



OPEN

DATA DESCRIPTOR

De novo transcriptome assembly of the *Perna viridis*: A novel invertebrate model for ecotoxicological studies

V. G. Vysakh^{1,2}, Sandhya Sukumaran¹✉ & Gopalakrishnan A.¹

Mussels, particularly *Perna viridis*, are vital sentinel species for toxicology and biomonitoring in environmental health. This species plays a crucial role in aquaculture and significantly impacts the fisheries sector. Despite the ecological and economic importance of this species, its omics resources are still scarce. We generated a gill-specific reference transcriptome for *Perna viridis* using 292 million short Illumina reads from eight pooled gill tissue samples isolated from twenty-four individuals. The Trinity assembler generated 438,842 transcripts with an N50 of 1,958 bp. Several databases were employed in the annotation process. This dataset greatly expands the omics resources of bivalve databases and advances our knowledge of transcriptomics, molecular biology, environmental toxicology, and cancer research.

Background & Summary

Mussels, as bivalves, display a range of diverse adaptations evolved to manage the stressors associated with their sedentary lifestyle in challenging environments^{1,2}. These stressors include variations in salinity and temperature, as well as exposure to pollution and parasites^{2–4}. This remarkable adaptability is reflected in their gene expression pattern and genomic data profiles, making mussels valuable model organisms for studying genomic evolution^{1,5–7}.

The Asian green mussel (*Perna viridis*) holds significant economic importance due to its ecological role and value in fisheries. Beyond their foundational role in aquatic ecosystems, mussels exhibit notable phenotypic plasticity, allowing them to alter physiological and physical traits in response to environmental changes^{8–10}. This plasticity, linked to genetic variations, enables diverse trait expressions under varying conditions. The genomic architecture of mussels responds to stressors such as parasitic infections (e.g., *Perkinsus spp.*) and pollutants, ranging from nano-micro scale to bulk contaminants^{11–16}. Their high genetic diversity facilitates local adaptation, with populations developing traits that enhance survival in specific environments, including pollutant tolerance. Whole-genome duplications further increase genetic variability, providing a broader range of potential adaptations. Additionally, mussels demonstrate compensatory mechanisms, such as enhanced biomineralisation, to cope with environmental stressors^{17–19}. Therefore, the genomic plasticity of mussels enables effective adaptation through phenotypic changes, gene regulation, and genetic diversity, contributing to their ecological success. Understanding these mechanisms is crucial for the conservation and management of mussel populations in aquaculture^{5,17}.

Tissue-specific transcriptome assembly is particularly critical for emerging model organisms like *Perna viridis*. Such assemblies are invaluable for various application-level studies, including differential gene expression analysis, genome-wide association studies (GWAS), and comprehensive genome-wide characterisations. Transcriptome assembly is often more advantageous than genomic assembly for differential gene expression research, as it enables accurate characterisation of alternative splicing events and expression patterns that may be obscured in genome assemblies. This is particularly important for sentinel mussels like *Perna viridis*, which serve as vital bioindicators in biomonitoring and ecotoxicological studies^{20–28}.

¹Marine Biotechnology Fish Nutrition and Health Division, Central Marine Fisheries Research Institute, Post Box No 1603 Ernakulam North PO., Kochi, 682018, Kerala, India. ²Mangalore University, Mangalagangotri, Mangalore, 574199, Karnataka, India. ✉e-mail: sandhyasukumarancmfri@gmail.com

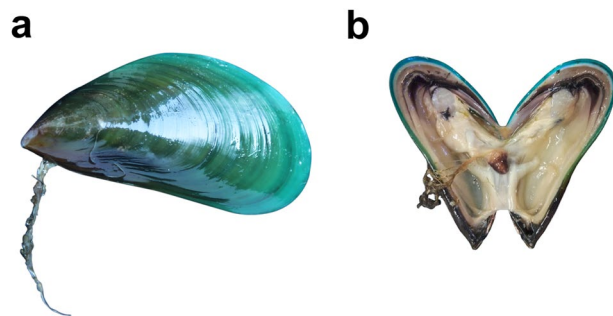


Fig. 1 *Perna viridis*. Photo highlighting key anatomical features relevant to the study.

Tissue-specific transcriptome and expression profiling are indispensable in this context. The gill, in particular, is a vital organ for mussels, serving as the primary site of exposure to environmental pollutants and chemicals, including nanoparticles and microplastics^{10,14,29,30}. It is also a key target for infections from parasitic and bacterial organisms. Numerous studies have highlighted tissue damage and the impact of contaminants on gill tissues, underscoring the need for detailed transcriptomic data^{10,31}. Furthermore, *Perna viridis* has recently gained recognition as a model organism for cancer research based on emerging genomic data, highlighting the future importance of tissue-specific transcriptome resources in advancing cancer biology²⁴.

Given the ecological and biomedical significance of *Perna viridis*, we developed a gill tissue-specific transcriptome to enrich the molluscan database. This resource is crucial for ecotoxicological and biomonitoring studies and will support future investigations into the molecular mechanisms underlying pollutant responses and cancer biology.

