Transcriptomics reveals tissue/organ-specific differences in gene expression in the starfish

*Patiria pectinifera*

Chan-Hee Kim\(^a\), Hye-Jin Go\(^a\), Hye Young Oh\(^a\), Yong Hun Jo\(^b\), Maurice R. Elphick\(^c\), and Nam Gyu Park\(^a,\dagger\)

\(^a\)Department of Biotechnology, College of Fisheries Sciences, Pukyong National University, 45 Yongso-ro, Nam-gu, Busan 48513, Korea

\(^b\)Division of Plant Biotechnology, Institute of Environmentally-Friendly Agriculture (IEFA), College of Agriculture and Life Sciences, Chonnam National University, Gwangju 500-757, Korea

\(^c\)School of Biological and Chemical Sciences, Queen Mary University of London, London, E1 4NS UK

\(^\dagger\)Corresponding author: Nam Gyu Park, Department of Biotechnology, College of Fisheries Sciences, Pukyong National University, 45 Yongso-ro, Nam-gu, Busan 48513, Korea, Telephone: +82 51 6295867, Fax: +82 51 6295863, e-mail: ngpark@pknu.ac.kr
Abstract

Starfish (Phylum Echinodermata) are of interest from an evolutionary perspective because as deuterostomian invertebrates they occupy an “intermediate” phylogenetic position with respect to chordates (e.g. vertebrates) and protostomian invertebrates (e.g. Drosophila). Furthermore, starfish are model organisms for research on fertilization, embryonic development, innate immunity and tissue regeneration. However, large-scale molecular data for starfish tissues/organs are limited. To provide a comprehensive genetic resource for the starfish Patiria pectinifera, we report de novo transcriptome assemblies and global gene expression analysis for six P. pectinifera tissues/organs – body wall (BW), coelomic epithelium (CE), tube feet (TF), stomach (SM), pyloric caeca (PC) and gonad (GN). A total of 408 million high-quality reads obtained from six cDNA libraries were assembled de novo using Trinity, resulting in a total of 549,625 contigs with a mean length of 835 nucleotides (nt), an N50 of 1,473 nt, and GC ratio of 42.52%. A total of 126,136 contigs (22.9%) were obtained as predicted open reading frames (ORFs) by TransDecoder, of which 102,187 were annotated with NCBI non-redundant (NR) hits, and 51,075 and 10,963 were annotated with Gene Ontology (GO) and Kyoto Encyclopaedia of Genes and Genomes (KEGG) using the Blast2GO program, respectively.

Gene expression analysis revealed that tissues/organs are grouped into three clusters: BW/CE/TF, SM/PC, and GN, which likely reflect functional relationships. 2,408, 1,727, 2,667, 3,321, 2,687, and 8,560 specifically expressed genes were identified for BW, CE, TF, SM, PC and GN, respectively, using the ROKU method. This study provides a valuable transcriptome resource and novel molecular insights into the functional biology of different tissues/organs in starfish as a model organism.

Keywords: transcriptome, multiple tissues, differential expressed genes, echinoderm, starfish, Patiria pectinifera
**Introduction**

Starfish are deuterostomian invertebrates belonging to the phylum Echinodermata that are recognized as fascinating animals with many features of special interest (Arnone et al., 2016). These include their status as a canonical example of a keystone species in ecology (Paine, 1966) and as model organisms for research on neuroendocrinology (Semmens et al., 2016), innate immunity (Franco et al., 2011) and tissue regeneration (Thorndyke et al., 2001).

