DE NOVO ASSEMBLY AND ANNOTATION OF PERNA INDICA TRANSCRIPTOME USING NEXT GENERATION SEQUENCING TECHNOLOGY

Sujitha Mary, Sandhya Sukumaran, Wilson Sebastian, A Gopalakrishnan
ICAR-Central Marine Fisheries Research Institute,
Ernakulam North P. O., Kochi - 682 018, Kerala, India .
:sujithamary@gmail.com

(MS Received: 08.08.2021; MS Revised: .16.09.2021; MS Accepted: 17.09.2021)

MS 2779 **(RESEARCH PAPER IN MARINE FISHERIES)**

Abstract

Indian coastline is bestowed with a rich diversity of molluscan species of which mussels belonging to the genus Perna (Family; Mytilidae) occupy a major position. Aquaculture importance of these species is globally recognized. They are extensively used as pollution indicators as they can accumulate contaminants at levels higher than that present in water. This property has made them valuable biomarkers of environmental pollution. The scarcity in the genetic resources of this species is one of the major hindrances in its wide usage as a biomonitor. Here we sequenced the first transcriptome of Perna indica, collected from the southernmost part of Kerala coast, using NGS illumina HiSeq 2500 technology. The de novo assembly of Perna indica transcriptome generated 157203 contigs and 132641 non-redundant genes with a minimum length of 201bp. The blast search and InterPro database scan resulted in assigning domains, ontology and pathway mapping to these genes which helped in identifying genes involved in metabolism, immune related genes, proteins and domains associated with these genes. 84% of the transcripts which had atleast one blast-hit showed similarity to species from bivalves. 12420 transcripts with GO terms and 17835 transcripts with 395 KEGG pathways were identified in the assembled transcriptome of Perna indica.

Keywords: *NGS, transcriptome, mussel, ontology, pathway mapping*

Introduction

The brown mussel *Perna indica* belongs to the family Mytillidae,true mussels in class bivalvia (Mollusca; Bivalvia; Lamellibranchia; Mytiloida; Mytilidae; Perna; *Perna indica*).The genus Perna, Philipson 1788, includes green and brown mussels from tropical, subtropical, warm and cold temperate regions of both hemispheres. The brown mussel exhibited different characteristics from that of *Perna viridis* and so it was named as a new species *Perna indica* (Kuriakose et.al, 1976). The comprehensive reviews in the genus *Perna* concluded that it has only three valid species *Perna canaliculus*, *Perna viridis* and *Perna perna* (Sidall 1980, Wood et.al.2007). Earlier green and brown mussels seen along the Indian coasts were grouped under *Mytillus* genus. But later studies revealed that they showed more similarity to the generic characteristics of *Perna* genus, and started considering green mussel *M. viridis* as *P. viridis*. In southernmost India two species are widely seen; green mussel *Perna viridis* and brown mussel *Perna indica* with affinity to African brown mussel *Perna perna* (Sheela et.al. 2013).After a series of controversial research conducted by different groups recently it has been suggested that *Perna indica* is not a distinct species and it is an introduced population of *Perna perna* (Divya et.al, 2009, Gardener et.al.,2016). The native versus introduced status of *P.indica* can be confirmed only with a detailed understanding for genetic structure of the genus in *Perna*.

The limited genomic resource in mussels prevents their usage as biomonitors for environmental pollution. Currently, there are only very few whole genomes or transcriptome sequences of mytillidae family in publicly available sequence databases. In this study, we sequenced the transcriptome of *Perna indica* collected from the southernmost part of Indian coast using NGS technology. The assembly and functional annotation of *Perna indica* mussel is described here. *Perna indica* transcriptome characterization will have important utility in its use as a bioindicator of pollution as well as in comparative studies of bivalves. The genomic resources for *Perna indics* consist only of the complete mitochondrial genome and a few COX I genes (Uliano et.al, 2016). So the output of the present study will form a valuable genetic resource for further evolutionary, taxonomic and biomonitoring studies in *Perna* genus and other bivalves.