Methods

Sample collection. Wild adult Asian green mussels (*Perna viridis*) weighing 40–50 g and having shell lengths of 6 ± 0.41 cm were collected from the coasts of Munambam, Arabian Sea (10.1772° N, 76.1655° E) (Fig. 1). A total of 24 individuals were gathered. Gill tissues were snap-frozen in liquid nitrogen and stored at -80°C until further processing. All animal experimentation protocols were approved by the Institutional Animal Ethical Committee of ICAR-Central Marine Fisheries Research Institute (CMFRI), Kochi, and adhered to the ARRIVE guidelines (<http://arriveguidelines.org>).

RNA extraction, library preparation and sequencing. Total RNA was extracted from each individual using TRIzol Reagent (Invitrogen, USA) following the manufacturer's instructions. To remove any genomic DNA, the RNA samples were treated with RNase-Free DNase I (Qiagen, USA). RNA samples with an OD 260/280 ≥ 1.8 and an RNA integrity number ≥ 7 were selected for further analysis. To ensure representativeness, we combined equal volumes of high-quality RNA from three distinct individuals to produce a unified sample. This process was repeated across eight samples, each is a pool of three individuals, and subsequently used for cDNA synthesis and sequencing. RNA quality and quantity were assessed using 1% denaturing agarose gel electrophoresis and the Agilent 4200 Bioanalyzer (Agilent Technologies, USA). cDNA libraries were constructed following the TruSeq RNA Sample Preparation Kit v2 protocol (Illumina, Cat. No. RS-122-2001 and/or RS-122-2002). Library quality was evaluated using the Agilent 2100 Bioanalyzer (Part. No. G2939BA), and concentration was measured with the KAPA Library Quantification Kit (Cat. No. KK4824). Sequencing was performed on an Illumina NovaSeq. 6000 platform, utilising paired-end mode with 150 bp reads.

Pre-assembly processing. All bioinformatics analyses were conducted using the high-performance computing system “MY FISH” provided by CMFRI. The workflow of the bioinformatics pipeline is represented in Fig. 2. A total of 292,823,639 raw reads were generated after sequencing (Table 1). Raw data quality was assessed using FastQC software version 0.11.8³². The quality results for all samples were aggregated into a single report using the MultiQC v1.12³³. Adaptor sequences and low-quality reads (PHRED score < 20) were removed using the fastp tool v1.12³⁴. A total of 244,726,426 cleaned reads were maintained to build the *de novo* transcriptome assembly, representing 83.5% of the raw reads. Metrics for the generated transcriptome data are summarised in Table 2.

De novo transcriptome assembly. Trinity (version 2.11), a *de novo* assembly tool for eukaryotes based on de Bruijn graphs, was used to optimise the transcriptome assembly, reduce chimeric transcripts, and improve reliability³⁵. Trinity was employed to assemble the RNA-seq data with default parameters. This approach resulted in a total of 438,842 transcripts (Table 2). Following the assembly process, the assembled transcripts were subjected to post-assembly quality assessment using tools such as BUSCO v5.4.4³⁶ (Benchmarking Universal Single-Copy Orthologs) and TransRate v1.0.3³⁷ to assess assembly quality and completeness. Potential contaminants were removed using NCBI-FCS and Kraken 2^{38,39}.

Assembly thinning and redundancy reduction. Two distinct approaches were employed for assembly thinning to enhance the efficiency and quality of transcriptomic analysis. Firstly, CD-HIT-est v. 4.8.1⁴⁰ clustering was utilised to group similar isoforms and select representative sequences for each gene cluster. This method effectively reduces redundancy while maintaining crucial sequence information, thereby simplifying the dataset

