BIOINFORMATICS LABORATORY
SOONCHUNHYANG UNIVERSITY
Genome Information Research

Transcriptomic analysis of endangered freshwater mussel *Cristaria plicata* provides valuable resource for species conservation

Bharat Bhusan Patnaik[1,2†], Tae Hun Wang[1†], Se Won Kang[1], Hee Ju Hwang[1], So Young Park[1], Eun Bi Park[1], Jong Min Chung[1]
Dae Kwon Song[1], Changmu Kim[3], Soonok Kim[3], Jun Sang Lee[4], Yeon Soo Han[5], Hong Seog Park[6], Yong Seok Lee[1*]

[1]*Department of Life Science and Biotechnology, College of Natural Sciences, Soonchunhyang University, Asan, Chungnam, 31538, Korea*
[2]*Trident School of Biotech Sciences, Trident Academy of Creative Technology (TACT), Bhubaneswar- 751024, Odisha, India*
[3]*National Institute of Biological Resources, Incheon, 22689, Korea*
[4]*Institute of Environmental Research, Kangwon National University, Chuncheon, Gangwon, 24341, Korea*
[5]*College of Agriculture and Life Science, Chonnam National University, Gwangju, 61186, Korea*
[6]*Research Institute, GnC BIO Co., LTD., Daejeon, 34069, Korea*

ABSTRACT

The freshwater mussel *Cristaria plicata* (Bivalvia: Eulamellibranchia: Unionidae) has been assessed as endangered by the Korean Red List of Threatened Species and Data deficient by International Union for Conservation of Nature and Natural Resources (IUCN) Red List of Threatened species. The number of individuals been dwindling in recent times, due to indiscriminate collection of specimens and loss of natural habitats. In order to understand the strategic indices for the conservation of the species, we conducted *de novo* transcriptome sequencing, assembly, and annotation analysis using Illumina HiSeq2500 next-generation sequencing (NGS) technology, Trinity assembler, and BLAST2GO analysis, respectively. We obtained 98.31% of high-quality reads from a total of 286,152,584 raw read sequences. The assembly generated a total of 453,931 contigs having a mean length of 731.2 and N50 length of 1,254. The contig sequences were clustered to 374,794 unigenes with a mean length of 737.1 and N50 length of 1,262. A 100% coverage of *C. plicata* mitochondrial genes within two unigenes validated the quality of the assembler. The BLAST top-hit distribution of unigenes against PANMDB (79,960 hits) showed maximum homology to other molluscs. The NCBI-KOG annotation showed the maximum preference of the transcriptome towards Cellular Processes and Signaling mechanisms with a total of 4,916 belonging to the signal transduction mechanism category. BLAST2GO analysis was helpful to decipher the putative genes related to immunity and reproduction that would be beneficial for protection of the population.
Keywords: Cristaria plicata; Transcriptome; Endangered species

Materials and Methods

◎ **Sample name:** ***Cristaria plicata*** (Korean Red List of Threatened Species)

◎ **Ethics Statement**
- permission (Ref. No. 2014-10) from the Guem River Basin Environmental Office.

◎ **Construction of mRNA-seq library and Illumina Sequencing**
- The mRNA-seq library was constructed using the mRNA-seq sample preparation kit (Illumina, San Diego, CA). cDNA synthesis and sequencing on the Illumina HiSeq 2500 sequencing platform.

◎ ***De novo* assembly**
- Before *de novo* transcriptome assembly, the raw reads were cleaned by removing adaptoronly reads, repeated reads and low-quality reads using Sickle softwares. Assembly using the Trinity program. Contigs were further assembled to unigenes (having 94% identity, 30 bp overlap) with the sequence clustering software TGICL.

◎ **Transcriptome annotation and discovery**
- We constructed a unique reference dataset that combined protein sequence data of Arthropoda, Nematoda, and Mollusks downloaded from the Taxonomy browser of NCBI nr database. The sequences stored as PANM-DB (http://malacol.or.kr/blast/) by using the formatdb program. The assembled *C. plicata* unigene sequences were searched against the PANM-DB reference database using BLASTx algorithm with an E-value threshold of 1.0E-5. GO, KEGG annotation using BLAST2GO software.

◎ **Identification of candidate genes related to immune response**
- The identification of potential genes in *C. plicata* involved in immune responses, sex determination, and reproduction was performed according to a keyword search of our BLASTx annotation results in PANM-DB.

◎ **Microsatellite markers discovery**
- The assembled unigenes of *C. plicata* were searched for Simple Sequence Repeats (SSR) motifs using the software MIcroSAtellite (MISA).

Results and Discussion

◎ Illumina HiSeq 2500 sequencing platform generated a total of 286,152,584 raw reads with about 36,055,225,584 number of bases. After a quality assessment wherein the raw reads were filtered to remove adapter sequences, low quality reads, and unambiguous sequences, a total of 281,322,837 clean reads of an average length of 124.1 bases were obtained.