Recently, transcriptome sequencing of emerging model marine organisms has proven to be an efficient method for relatively low cost gene discovery and analysis of differential gene expression (Martin and Wang, 2011). For these reasons, several research groups have utilized high throughput sequencing technology for molecular level characterization of various biological processes in several starfish species, including *Asterias amurensis* (Richardson and Sherman, 2015), *Asterias rubens* (Semmens et al., 2013; Semmens et al., 2016), *Acanthaster planci* (Stewart et al., 2015) and *Coscinasterias muricata* (Gabre et al., 2015). However, these transcriptomic studies were limited to analysis of whole larvae or just one adult starfish tissue/organ. Starfish have many different tissues and organs that are responsible for a variety of biological processes. For example: 1. the mechanical state of the body-wall determines body stiffness and posture in starfish (Motokawa, 2011), 2. the coelomic epithelium is involved in wound healing, regeneration, and hematopoiesis (Gabre et al., 2015), 3. the stomach (pyloric and cardiac stomach) and pyloric caeca enable intake, digestion, absorption and storage of nutrients (Ferguson, 1964), 4. a multitude of tube feet enable locomotion, utilising secretion of adhesive materials for adhesion (Hennebert et al., 2011), and 5. gonads are, of course, essential for reproduction (Stewart et al., 2015). It is not known, however, if the physiological roles of different tissues/organs in starfish are reflected in their gene expression profiles. Therefore, it is of interest to comprehensively analyze differentially expressed genes (DEGs) in starfish tissues/organs.

The starfish species *Patiria pectinifera* is widely distributed in the northern Pacific Ocean and has been used as a model organism for studying different aspects of starfish physiology and is also of interest from economic and environmental perspectives (Kim et al., 2016; Mita et al., 2009). In this study, we characterise the *P. pectinifera* transcriptome by RNA-seq employing paired-end Illumina HiSeq™ 2500 sequencing technology and subsequent de novo assembly to generate a comprehensive set of reference contigs for gene discovery and for analysis of DEGs among six different tissues/organs. Our study provides a genetic resource for future comparisons with other echinoderm transcriptomes and for functional analysis of gene expression.
2. Data description

2.1. RNA isolation and illumina sequencing

Live specimens of the starfish species *Patiria pectinifera* (approximate diameter 8 cm) were collected at low tide from the coast of Cheongsapo of Busan, Korea (Table 1). Approval by the local institution/ethics committee was not required for this work because experimental work on starfish is not subject to regulation and *P. pectinifera* is not an endangered or protected species. Six tissues/organs, gonad (GN), pyloric caeca (PC), coelomic epithelium (CE), stomach (SM), tube feet (TF), and body-wall (BW, excluding CE) were dissected from six individual specimens of *P. pectinifera* and then total RNA was extracted using RNeasy Total RNA Isolation kit (Qiagen, USA) according to the manufacturer’s instructions. The concentration, quality, and integrity of RNA preparations were determined using a NanoDrop-2000 spectrophotometer (Thermo, USA) and a Bioanalyzer 2100 (Agilent Technologies, USA). Then the RNA preparations were disrupted into short fragments. Double-stranded cDNA was synthesized with sequencing adapters using Illumina TruSeq™ RNA Library Prep Kit v2 (San Diego, CA, USA) following the manufacturer’s instructions. Finally, six RNA-seq libraries were subjected to paired-end sequencing with a read length of 2×101 nucleotides on an Illumina HiSeq 2500 platform. Illumina HiSeq 2500 produced a total of 417,972,264 reads representing a total of 42,215,198,664 nucleotides from six tissues/organs, with the maximum number of reads (75,840,222) generated from the SM library and the minimum number of reads (63,360,640) generated from CE library (Table 1). The raw reads were deposited in the Sequencing Read Archive (SRA) of NCBI with accession numbers SRR5229423, SRR5229424, SRR5229425, SRR5229426, SRR5229427, SRR5229428 for TF, SM, CE, PC, GN and BW, respectively.

