Blast2GO software and the threshold parameter of significant difference is $Q < 0.05$. Briefly, a total of 4,831 unigenes were successfully mapped to existing gene categories and all unigenes were categorized into 52 functional groups (Fig. 2). Among these functional groups, the terms of “cell”, “catalytic activity” and “metabolic process” were dominant in “cellular component”, “molecular function” and “biological process”, respectively.

COG classification was delineated by comparing protein sequences encoded in the complete genome and used to further evaluate the major phylogenetic lineages (Fig. S1.A). In this study, a total of 4562 unigenes were mapped 25 different COG categories. Among these categories, the largest COG group was the R category, representing “General function prediction only”, followed by the L category, K category, G category, which representing “Replication, recombination and repair”, “Transcription” and “Translation, ribosomal structure and biogenesis”, respectively. KOG classification was used to analysis and 9348 unigenes were mapped 25 different KOG categories (Fig. S1.B). Among these categories, the first three largest groups were R category, T category and O category, which representing “General function prediction only”, “Signal transduction mechanisms” and “Posttranslational modification, protein turnover, chaperones”, respectively. EggNOG database was a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences, its covers a range of protein sequences far beyond the COG and KOG databases (Fig. S1.C). As a result, a total of 12,695 unigenes were mapped 25 different EggNOG categories and a large number unigenes were dominant in three terms, including R category (2,377 unigenes), S category (2,185 unigenes) and T category (1,455 unigenes), which representing “General function prediction only”, “Function unknown” and “Signal transduction mechanisms”, respectively.

Then, KEGG database analysis was used to estimate the biochemical metabolic pathways and function of gene product. In the present study, all unigenes were searched in KEGG database and a total of 6,582 unigenes were assigned to 272 KEGG pathways. Result showed that annotated sequences were involved in metabolic pathways, immune regulatory pathway, cellular processes pathway and genetic information processing pathway.

In order to further quantitative assessment of the assembly and annotation completeness, we applied the BUSCO software package with default setting and downloaded from Ensembl Metazoa as a reference. The result showed that 91.5% protein-coding genes were found in our assembled transcripts, the proportions of mismatching genes were calculated and only 5.1%. The BUSCO analysis shows that this transcriptome assembly will provide a good basis for further transcriptomic research of the *S. maindroni*.

2.5. Predict the unigene structure

We predicted the structure of all unigenes in the present study, and all unigenes were analyzed using TransDecoder software and MIcroSAtellite

Please cite this article as: Tian, K., et al., *De novo* assembly and annotation of the whole transcriptome of *Sepiella maindroni*, Mar. Genomics (2017), http://dx.doi.org/10.1016/j.margen.2017.06.004