Table 1. Transcriptome assembly statistics of *C. plicata* visceral mass using the Trinity analysis

Total number of raw reads		Contig information		Unigene information	
- Number of sequences	286,152,584	- Total number of contig	453,931	- Total number of unigenes	374,794
- Number of bases (bp)	36,055,225,584	- Number of bases (bp)	331,930,879	- Number of bases (bp)	276,264,683
Total number of clean reads		- Mean length of contig (bp)	731.2	- Mean length of unigene (bp)	737.1
- Number of sequences	281,322,837	- N50 length of contig (bp)	1,254	- N50 length of unigene (bp)	1,262
- Number of bases (bp)	34,909,374,303	- GC % of contig (%)	36.62	- GC % of unigene (%)	36.47
- N50 length of contig (bp)	126	- No. of large contigs (≥500bp)	151,695	- Length ranges (bp)	212-68,788

Table 2. Functional annotation of unigenes of the *C. plicata* transcriptom

Databases	All annotated transcripts	≤300 bp	300-1000 bp	≥1000 bp
PANM-DB	79,960	14,480	30,748	34,732
UNIGENE DB	13,934	1,848	3,721	8,365
KOG	40,196	4,763	11,445	23,988
GO	23,246	2,593	5,625	15,028
KEGG	4,776	483	927	3,366
All annotated	84,274	15,700	33,108	35,466

◎ Out of a total of 374,794 unigenes, 79,960 (21.33%), 40,196 (10.72%), and 13,934 (3.72%) unigenes showed similarity to sequences in PANM DB, Unigene DB, and KOG DB, respectively. In total, 84,274 (22.49%) annotated transcripts were found within the clustered unigenes of *C. plicata* visceral mass transcriptome.

◎ According to the analysis, highest isogeny was observed to the oyster, *Crassostrea gigas* (32,609 unigenes, 40.78%) followed by 12,065 (15.09%) unigenes to the owl limpet, *Lottia gigantea*. As expected, most of the unigene hits belonged to molluscans and other arthropod proteins. A summary of the top-hit InterPro domains identified 1,374 unigenes showing zinc finger, C2H2-like domain. The zinc finger domains are a regular feature in molluscs, insects and other crustacean groups that participate in important cell processing functions including signal transduction and transcriptional regulation. The C2H2-like zinc finger proteins are the most common DNA-binding motifs present in prokaryotic and eukaryotic transcription factors. The KOG chart reveals that the unigenes distributed to different functional categories cluster to four major groups that includes information storage and processing, cellular processes and signaling, metabolism, and poorly characterized. In GO based annotation, of the unigenes, 21,189 were assigned to molecular function, 11,419 to biological process and 6,391 to cellular component function. By performing BLASTX against KEGG database using BLAST2GO suite, we assigned 4,776 unigenes and 709 enzymes to important metabolic, cellular and immune function pathways.

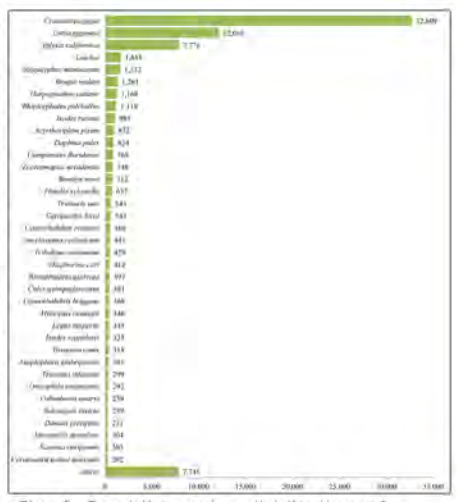

Fig. 1. Top-hit species distribution of *C. plicata* visceral mass unigenes against PANM-DB.