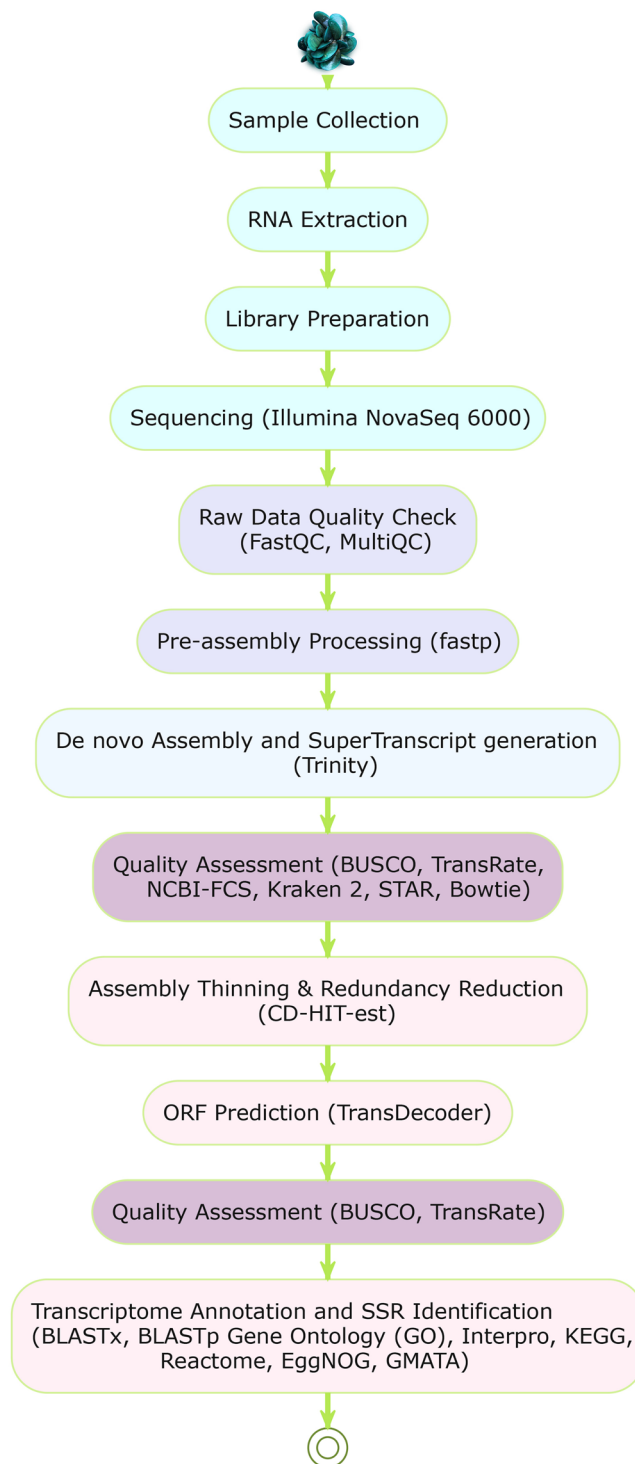


Fig. 2 Workflow of the bioinformatic pipeline. From raw data to annotated transcripts for the *de novo* transcriptome assembly of Gill tissues from *Perna viridis*.

for downstream analysis. Secondly, SuperTranscript was applied to construct a hybrid transcript by stitching together all unique exons from various isoforms into a continuous linear sequence. Although this hybrid transcript may not exist in a natural biological context, it provides a comprehensive representation of all possible exonic regions, ensuring no sequence information is lost. Together, these methods complement each other by consolidating redundant data and integrating all relevant sequence information, thus improving the depth and accuracy of transcriptome analysis⁴¹.

The initial assembly output from Trinity was processed using CD-HIT-est v. 4.8.1, a hierarchical clustering tool designed to eliminate redundant and fragmented transcripts often encountered in *de novo* assemblies,

Sample	Read Orientation	Number of Reads Obtained	Poor Quality Reads	GC%	Mean Read Length (bp)
PV1	R1	16,115,540	0	37	151
	R2	16,115,540	0	37	151
PV2	R1	17,362,360	0	37	151
	R2	17,362,360	0	37	151
PV3	R1	18,266,798	0	38	151
	R2	18,266,798	0	38	151
PV4	R1	18,594,860	0	37	151
	R2	18,594,860	0	37	151
PV5	R1	20,004,803	0	38	151
	R2	20,004,803	0	38	151
PV6	R1	27,755,105	0	37	151
	R2	19,699,776	0	37	151
PV7	R1	27,755,105	0	37	151
	R2	27,755,105	0	37	151
APD3	R1	24,579,973	0	39	151
	R2	24,579,973	0	39	151

Table 1. Summary of sequencing read statistics. Summary of sequencing read statistics, including total reads, read length distribution, and quality metrics.

Metric	Trinity	CD-HIT-est (Unigens)	SuperTranscript
Total Transcripts	438,842	378,672	308,707
N90	345	315	290
N70	890	687	565
N50	1,958	1,554	1,223
N30	3,471	3,005	2,920
N10	6,742	6,171	7,115
GC Content (%)	33.96	34%	33%
Bases with N (Ambiguity)	0	0	0
Proportion of Bases N	0	0	0
Number of Transcripts with ORF	63,033	43,324	19,464
Mean ORF Percentage	39.3%	38.12%	31.41%
Mean Sequence Length (bp)	965.73	842.29	753.18
Total Number of Bases	423,808,106	318,952,430	232,512,581
Sequences Under 200 bp	23	12	0
Sequences Over 1,000 bp	108,162	77,183	48,461
Sequences Over 10,000 bp	1,137	691	791
Smallest Sequence Length (bp)	179	179	201

Table 2. Transcriptome Assembly Metrics. Key assembly metrics for the *Perna viridis* transcriptome, including transcript count and average length.

thereby generating unique gene sequences. CD-HIT-est was executed with default parameters, setting a similarity threshold of 95%. This was followed by a second validation phase of the CD-HIT-est output, as detailed in Table 2. Subsequently, the CD-HIT-est output was analysed using TransDecoder v. 5.7.0⁴², a standard tool for identifying long open reading frames (ORFs) in assembled transcripts. TransDecoder was run with default settings, which include ORF prediction on both strands of the assembled transcripts, irrespective of the specific sequencing library used. It also ranks ORFs based on their completeness by examining the presence of AA codons upstream of a start codon (AUG) and checking for an in-frame stop codon to determine the completeness of the 5' end. The “Longest ORF” rule was applied to select the longest 5' AUG as the translation start site relative to the in-frame stop codon⁴².