2.2. De novo assembly and functional annotation

After sequencing was completed, Illumina TruSeq adapter sequences, low-quality bases, and reads under minimum length were trimmed from the reads using CutAdapt v1.10 with -q 20, -m 30 parameters (Martin, 2011). Then, reads were filtered into clean reads. From these reads, contamination removal was performed by Bowtie2 v2.2.9 against the bacterial and ocean metagenome databases downloaded from NCBI ([ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria](ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria), [ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria_DRAFT](ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria_DRAFT), [ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/204/965/GCA_000204965.1_ASM20496v1](ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/204/965/GCA_000204965.1_ASM20496v1)). A total of 408,167,516 reads (97.66%) were obtained and were subjected to de novo assembly. The assembly produced a total of 549,625 contigs of a total length of 459,072,158 nt, with a mean length of 835 nt, an N50 of 1,473 nt, and GC ratio of 42.52% in a size range of 224 to 35,675 nt, using
Trinity Assembler (v2.0.6) (Grabherr et al., 2011), which is optimized to present the best *de novo* results compared to a variety of assemblies with paired-end reads. This Transcriptome Shotgun Assembly (TSA) project has been deposited at DDBJ/EMBL/GenBank under the accession GFOQ00000000. The version described in this paper is the first version, GFOQ01000000. Subsequently, the output from the assembler was processed to predict open reading frames (ORFs) using TransDecoder v3.0.0, generating a total of 126,136 ORFs with a total length of 145,293,354 nt, a mean length of 1,151 nt, an N50 of 1,569 nt, and GC ratio of 51.272% in a size range of 297 to 33,894 nt. The predicted ORFs were used in subsequent stages of annotation and analysis, and are available in the supplementary material (Appendix A).

To identify the function of predicted genes, all the predicted ORFs were aligned against the NCBI non-redundant (NR) databases using a cut off e-value of $< 10^{-3}$, returning 102,187 hits (supplementary material 1). The top 10 species with hits are shown in Fig. 1A. Most of the ORFs aligned to the genome sequence of the sea urchin *Strongylocentrotus purpuratus* (48.08%), followed by the acorn worm *Saccoglossus kowalevskii* (15.91%), and the lancelet *Branchiostoma floridae* (7.302%). This result is as expected because as an echinoderm *P. pectinifera* is a deuterostomian invertebrate, a superphylum that includes two chordate subphyla that are closely related to vertebrates (Urochordata and Cephalochordata) and the Ambulacraria (Hemichordata and Echinodermata) (Adoutte et al., 2000). To further investigate the biological functions of the predicted ORFs they were annotated to the Gene Ontology (GO) database for biological process (BP), cellular component (CC), and molecular function (MF) and to the Kyoto Encyclopedia of Genes and Genomes (KEGG) database for enzyme commission (EC) number assignment via mapping to the KEGG pathway map, using Blast2Go program (Conesa et al., 2005). GO classification revealed that 51,075 (40.49%) annotated ORFs were categorized into 30 functional groups, and 10,963 (8.69%) transcripts had significant matches in the KEGG pathways (Fig. 1B and C and supplementary material 1).

### 2.3. Differentially expressed genes (DEGs) analysis

The expression values of all samples were estimated to a fragments per kilobase per million mapped reads (FPKM) value by RSEM (Li and Dewey, 2011) and were further normalized by the quantile normalization method with the R-package preprocessCore. A total of 91,124 transcripts, approximately three-fourths of the predicted ORFs in the transcriptome of *P. pectinifera*, were expressed in at least one tissue/organ based on the following criteria (FPKM $\geq$ 0.3 and Expected read count $\geq$ 5), and 23,412 of the expressed transcripts were present in all tissues/organs (supplementary material 1). The tissue-specific transcripts (i.e. significantly higher expression in a single tissue relative to all tissues) were defined using the ROKU method based on the Shannon entropy and
outlier detection to scan expression profiles for ranking tissue-specific genes using Akaike’s information criterion (AIC) procedure (Kadota et al., 2006). The upper-limit option specifying the maximum 25 percentage of tissue transcripts as outliers to each gene was set as 0.25. A total of 21,370 tissue/organ-specific transcripts were detected as follows: BW (2,408), GN (8,560), PC (2,687), CE (1,727), SM (3,321), and TF (2,667) (Fig. 2A and supplementary material 2). Differentially expressed genes (DEGs) were identified by pairwise comparisons between two tissues/organs using the edgeR program (Robinson et al., 2010). For the non-replicated samples, we set the biological coefficient of variation (BCV) parameter as 0.4 and the parameters for DEGs were set as a log fold change (logFC ≥ 4) and a false discovery rate (FDR ≤ 0.01). The highest number of DEGs (27,840) was detected in the GN vs. PC comparison, representing 17,014 of up-regulated genes in GN and 10,826 of up-regulated genes in PC, and the lowest number of DEGs (12,385) was detected in the BW vs. CE comparison, representing 6,010 of upregulated genes in BW and 6,375 of up-regulated genes in CE. The DEGs between different tissue/organ samples are summarized in Fig. 2B and details are shown in supplementary material 3. To assess transcriptome similarity between tissues/organs, principal component analysis (PCA) and hierarchical clustering were performed (Fig. 2C, D and E). Both analyses revealed three discrete groupings reflecting their biological functions: the body wall (BW) and its associated tissues/organs (CE and TF) constituting epidermis, organs involved in feeding and digestion (PC and SM), and an organ (GN) involved in reproductive functions formed a separate, stand-alone cluster.