Materials and methods

Sample collection and rna extraction

Live *Perna indica* samples were collected from CMFRI, Vizhinjam, Kerala. nd transported them to the laboratory in a wet gunny bag. They were placed in a clean tank with aerated artificial sea water (32ppt salinity). After feeding with mixed microalgal feed, tank was cleaned and filled with seawater (32ppt salinity) again. These samples were maintained for 14days and then dissected to obtain the target tissues, gill, foot, adductor muscle and mantles for sequencing. RNA was extracted from all four tissues using TRIzol Reagent following the manufacturer's protocol (Invitrogen, USA). Total RNA was extracted from the tissues separately and pooled an equal amount for RNASeq. The quality of extracted RNA was checked using 1% agarose gel electrophoresis. Quantity and concentration were determined using a nanodrop-2000 spectrophotometer (Thermofisher, USA). RNA integrity was checked using the Qubit and Agilent Tapestation method. Samples with Rin >=8.0 were selected for cDNA library preparation.

cDNA library preparation and sequencing

RNA was converted to cDNA and prepared library using TrueSeq RNA sample preparation kit (Poly A) according to manufacturer's protocol. The procedure involves the isolation of mRNA from total RNA using polydT capture oligo/bead fragmented and converted to cDNA. The cDNA fragments are end-repaired to known oligonucleotides (adapters) to generate the library. The library was then sequenced on NGS machine illumina HiSeq2500 to produce millions of paired-end short reads (Scigenom labs,Kochi,Kerala).

Transcriptome de novo assembly

Prior to assembly, the base quality, adapter contamination and kmer frequencies of the illumina reads should be known. The quality of reads were checked using FastQC (Andrew, 2015), to derive a quick impression on the quality metrics of our raw reads. The reads that contain erroneous kmers will result in lowly expressed transcripts which affects the assembly adversely. So erroneous Kmer correction was accomplished using Rcorrector (Li et.al, 2015) which was designed specifically for RNAseq data. Illumina sequencing demultiplexing will not remove adapter content completely. So these bases and other low-quality bases were trimmed using Trim-Galore (Kreuger, 2015), which is a wrapper for cutadapt (Martin, 2011). The high-quality reads obtained after cleaning were then assembled with Trinity version 2.9.1 (Grabherr et.al, 2011) using the default options. The transcripts obtained after assembly were clustered using CD-HIT version 4.6.8 (Li et.al, 2006) and the output was used for further downstream analysis. The contigs were scanned against SILVA database to control ribosomal RNA contamination (Pruesse et.al, 2007).

Transcriptome annotation

The candidate coding regions in non-redundant transcripts were identified using Transdecoder version 5.5. The similarity search using Blastp and homology search using Pfam were integrated to increase the sensitivity for capturing ORFs with functional significance. Blast search with the candidate coding regions as query against NCBI's nr, nt and Swissprot databases were done to reveal the similarity of transcripts with the functionally known proteins deposited in public databases. The predicted protein sequences were searched against HMM profiles (Pfam database version 33.1) using HMMER v3.3 to obtain Pfam domain annotation in the coding region of the *Perna indica* transcriptome. The assembled transcripts were also processed by InterProScan to predict the domain signatures and GO ontology terms. KEGG pathways were assigned to transcripts using KEGG automatic annotation server (KAAS;http://www.genome.jp/kegg/kaas/)

Results and discussion

Sequencing and assembly evaluation

The first comprehensive reference transcriptome of mussel *Perna indica* was generated in this study. It will be a valuable genomic resource for the molecular level research that is undergoing in mussels as well as in bivalves. The raw reads obtained from illumina HiSeq sequencing were around 63705960. After quality trimming by Trim-Galore and Rcorrector we could retain 93% of illumina reads. This data tells that it is suitable to produce a well-assembled transcriptome which will be a strong genomic resource for *Perna indica* mussel. The *denovo* assembly generated 157203 contigs which produced 132641 unigenes after clustering using CD-Hit. The summary of transcriptome assembly data is shown in Table 1. The maximum and minimum length of the transcript is 37616bp and 201bp respectively. The length distribution of transcripts can be seen in Fig.1 which indicates the maximum number of transcripts has their length between 200bp and 350bp. There were only 1695 transcripts having a length of more than 5000bp.

Functional annotation

The transcripts with potential for coding were identified using the Transdecoder program.

There were 28692 transcripts which were differentiated as complete, internal, 5' prime partial and 3' prime partial ORFs. These non-redundant protein-coding transcripts which were predicted to be complete were considered for downstream analysis. InterProScan analysis predicted 18625 assembled transcripts with putative functions for protein domains. All these sequences were also aligned against public databases such as NCBI nr, NCBI nt and Swissprot using Blast program to identify the functions of assembled transcripts. 19940 transcripts had at least one hit in blast similarity search, in which 84% of them showed significant matches with bivalves. The species distribution of top blast-hits is shown in Fig. 2.