Assessment of transcriptome assembly quality. The quality of the transcriptome assembly was evaluated using multiple analytical tools. Initially, TransRate v. 1.0.3³⁷ was used to assess the preliminary assembly outputs from Trinity, CD-HIT clustering, and the supertranscript. This tool provided essential metrics to identify potential errors and assess the quality of the assembled transcriptomes.

To measure transcriptome completeness, BUSCO v. 5.4.4³⁶ was employed. The assessment was conducted against three ortholog databases: Metazoa (metazoa_odb10), Mollusca (mollusca_odb10), and Eukaryota (eukaryota_odb10)⁴³. The results, detailed in Table 3, demonstrated high levels of completeness, with significant

Lineage Dataset	Total BUSCOs	Complete Single-copy	Complete Duplicated	Fragmented	Missing	Percentages
Mollusca	5295	2,872 (54.24%)	2,132 (40.26%)	69 (1.3%)	222 (4.19%)	Complete: 94.5% Fragmented: 1.3% Missing: 4.19%
Eukaryota	255	167 (65.49%)	87 (34.12%)	1 (0.39%)	0	Complete: 99.61% Fragmented: 0.39% Missing: 0%
Metazoa	954	600 (62.89%)	351 (36.79%)	0	3 (0.31%)	Complete: 99.69% Fragmented: 0% Missing: 0.31%

Table 3. BUSCO Completeness Assessment. BUSCO completeness percentages for *Perna viridis* across Mollusca, Eukaryota, and Metazoa lineages, including single-copy and duplicated genes.

percentages of complete gene representation across all databases. Specifically, the completeness of BUSCOs varied from 62.89% for single-copy and 36.79% for duplicated genes in Metazoa to 65.49% and 34.12% in Eukaryota, respectively.

Additionally, the trimmed raw reads were mapped to the final *de novo* transcriptome assembly using STAR v2.7.8a⁴⁴, a well-established gene alignment tool. The mapping efficiency, reflected as a read support value, achieved 89%, further validating the high quality of the transcriptome assembly. To further validate this approach, additional mapping was performed using Bowtie v0.7.17⁴⁵, yielding an average mapping rate of above 95%. This high mapping rate underscores the accuracy and reliability of the assembly.

Transcriptome annotation. The predicted protein-coding regions were annotated through homology searches using BLASTx and BLASTp against multiple databases, including NCBI nr and UniProtKB, with an e-value threshold of 1e-5. Gene Ontology (GO), pathway and Enzyme code annotations were performed using EggNOG⁴⁶, KEGG⁴⁷, Reactome⁴⁸, and Blast2GO v5.2.5⁴⁹, assigning transcripts to specific cellular components, functions, and biological processes^{46,48,50-54}. Functional protein domains were identified with InterProScan v4.0⁵¹, which integrates predictive models from various databases such as Gene3D, PANTHER, Pfam, and others.

Identification of Simple Sequence Repeats (SSRs) in the transcriptome. Simple sequence repeats (SSRs) within the generated transcriptome were identified using GMATA (Genome-wide Microsatellite Analyzing Towards Application)⁵⁵. SSRs with repeat units ranging from 2 to 10 base pairs and a minimum repeat number of 5 were analysed to characterise the repetitive sequences present in the transcriptome.

Data Records

The curated transcriptome assembly has been submitted to NCBI GenBank under Bioproject PRJNA1149310 with accession number GKYN00000000⁵⁶. The high-quality sequence data, devoid of vector contamination, has been deposited in the NCBI Sequence Read Archive (SRA)⁵⁷⁻⁶⁴. The annotated transcriptome assembly, including the Gene Ontology (GO) annotations, is available on Figshare⁶⁵.

Technical Validation

RNA and library quality control. Only RNA samples with confirmed high quality and integrity (concentration ≥ 7 ng/ μ L and a flat baseline on the Bioanalyzer) were used for pooling. It is important to note that the RNA Integrity Number (RIN) is not a reliable measure of RNA integrity in mussels, as molluscs, in general, display different numbers of peaks due to the presence of hidden breaks in rRNA. Prior to multiplexing, libraries were checked for fragment distribution and concentration to ensure that all sequencing criteria were met.

Read and *de novo* assembly basic statistics. Sequencing yields, assembly, and annotation statistics, along with completeness results, are presented in Tables 2 and 3. Briefly, a total of approximately 292 million raw reads were obtained, with an average of 6.25 Gb data per sample (Table 1). The raw paired-end reads possess a high-quality Q30 score. FASTQC results confirmed that the cleaned reads passed the minimum quality standards, with over 83.5% of reads retained after filtering and adapter removal, ensuring high-quality *de novo* assemblies.