Results and Discussion

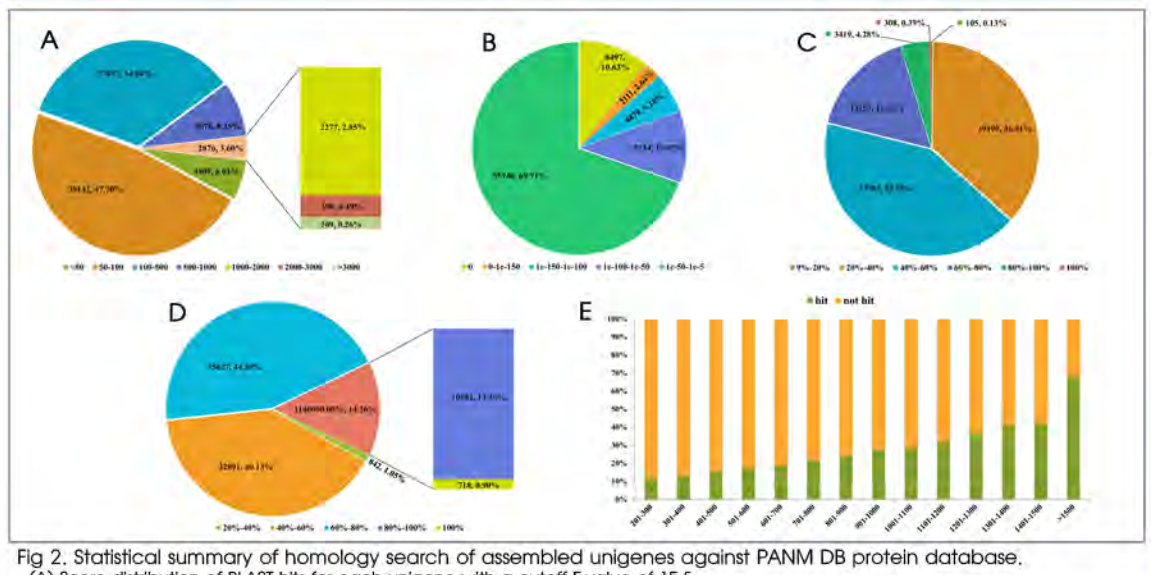

Fig 2. Statistical summary of homology search of assembled unigenes against PANM DB protein database.
(A) Score distribution of BLAST hits for each unigene with a cutoff E-value of 1E-5.
(B) E-value distribution of each unigene using BLAST hits with a cutoff E-value of 1E-5
(C) Identity distribution of the top BLAST hits for each unigene. (D) Similarity distribution of the top BLAST hits for each unigene.
(E) Length of unigenes compared with hits or without hits.

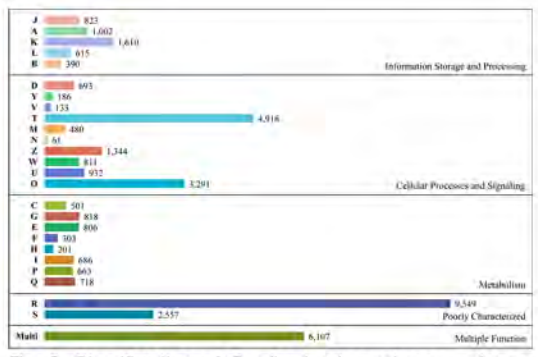

Fig. 3. Classification of *C. plicata* visceral mass unigenes based on sequence homology to entries in Clusters of orthologous groups (KOG) database at E-value of 1E-5.

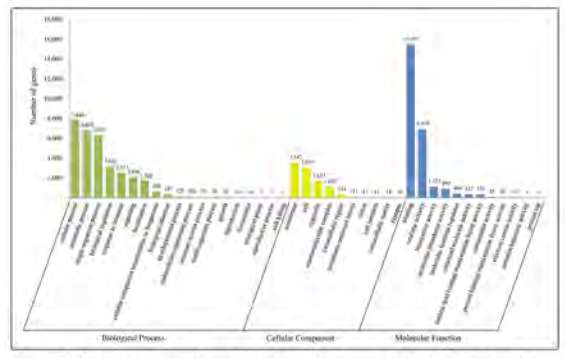

Fig. 4. Functional classification of assembled *C. plicata* unigene sequences based on GO categorization.

Fig. 5. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis.

Table 3. List of the top-hit 40 Interpro domains in *C. plicata* transcriptome

Table 4. List of the top-hit 40 Interpro domains in *C. plicata* transcriptome.

Repeat Numbers	Motif length							Total
	Di	Tri	Tetra	Penta	Hexa	Hepta	Octa	
3	732,862	129,130	20,848	3,199	1,319	108	61	887,527
4	74,489	13,076	3,729	221	81	2	3	91,601
5	13,084	4,021	1,659	58	1		3	18,826
6	5,870	1,735	1,548	11	1			9,163
7	2,848	1,159	166	11	1	1		5,187
8	2,968	987	195	2				4,152
9	1,978	363	187	4				2,432
10	1,621	311	134	6				2,072
11	2,209	256	116	1				2,582
12	1,684	248	107	1				2,032
13	597	166	117	6				886
14	755	193	83	3				1,034
15	726	128	98	2	1			955
16	686	111	125					922
17	617	102	87					806
18	555	80	56	3				694
19	473	101	34					608
20	476	95	27	1				599
21	3,614	185	132	3				3,934
Total	849,112	152,331	29,448	3,538	1,405	111	67	1,036,012

◎ A total of 282,790 unigenes containing 1,063,663 identified SSRs were found, with 190,214 of sequences containing more than one cSSR. After eliminating the Mono-nucleotide repeats (27,630 number of SSRs) and Deca-nucleotide repeats (21 number of SSRs), we obtained a total of 1,036,012 SSRs. The SSRs obtained included the most abundant Di-nucleotide repeats (849,112), followed by Tri- (152,331), Tetra- (29,448), Penta- (3,538), Hexa- (1,405), Hepta- (111), and Octa605 nucleotide repeats (67).