The protein function can be easily identified by finding the domains present in it. Here the search using Hmmscan predicted 6304 Pfam domains from the scanned protein coding sequences.5699 sequences were having at least one domain. Top 20 hits are shown in Table 2. The domains that occurred in more number were the immune related domains and domains that help in phosphorylation and ATPase binding.

InterProScan tool was used to predict pathways and GO terms associated with the protein-coding sequences among the assembled transcripts of *Perna indica*. It assigned 58223 unique InterPro annotations to these transcripts. Domains like immunoglobulin-like fold, Protein Kinase, Ankyrin repeat, P-loop containing nucleoside tri-phosphate and zinc finger domain were assigned to the maximum number of transcripts. Table 3 shows the domains which occurred more frequently in InterProScan annotations of assembled *Perna indica* transcriptome sequences.

Gene Ontology terms were assigned according to InterProScan pathway prediction. A total of 12420 assembled transcripts were assigned at least one well-defined GO term. They were further classified into three categories namely cellular component, molecular function and biological process. The higher number of GO terms were linked with molecular function (64%), biological process (26%) and cellular components (9%) respectively. WEGO generated graphical representation of GO terms is shown in Fig.3.

For KEGG annotations 17835 assembled transcripts were assigned 395 KEGG pathways which belong to Metabolism (1838 genes), Genetic information processing (550 genes), Environmental information processing (1206 genes), cellular processes (963 genes), Organismal systems (1955 genes), Human diseases (3570 genes). The percentile distribution of genes under these main categories are shown in Fig.4. In each of the above category the most representative terms were Carbohydrate metabolism, Folding, sorting and degradation, Signal transduction, Transport and catabolism, Endocrine system, Neurodegenerative diseases respectively.

Data availability

The NGS illumina-seq run generated raw data has been deposited in NCBI SRA database in BioProject PRJNA692331 with SRA accession SRR15968044.

Conclusion

We have conducted the first exhaustive transcriptome sequencing study of Perna *indica* collected from the southern part of the Kerala coast. The illumina HiSeq sequencing generated 63705960 raw reads which was assembled into 157203 contigs clustered as 132641 unigenes. 12420 genes were assigned GO terms and 395 KEGG pathways were assigned to 17835 genes. Transcriptome sequence data and annotation are valuable as *Perna indica* mussels can be used as a biomonitor in environmental pollution.

Acknowledgment

This work was supported by the grant entitled Back-to-Lab Post-doctoral fellowship funded by Kerala State Council for Science Technology and Environment (KSCSTE).

References

1. **Andrews, S.2015**. FastQC: a quality control tool for high throughput sequence data http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
2. **Divya, P., Gopalakrishnan, A., John, L., Thomas, P.C., and Lakra, W.S. 2009.** Mitochondrial DNA (Cytochrome c oxidase I) sequencing of Indian marine mussel *Indian J. Fish*., 56(3): 223-226.
3. **Gardner, J .P.A., Jamila, P., George, S. and Edward, J.K.P. 2016**. Combined evidence indicates that Perna indica.
4. **Kuriakose and Nair 1976** is *Perna indica* (Linnaeus, 1758) from the Oman region introduced into southern India more than 100 years ago. *Biol.Invasions*, 18:1375–1390.
5. **Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L.,Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A.,Rhind, N., Di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman,**
6. **N., Regev, A.2011.** Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*, 29(7):644-52.
7. Krueger, F.2015. Trim Galore http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/
8. **Kuriakose, P.S. and Nair, N.B.1976.** The genus *Perna* along the coasts of India with the description of a new

species *Perna indica*. *Aquat. Biol*., 1: 25-36

9. **Li, S and Liliana, F.2015.** Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. *GigaScience*, 4:48
10. Li, W, Godzik, A.2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* ,22:1658–9.
11. **Martin and Marcel.2011.** Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*,17(1): 10-12.
12. **Pruesse, E, Quast, C, Knittel, K, Fuchs, B.M, Ludwig, W.G, Peplies, J, Glöckner, F.O.2007.** SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB.*Nucl. Acids Res*., 35:7188-7196.
13. **Sheela, T and Patterson, E.J.K. 2013**. Molecular phylogenetic analysis of two closely related marine Indian mussels of genus Perna (PHILIPSSON, 1788) based on mitochondrial (COI) and nuclear (ITS) genes. *Journal of Aquatic Biology & Fisheries*,1: 123-139.
14. **Siddall, S.E. 1980.** A classification of the genus Perna (Mytilidae). *Bull. Mar. Sci*., 30: 858–70.
15. **Uliano-Silva, M, Americo, J, Bastos, A.S, Furtado, C, Rebelo, M.F, Prosdocimi, F.2016**.Complete mitochondrial genome of the brown mussel *Perna indica* (Bivalve, Mytilidae). *Mitochondrial DNA A DNA Mapp Seq Anal*, 6:3955-3956.
16. **Wood, A.R., Apte, S., MacAvoy, E.S. and Gardner, J.P.A. 2007**. Molecular phylogeny of the marine mussel genus Perna (Bivalvia: Mytilidae) based on nuclear ITS (1&2) and mitochondrial (COI) DNA sequences. *Mol. Phylogenet. Evol.*, 44 (2): 685-698.