2.4. Conclusions

Although several studies have reported transcriptomic analysis of starfish species, there have been no studies that investigated differentially expressed genes in different starfish tissues/organs. In this study, RNA-seq employing paired-end Illumina HiSeq™ 2500 sequencing technology was used to generate transcriptome data and profile global gene expression in six tissues/organs from the starfish *P. pectinifera*. Our results revealed that numerous genes are differentially expressed in the six tissues/organs, which group together into three clusters in accordance with biological functions. These data provide a basis for identification of gene networks associated with specific biological functions in adult starfish tissues/organs. Therefore, the *P. pectinifera* transcriptome data obtained in this study provide a valuable resource for future research on many aspects of starfish biology, including gene/protein evolution, neuroendocrinology, innate immunity, tissue regeneration, and ecology.

Acknowledgments

This research was supported by Korea Ministry of Environment (MOE) as “Eco-innovation Program (201300030002) and co-supported by Basic Science Research Program through the National
Research Foundation of Korea (NRF) funded by the Ministry of Education (2016R1D1A3B03934086).

The authors declare that they have no conflicts of interest with the contents of this article.
References


Figure legends

**Fig. 1** Annotation of the *P. pectinifera* transcriptome to NCBI NR, Kyoto Encyclopedia of Genes and Genomes (KEGG), and Gene Ontology (GO) database. A, Distribution of top 10 species with hits with a cut off e-value of < $10^{-3}$. B, Top 10 GO terms for each of the GO categories: molecular function (MF), biological process (BP), and cellular component (CC). C, Top 10 processes in KEGG pathways with hits using enzyme commission (EC) annotations.

**Fig. 2** Statistics of differentially expressed genes (DEGs) in tissues/organs of *P. pectinifera*. A, The number of tissue specific genes in six tissues/organs in *P. pectinifera*. B, The number of differentially expressed genes between different tissue/organ samples, with specific criteria ($\log_{2}FC \geq 4$ and FDR < 0.01). C, Principal component analysis based on all expressed genes showing three distinct groups of tissues. D, Cluster dendrogram between tissues/organs with AU/BP values (%) using Ward’s method. E, Correlation heatmap of the transcriptome between tissues/organ of *P. pectinifera*. BW, body-wall; GN, gonad; PC, pyloric caeca; CE, coelomic epithelium; SM, stomach; and TF, tube feet.
Table 1 MixS descriptors and statistics for the sequencing and de novo assembly of the starfish *Patiria pectinifera* transcriptome
Appendix A. Supplementary data legends

**Supplementary material 1** List of the predicted open reading frames (ORFs) from the *de novo* assembled transcriptome of *P. pectinifera*, including the results of annotations to NCBI NR, Gene Ontology (GO), and Kyoto Encyclopedia of Genes and Genomes (KEGG) database.

**Supplementary material 2** List of tissue-specific genes including the results of BLAST, GO, and KEGG annotations.

**Supplementary material 3** Differentially expressed genes (DEGs) in *P. pectinifera* transcriptome, based on comparison of six tissues/organs.
**A**

Pie chart showing the distribution of different species with Strongylocentrotus purpuratus as the major component (61%).

