Turk. J. Fish.& Aquat. Sci. 26(3), TRJFAS26728
https://doi.org/10.4194/TRJFAS26728

PROOF
RESEARCH PAPER

Turkish Journal of FISHERIES and AQUATIC SCIENCES

Comparative Transcriptome Sequencing Provides a Reference Resource for Expressed Genes and Associated Markers in Giant Tiger Prawn, *Penaeus monodon* Muscle Tissue

Vindhya Mohindra[1,*], Labrechai Mog Chowdhury[1], Nishita Chauhan[1], Neha Shukla[1], Alisha Paul[1], V. S. Basheer[2], Santosh Kumar[1], Joykrushna Jena[3]

[1]ICAR-National Bureau of Fish Genetic Resources, Canal Ring Road, Dilkusha, Lucknow- 226002, India.
[2]ICAR-National Bureau of Fish Genetic Resources, Centre for Peninsular Aquatic Genetic Resources, CMFRI Campus, Kochi–682 018, Kerala.
[3]Indian Council of Agricultural Research (ICAR), Krishi Anusandhan Bhawan - II, New Delhi, 110 012, India.

How to Cite

Mohindra, V., Chowdhury, L.M., Chauhan, N., Shukla, N., Paul, A., Basheer, V.S., Kumar, S., Jena, J. (2026). Comparative Transcriptome Sequencing Provides a Reference Resource for Expressed Genes and Associated Markers in Giant Tiger Prawn, *Penaeus monodon* Muscle Tissue. *Turkish Journal of Fisheries and Aquatic Sciences*, *26(3)*, *TRJFAS26728.* https://doi.org/10.4194/TRJFAS26728

Article History
Received 06 September 2024
Accepted 28 July 2025
First Online 22 August 2025

Corresponding Author
E-mail: vindhyamohindra@gmail.com

Keywords
Differentially expressed genes
Gene ontology
Glycolysis pathway
Simple sequence repeats
Single nucleotide polymorphism

Abstract

The shrimp industry has seen remarkable growth over the last four decades in India and the culture of P. *monodon* is experiencing a resurgence due to shifts in market demand, ecological considerations and broodstock and farming management. This study employs a transcriptomics approach to investigate the gene expression profiles of *Penaeus monodon* from two different origins. Large-scale transcriptomes generated 130,684 super-transcripts (GC 39.18%), with 92.94% BUSCO completeness and 39,974 have functional annotations, and 125 GO terms were identified. A total of 41,108 annotated transcripts examined, 19,637 SSR were identified and SNP mining resulted in identification of 69,028 SNPs with 26,785 indels, which included SNP within the genes of known function were found to be 504 nonsynonymous and 176 synonymous. Differentially expressed 79 genes between two locations fell under 12 GO terms, of which 4 GO terms were under biological process with 5 genes under glycolytic process and 4 under gluconeogenesis. Thus, the cSSRs and cSNP associated with coding genes identified could be a part of the marker set, which would be useful as marker loci or candidate genes to be use in construction of smaller SNP arrays for pedigree analysis and/or genomic selections.

Introduction

The aquaculture of shrimps and prawns has made significant advances during the last four decades in many parts of the world, including India. In the year 2023, the world produced around 5.6 million tons (mmt) of farmed shrimp and projected for 7.28 million tons in 2025 (Rossi et al., 2024). Majority of overall export of shrimp from India (nearly 84%) is contributed by the single exotic species, *Litopenaeus vannamei* (white leg shrimp) and the production traditional black tiger shrimp (*Penaeus monodon*) reduced to 27 600 tonnes in 2020-21 (FAO, 2024). However, India is witnessing a notable resurgence *of P. monodon,* due to a strategic shift driven by market demands, ecological considerations, availability of improved broodstocks and advancements in farming techniques, which offer cost-effective production and larger sizes (https://www.fao.org/fishery/en/facp/ind?lang=es). However, challenges of management of diseases and potential negative consequences of environmental damages are still cause of concern for the shrimp farming industry.

Most of the hatchery operators of *P. monodon* still prefer to use wild-caught broodstock (Nguyen et. al., 2025), which has led to the heavy demand for female broodstock from the wild. Thus, development of management strategies for *P. monodon* exploited wild

populations becomes a necessity, based on genetic stock structure information through molecular markers. In recent years, the genetic diversity and population structure of *P.monodon* were assessed with single-nucleotide polymorphisms (SNPs) (Vu et al. 2020) and microsatellites (Wong et al. 2021). It has also been reported that disease resistance and adaptability of the natural populations are affected by their level of genetic diversity (Li et. al., 2024). Thus, the molecular markers with their potential associations with a specific trait becomes a marker of choice, that would facilitate the genetic diversity studies, as well as the selective breeding programs.