Table 1 Assembly Statistics

No: of raw reads	63705960
No: of raw reads after quality filtering	59828633
No: of assembled transcripts	157203
No: of non-redundant transcripts	132641
N50(bp)	1725
N90(bp)	316
N10(bp)	7292
Ex90N50(bp)	2765
Maximum Length of bp	37616
Minimum length of bp	201
%GC	35
Mean orf percent	51.28

Table 2 Pfam Domain hits in coding sequences of *Perna indica*

Domain Name	No:of hits
AAA domain	181
Immunoglobulin domain	140
Protein kinase domain	135
Protein tyrosine and serine/threonine kinase	128
Ankyrin repeat	121
Immunoglobulin I-set domain	112
AAA ATPase domain	105
Immunoglobulin V-set domain	105
Protein of unknown function (DUF1664)	97
Spc7 kinetochore protein	89
RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)	87
Fibronectin type III domain	85
PH domain	81
Ephrin type-A receptor 2 transmembrane domain	79
Zinc finger, C2H2 type	77
C2H2-type zinc finger	77
EGF-like domain	71
WD domain, G-beta repeat	70
Tetratricopeptide repeat	68
Zinc finger, C3HC4 type (RING finger)	68

Table 3 InterPro Annotations in *Perna indica*

Accession	Description
IPR027417	P-loop containing nucleoside triphosphate hydrolase
IPR013783	Immunoglobulin-like fold
IPR036770	Ankyrin repeat-containing domain superfamily
IPR020683	Ankyrin repeat-containing domain
IPR002110	Ankyrin repeat
IPR036179	Immunoglobulin-like domain superfamily

IPR011009	Protein kinase-like domain superfamily
IPR015943	WD40/YVTN repeat-like-containing domain superfamily
IPR007110	Immunoglobulin-like domain
IPR013087	Zinc finger C2H2-type
IPR013083	Zinc finger, RING/FYVE/PHD-type
IPR016024	Armadillo-type fold
IPR000719	Protein kinase domain
IPR036322	WD40-repeat-containing domain superfamily
IPR032675	Leucine-rich repeat domain superfamily
IPR036236	Zinc finger C2H2 superfamily
IPR000742	EGF-like domain
IPR011993	PH-like domain superfamily
IPR011992	EF-hand domain pair
IPR001680	WD40 repeat
IPR003599	Immunoglobulin subtype
IPR017986	WD40-repeat-containing domain
IPR043502	DNA/RNA polymerase superfamily
IPR012677	Nucleotide-binding alpha-beta plait domain superfamily
IPR035979	RNA-binding domain superfamily
IPR002048	EF-hand domain
IPR000315	B-box-type zinc finger
IPR001841	Zinc finger, RING-type
IPR017441	Protein kinase, ATP binding site
IPR011990	Tetratricopeptide-like helical domain superfamily