**B**

Bar charts illustrating the number of transcripts for biological processes, cellular components, and molecular functions.

**C**

Bar chart showing the number of transcripts for various metabolic pathways and processes, with Biosynthesis of antibiotics being the most represented (121 transcripts) followed by Purine metabolism (54 transcripts) and Amino sugar and nucleotide sugar metabolism (35 transcripts).
Table 1 MixS descriptors and statistics for the sequencing and *de novo* assembly of the transcriptome of the starfish *Patiria pectinifera*

<table>
<thead>
<tr>
<th>MixS descriptors</th>
<th>Item</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Investigation type</td>
<td>Eukaryote</td>
<td>Biome</td>
</tr>
<tr>
<td>Project name</td>
<td>PRJNA371229</td>
<td>Feature</td>
</tr>
<tr>
<td>Lat_lon</td>
<td>35.16°N, 129.19°E</td>
<td>Material</td>
</tr>
<tr>
<td>Geo_loc_name</td>
<td>South Korea: Cheongsapo</td>
<td>Temp.</td>
</tr>
<tr>
<td>Collection_date</td>
<td>2015-05-12T14:00 +02:00</td>
<td>Seq_meth</td>
</tr>
</tbody>
</table>

**Sequencing (Illumina HiSeq2500; paired-end, 2 × 101) stats**

<table>
<thead>
<tr>
<th>Process</th>
<th>Tissues</th>
<th>Body-wall</th>
<th>Coelomic epithelium</th>
<th>Gonad</th>
<th>Pyloric caeca</th>
<th>Stomach</th>
<th>Tube feet</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw read</td>
<td>Number</td>
<td>68,676,738</td>
<td>63,360,640</td>
<td>69,973,024</td>
<td>68,743,218</td>
<td>75,840,222</td>
<td>71,378,422</td>
<td>417,972,264</td>
</tr>
<tr>
<td></td>
<td>Total size (bp)</td>
<td>6,936,350,538</td>
<td>6,399,424,640</td>
<td>7,067,275,424</td>
<td>6,943,065,018</td>
<td>7,659,862,422</td>
<td>7,209,220,622</td>
<td>42,215,198,664</td>
</tr>
<tr>
<td>Adapter trimming</td>
<td>Number, a(%)</td>
<td>67,107,554 (97.7)</td>
<td>61,920,652 (97.6)</td>
<td>68,288,434 (97.8)</td>
<td>67,210,592 (97.7)</td>
<td>74,127,038 (97.7)</td>
<td>69,813,734 (97.8)</td>
<td>408,468,004 (97.7)</td>
</tr>
<tr>
<td>Contamination removal</td>
<td>Number, b(%)</td>
<td>67,068,478 (97.7)</td>
<td>61,888,058 (97.7)</td>
<td>68,209,202 (97.7)</td>
<td>67,180,392 (97.7)</td>
<td>74,104,076 (97.7)</td>
<td>69,717,310 (97.7)</td>
<td>408,167,516 (97.7)</td>
</tr>
</tbody>
</table>

a(%), (number of adapter trimmed read /number of raw read)×100 
b(%), (number of contamination removed read /number of adapter trimmed read)×100

**Assembly (De novo assembly; Trinity 2.0.6) stats**

<table>
<thead>
<tr>
<th>Process</th>
<th>Number of contigs</th>
<th>Total size (bp)</th>
<th>Mean size (bp)</th>
<th>Minimum size (bp)</th>
<th>Maximum size (bp)</th>
<th>N50 (bp)</th>
<th>GC, %</th>
<th>Data accessibility</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>De novo</em></td>
<td>549,625</td>
<td>459,072,158</td>
<td>835</td>
<td>224</td>
<td>35,675</td>
<td>1,473</td>
<td>42.5</td>
<td>DDBJ/EMBL/GenBank</td>
</tr>
</tbody>
</table>