The resulting transcriptome assembly comprised 438,842 transcripts, with a mean contig length of 965 bp. The N50 values exceeded 1,000 bp across all assemblies, reflecting high-quality data. The smallest transcript measured 272 bp, and the longest reached 35,440 bp. BUSCO analysis indicated high completeness, with 89% read support value in read mapping. These results demonstrate the successful generation of a high-quality RNA-seq dataset for *Perna viridis*, with comprehensive assembly metrics provided in Table 2 and BUSCO results in Table 3.

Protein prediction and annotation. In the *Perna viridis* transcriptome, a total of 438,842 assembled transcripts were analysed for their open reading frames (ORFs). TransDecoder analysis revealed that 74.2% of the transcripts contained complete ORFs. Additionally, 12.4% of the transcripts had 5' partial ORFs, 7.3% had 3' partial ORFs, and 6.7% were classified as internal ORFs. The predicted proteins had an average minimum length of 85 amino acids (aa) and a mean length of 402 aa, with the longest proteins reaching up to 14,408 aa. On average, 73% of ORFs were complete, including both START and STOP codons.

A total of 63,369 transcripts with ORF were obtained from the *de novo* transcriptome assembly. Of these, 23,614 transcripts (37.3%) were successfully annotated with Gene Ontology (GO) terms, providing insights into their potential biological functions. Enzyme Commission (EC) codes were assigned to 11,472 transcripts (18.1%), indicating their roles in various biochemical pathways.

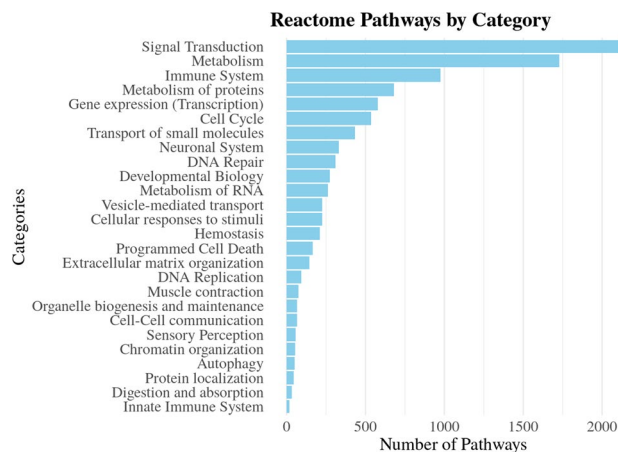


Fig. 3 REACTOME Pathway Annotations. Pathway annotations based on REACTOME, showing key biological processes.

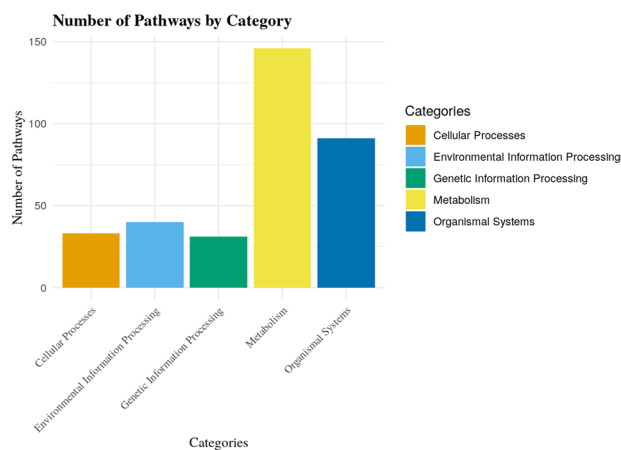


Fig. 4 KEGG Pathway Annotations. Pathway annotations using KEGG illustrate the pathways involved.

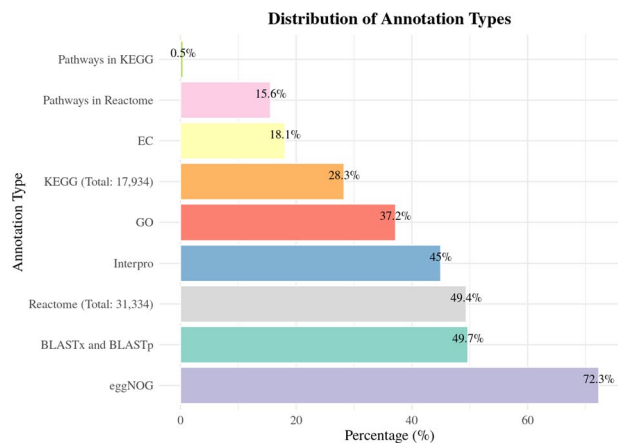


Fig. 5 Annotation Distribution. Distribution of functional annotations across the assembled transcripts, highlighting different categories.

BLAST annotations were achieved for 31,535 transcripts (49.8%), highlighting significant homology with known sequences. Additionally, 45,819 transcripts (72.3%) were annotated using EggNOG, a database for orthology prediction and functional annotation, which helped to further classify the transcripts based on evolutionary relationships and functional domains. InterProScan analysis yielded annotations for 28597 unigenes (45%).