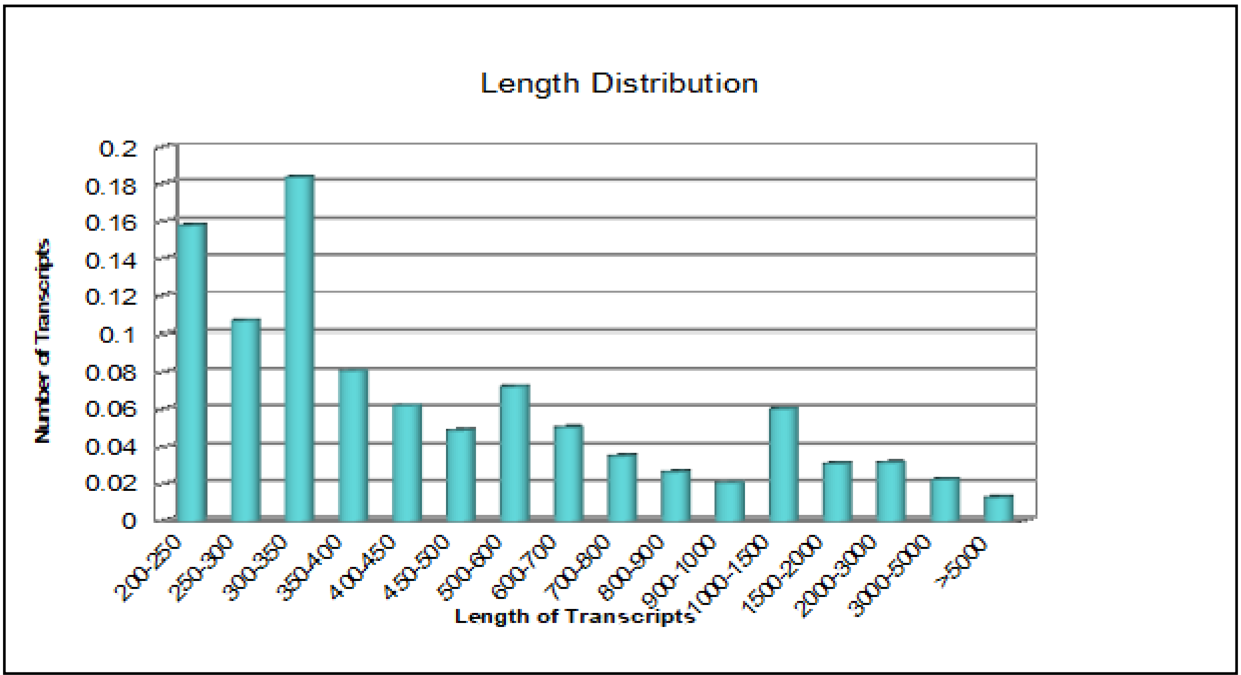

Fig.1 Length distribution in *P.indica* transcripts

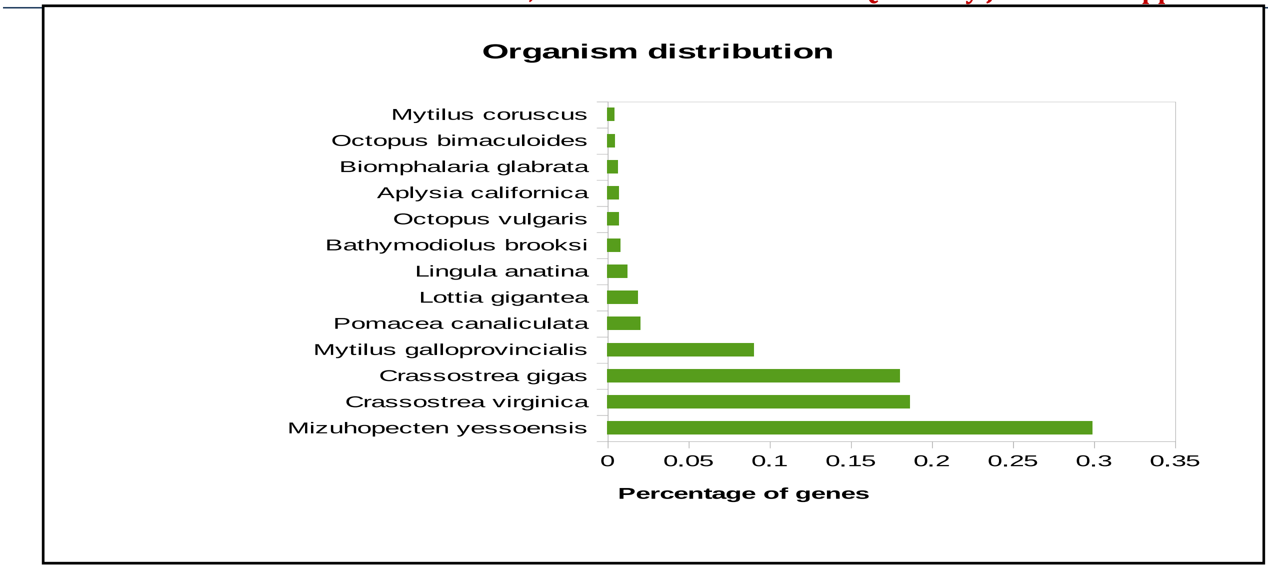

Fig. 2 Species distribution in top blast-hits of *Perna indica* transcriptome