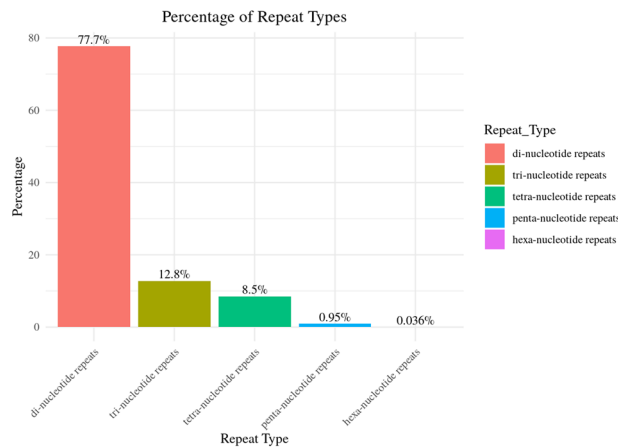


Fig. 6 SSR Repeat Distribution. Distribution and frequency of simple sequence repeats (SSRs) in the transcriptome.

In pathway analysis, 31,334 transcripts (49.4%) were linked to Reactome pathways, identifying a total of 9,923 distinct pathways (Fig. 3). Moreover, 17,934 transcripts (28.3%) were associated with KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways, resulting in the identification of 341 distinct KEGG pathways (Fig. 4). These annotations provide a comprehensive understanding of the molecular and biological functions of the identified transcripts, as well as their involvement in various cellular and metabolic pathways. Figure 5 represents the percentage of annotations. Single sequence repeats identified in transcriptome assembly are indicated in Fig. 6.

Code availability

No custom code was generated in this study. All software programs utilised in this study for *de novo* transcriptome assembly, pre- and post-assembly steps, and transcriptome annotation are listed below with their respective versions. Details on any non-default parameters used are also provided.

- 1 **FastQC** v0.11.8³².
- 2 **MultiQC** v1.12³³.
- 3 **fastp** v1.12³⁴ parameters: SLIDINGWINDOW:4:20, LEADING:5, TRAILING:5, MINLEN:25.
- 4 **Trinity** v2.11⁵.
- 5 **Kraken 2**³⁸.
- 6 **NCBI FCS**³⁹.
- 7 **CD-HIT-est** v. 4.8.1⁴⁰.
- 8 **STAR** v2.7.8a⁴⁴ (--alignIntronMax 1).
- 9 **Bowtie 2**⁴⁶.
- 10 **TransRate** v1.0.3³⁷.
- 11 **BUSCO** v5.4.4³⁶ parameters: dataset mollusca_odb10, Metazoa_odb, and Eukaryota⁴³.
- 12 **TransDecoder** v5.7.0⁴², parameters: default (open reading frame > 100 amino acids).
- 13 **BLAST 2.14.0**⁵⁰.
- 14 **UniProt**: (<http://www.uniprot.org/help/uniprotkb>).
- 15 **eggNOG-mapper** v2.1.12⁴⁶, parameters: -m diamond -evalue 0.001 -score 60 -pident 40 -query_cover 20 -subject_cover 20 -itype proteins -tax_scope auto -target_orthologs all -go_evidence all -pfam_realgn none -report_orthologs -decorate_gf yes -excel.
- 16 **Blast2GO v 5.2.5**⁴⁹.
- 17 **GMATA V 2**⁵⁵.

Received: 19 September 2024; Accepted: 17 January 2025;

Published online: 25 January 2025

References

1. Regan, T. *et al.* Ancestral Physical Stress and Later Immune Gene Family Expansions Shaped Bivalve Mollusc Evolution. *Genome Biol. Evol.* **13**, evab177 (2021).
2. Esposito, G., Pastorino, P. & Prearo, M. Environmental Stressors and Pathology of Marine Molluscs. *J. Mar. Sci. Eng.* **10**, 313 (2022).
3. Ma, Z., Fu, Z., Yang, J. & Yu, G. Combined Effects of Temperature and Salinity Affect the Surviv-AI of Asian Green Mussel (*Perna viridis*) through Digestive and Antioxidant Performance. *Antioxidants* **11** (2022).
4. Peteiro, L. G. *et al.* Responses to salinity stress in bivalves: Evidence of ontogenetic changes in energetic physiology on *Cerastoderma edule*. *Sci. Rep.* **8**, 8329 (2018).
5. Gallardo-Escárate, C. *et al.* Chromosome-Level Genome Assembly of the Blue Mussel *Mytilus chilensis* Reveals Molecular Signatures Facing the Marine Environment. *Genes* **14**, 876 (2023).
6. Yang, J.-L. *et al.* Chromosome-level genome assembly of the hard-shelled mussel *Mytilus coruscus*, a widely distributed species from the temperate areas of East Asia. *GigaScience* **10**, giab024 (2021).
7. Romero Picazo, D., Werner, A., Dagan, T. & Kupczok, A. Pangenome Evolution in Environmentally Transmitted Symbionts of Deep-Sea Mussels Is Governed by Vertical Inheritance. *Genome Biol. Evol.* **14**, evac098 (2022).

8. Cmfri, K. CMFRI Annual Report 2022. <https://eprints.cmfri.org.in/17732/> (2023).
9. Laxmilatha, P. Review of the Green Mussel *Perna viridis* Fishery of South West Coast of India. *Indian J. Mar. Sci.* **3**, 408–416 (2013).
10. Gomes-dos-Santos, A. *et al.* The gill transcriptome of threatened European freshwater mussels. *Sci. Data* **9**, 494 (2022).
11. Shamal, P. *et al.* *Perkinsus olsenii* in the short neck yellow clam, *Paphia malabarica* (Chemnitz, 1782) from the southwest coast of India. *J. Invertebr. Pathol.* **159**, 113–120 (2018).
12. Tanguy, A., Guo, X. & Ford, S. E. Discovery of genes expressed in response to *Perkinsus marinus* challenge in Eastern (*Crassostrea virginica*) and Pacific (*C. gigas*) oysters. *Gene* **338**, 121–131 (2004).
13. Azmi, N., Mazlan, A. G. & Zaidi, C. C. Apicomplexa-like parasites of economically important bivalves from Merambong shoals, Johor. *Malay. Nat. J.* **66**, 108–120 (2014).
14. Trevisan, R. *et al.* Gills are an initial target of zinc oxide nanoparticles in oysters *Crassostrea gigas*, leading to mitochondrial disruption and oxidative stress. *Aquat. Toxicol.* **153**, 27–38 (2014).
15. Yung, M. M. N., Mouneyrac, C. & Leung, K. M. Y. Ecotoxicity of Zinc Oxide Nanoparticles in the Marine Environment. in *Encyclopedia of Nanotechnology* (ed. Bhushan, B.) 1–17, https://doi.org/10.1007/978-94-007-6178-0_100970-1 (Springer Netherlands, Dordrecht, 2014).
16. Rahim, N. F. & Yaqin, K. Histological Alteration of Green Mussel *Perna viridis* Organs Exposed to Microplastics. *Squalen Bull. Mar. Fish. Postharvest Biotechnol.* **17**, 44–53 (2022).
17. Guo, F. *et al.* Genetic Diversity, Population Structure, and Environmental Adaptation Signatures of Chinese Coastal Hard-Shell Mussel *Mytilus coruscus* Revealed by Whole-Genome Sequencing. *Int. J. Mol. Sci.* **24**, 13641 (2023).
18. Kilikowska, A. *et al.* The Patterns and Puzzles of Genetic Diversity of Endangered Freshwater Mussel *Unio crassus* Philipsson, 1788 Populations from Vistula and Neman Drainages (Eastern Central Europe). *Life* **10**, 119 (2020).
19. Österling, M., Lopes-Lima, M., Froufe, E., Hadzihalilovic, A. H. & Arvidsson, B. The genetic diversity and differentiation of mussels with complex life cycles and relations to host fish migratory traits and densities. *Sci. Rep.* **10**, 17435 (2020).
20. Leung, P. T. *et al.* *De novo* transcriptome analysis of *Perna viridis* highlights tissue-specific patterns for environmental studies. *BMC Genomics* **15**, 804 (2014).
21. Zhang, X. *et al.* *De novo* assembly and comparative transcriptome analysis of the foot from Chinese green mussel (*Perna viridis*) in response to cadmium stimulation. *PLoS ONE* **12**, e0176677 (2017).
22. Inoue, K. *et al.* Genomics and transcriptomics of the green mussel explain the durability of its byssus. *Sci. Rep.* **11**, 5992 (2021).
23. Raghavan, V., Kraft, L., Mesny, F. & Rigerte, L. A simple guide to *de novo* transcriptome assembly and annotation. *Brief. Bioinform.* **23**, bbab563 (2022).
24. Sukumaran, S. *et al.* The chromosome level genome assembly of the Asian green mussel, *Perna viridis*. *Sci. Data* **11**, 930 (2024).
25. Benoit-Pilven, C. *et al.* Complementarity of assembly-first and mapping-first approaches for alternative splicing annotation and differential analysis from RNAseq data. *Sci. Rep.* **8**, 4307 (2018).
26. Deschamps-Francoeur, G., Simoneau, J. & Scott, M. S. Handling multi-mapped reads in RNA-seq. *Comput. Struct. Biotechnol. J.* **18**, 1569–1576 (2020).
27. Hölzer, M. & Marz, M. *De novo* transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. *GigaScience* **8**, giz039 (2019).
28. Burns, P. D., Li, Y., Ma, J. & Borodovsky, M. UnSplicer: mapping spliced RNA-Seq reads in compact genomes and filtering noisy splicing. *Nucleic Acids Res.* **42**, e25 (2014).
29. Saco, A., Rey-Campos, M., Novoa, B. & Figueras, A. Transcriptomic Response of Mussel Gills After a *Vibrio splendidus* Infection Demonstrates Their Role in the Immune Response. *Front. Immunol.* **11** (2020).
30. Marisa, I. *et al.* *In vivo* exposure of the marine clam *Ruditapes philippinarum* to zinc oxide nanoparticles: responses in gills, digestive gland and haemolymph. *Environ. Sci. Pollut. Res.* **23**, 15275–15293 (2016).
31. Sun, Z., Lou, F., Zhang, Y. & Song, N. Gill Transcriptome Sequencing and *De Novo* Annotation of *Acanthogobius ommaturus* in Response to Salinity Stress. *Genes* **11**, 631 (2020).
32. Andrews, S. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data (2010).
33. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarise analysis results for multiple tools and samples in a single report. *Bioinforma. Oxf. Engl.* **32**, 3047–3048 (2016).
34. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinforma. Oxf. Engl.* **34**, i884–i890 (2018).
35. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
36. Seppy, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol. Biol. Clifton NJ* **1962**, 227–245 (2019).
37. Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M. & Kelly, S. TransRate: reference-free quality assessment of *de novo* transcriptome assemblies. *Genome Res.* **26**, 1134–1144 (2016).
38. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
39. Astashyn, A. *et al.* Rapid and sensitive detection of genome contamination at scale with FCS-GX. *Genome Biol.* **25**, 60 (2024).
40. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinforma. Oxf. Engl.* **22**, 1658–1659 (2006).
41. Davidson, N. M., Hawkins, A. D. K. & Oshlack, A. SuperTranscripts: a data driven reference for analysis and visualisation of transcriptomes. *Genome Biol.* **18**, 148 (2017).
42. Haas, B. J. *et al.* *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
43. Kriventseva, E. V. *et al.* OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **47**, D807–D811 (2019).
44. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinforma. Oxf. Engl.* **29**, 15–21 (2013).
45. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
46. Huerta-Cepas, J. *et al.* eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).
47. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).
48. Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **46**, D649–D655 (2018).
49. Conesa, A. *et al.* Blast2GO: a universal tool for annotation, visualisation and analysis in functional genomics research. *Bioinformatics* **21**, 3674–3676 (2005).
50. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
51. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
52. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
53. Heid, E., Probst, D., Green, W. H. & Madsen, G. K. H. EnzymeMap: curation, validation and data-driven prediction of enzymatic reactions. *Chem. Sci.* **14**, 14229–14242.
54. Zhou, Y. *et al.* Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* **10**, 1523 (2019).