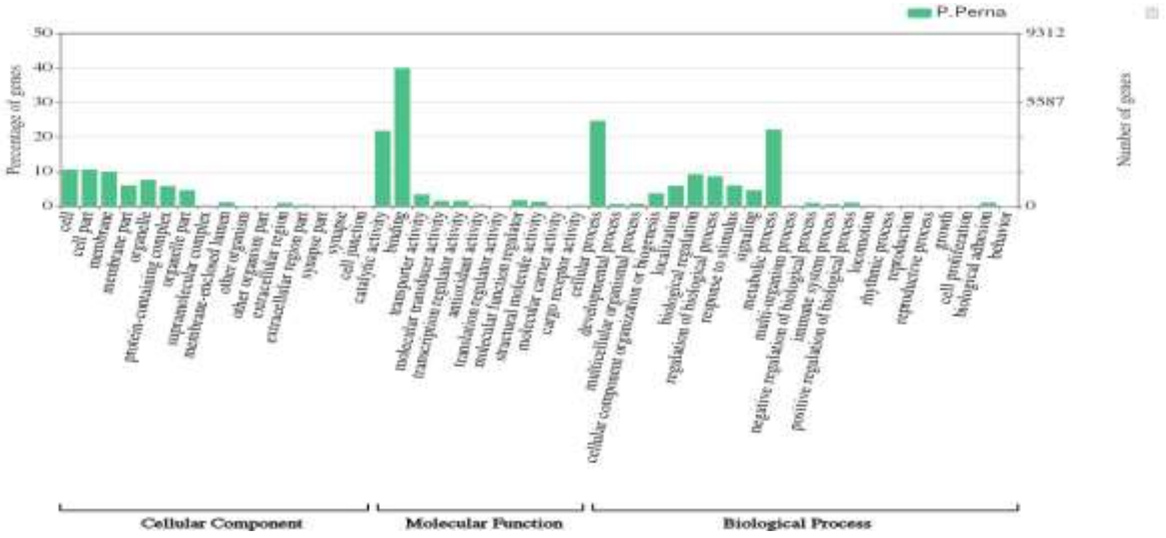

Fig. 3 Graphical representation of GO terms assigned to assembled transcripts

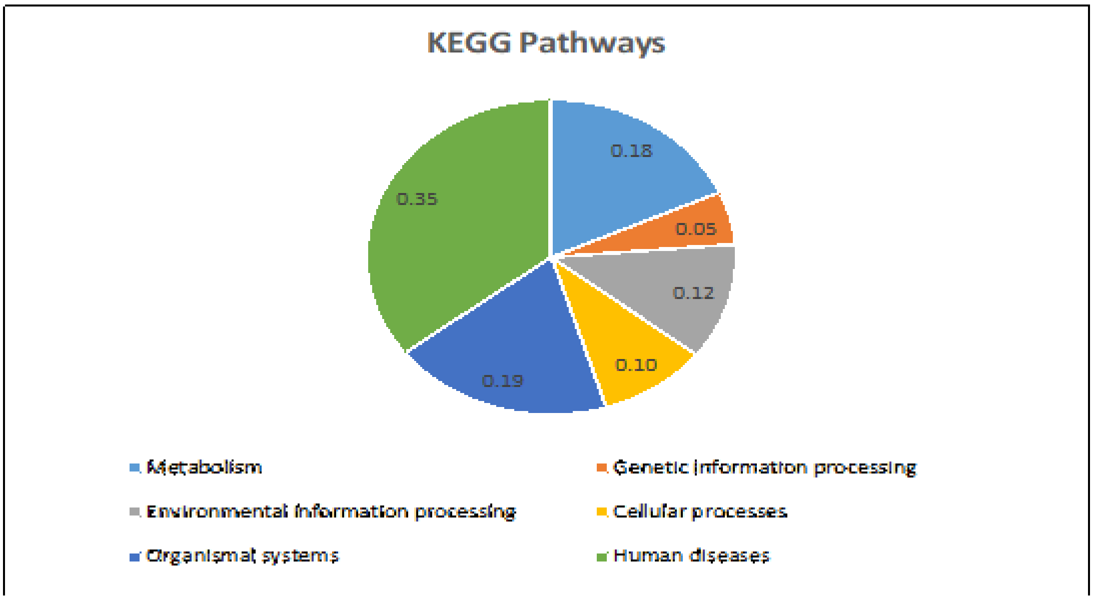

Fig. 4 Graphical representation of KEGG terms assigned to assembled transcripts