55. Wang, X. & Wang, L. GMATA: An Integrated Software Package for Genome-Scale SSR Mining, Marker Development and Viewing. *Front. Plant Sci.* **7** (2016).
56. Vysakh, G. *et al.* *Perna_v_gill_vysakh*, Gill-specific *de novo* transcriptome assembly of the Asian green mussel, *Perna viridis*: a novel invertebrate model for ecotoxicology and cancer studies. *NCBI GenBank* <https://identifiers.org/ncbi/insdc:GKYN00000000> (2024).
57. *NCBI Sequence Read Archive* <https://identifiers.org/insdc.sra:SRR30565980> (2024).
58. *NCBI Sequence Read Archive* <https://identifiers.org/insdc.sra:SRR30565978> (2024).
59. *NCBI Sequence Read Archive* <https://identifiers.org/insdc.sra:SRR30565981> (2024).
60. *NCBI Sequence Read Archive* <https://identifiers.org/insdc.sra:SRR30565979> (2024).
61. *NCBI Sequence Read Archive* <https://identifiers.org/insdc.sra:SRR30565982> (2024).
62. *NCBI Sequence Read Archive* <https://identifiers.org/insdc.sra:SRR30565983> (2024).
63. *NCBI Sequence Read Archive* <https://identifiers.org/insdc.sra:SRR30565984> (2024).
64. *NCBI Sequence Read Archive* <https://identifiers.org/insdc.sra:SRR30565985> (2024).
65. Vysakh, V. G. *et al.* Gill-Specific *De Novo* Transcriptome Assembly of the Asian green mussel, *Perna viridis*: A Novel Invertebrate Model for Ecotoxicology and Cancer Studies. *Figshare* <https://doi.org/10.6084/m9.figshare.27054268>.
66. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

Acknowledgements

The authors wish to express their sincere gratitude to Dr. Grinson George, Director of ICAR-CMFRI, Kochi, and Dr. Kajal Chakraborty, Head of the Division, for facilitating access to essential research resources. We extend our gratitude to Dr. Binoy, Manu VK and P. R. Abhilash for their valuable support during this research. This work is part of the Ph.D. thesis of VGV, who acknowledges the financial support from the UGC, India, for the fellowship received during this research. We are also grateful to acknowledge the MYFISH server facility for providing high-performance computing resources critical to our projects.

Author contributions

V.V.G. conducted sample collection, data processing, analysis and drafted the manuscript. S.S. & V.V.G. conceived, designed, and organised the project. A.G. provided critical revisions to the manuscript. All authors contributed to the final manuscript preparation and approved the submitted version.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025