With the rapid development of high-throughput sequencing technology, genes and transcriptional responses associated with traits have been identified using RNA-sequencing (RNA-seq) techniques in *P. monodon*, for hypotonic stress (Ji et. al.,2024), immune development in life stages (Angthong et. al.,2021), in early and late-vitellogenic broodstock females (Nguyen et. al., 2025) and response to acute hepatopancreatic necrosis disease (Soo et.al., 2019). Although gene expression variation within species is relatively common, the gene expression changes and pathways underlying the physiological adaptation in different locations with different environmental conditions is poorly understood. However, molecular markers associated with the traits and particularly for the stress responses related to the environmental variations in *P. monodon* is scarce.

Over the past few decades, genome-wide molecular markers, including SNPs, have recently become the predominant molecular markers utilized in aquaculture research (Zenger et al., 2019). However, nowadays, smaller informative marker subsets in SNP panels are being preferred for fast, vigorous, and lower cost for diversity analysis (Sukhavachana et al., 2021, Reverter et al., 2020) for commercial applications. Flanagan and Jones (2019) concluded that a relatively small number of markers could completely resolve pedigree in most situations. Studies have revealed the functional genes to be controlling the major developmental processes and the associated functional genetic variants are expected to have effects on phenotypes, and selection of these variants depends on the appropriate functional annotation of the data generated (Johnston et. al., 2024). Thus, incorporating the information of genetic functional variants in SNP panels, which are prioritized, have potential to increase accuracy of selection (Xiang et al., 2021) as compared to randomly selected SNPs.

The objective of the present study was to generate a reference genomic resource in the form of functional annotation transcriptome datasets from *P. monodon* specimens from diverse location and environmental conditions, representing muscle tissues and identification of functional variants (genes) and their associated molecular markers, for marker assisted breeding programmes.

Materials and Methods

Sample Collection

Penaeus monodon samples were collected from two grow-out commercial farms, at diverse locations, cultivating offspring from wild populations: population 1. collected and cultivating at Vypin, Kochi, Kerala (N 9° 99815', E 76° 23795') and 2. collected from Bay of Bengal and cultivating at Purba Medinipur, West Bengal (N 22° 05698', E 88° 140408') (Supplementary Figure 1). Specimens were immediately euthanized with MS222 (Sigma Aldrich, USA) and muscle tissues were collected and snap frozen, and stored in liquid nitrogen at the site of collection. Later, these were stored at -80°C in the laboratory, till use.

RNA Extraction, Transcriptome Sequencing and Assembly

The total RNA was extracted using TRIzol Reagent (Thermo Fisher Scientific, United States) from muscle tissues from 11 samples (6 from population 1 and 5 from population 2), and DNaseI (NEB, Massachusetts, USA) was added to prevent genomic DNA contamination. The total RNA of each sample was quantified and qualified by Agilent 2100 Bioanalyzer (Agilent Technologies, USA), and the integrity were checked through 1% agarose gel, followed by purification using RNA Clean XP (Beckman Coulter & Pasadena, USA). One µg total RNA with RIN >7 (Agilent Tapestation; Agilent Technologies, USA) was used for individual sample library preparation, using Illumina Paired end library sequencing kit (NEBNext® UltraTM RNA Library Prep Kit for Illumina®) and transcriptome sequencing was carried out using Illumina HiSeq2500 (Illumina Inc., USA). The concentration of the prepared library was tested by using Qubit 2.0 (Invitrogen) and the library was check for size distribution Agilent 2100 and sequenced with HiSeq 2500, according to the manufacturer's instructions. The paired end reads were filtered and trimmed with the minimum quality of Q15 Qphred=-10$\log_{10}$(e), using fastp v0.19.4 (Chen et al., 2018).

The sequence data from all samples were pooled together, to digitally normalize the data using BBNorm (https://genome.jgi.doe.gov) at K=25 and the transcriptome was assembled using software Trinity (Grabherr et. al., 2012). The assemblies were concatenated using CD-HIT-EST version 4.6 (Huang et. al., 2010) and the transcripts <90 bp in length were filtered using the Evidential Gene package (Gilbert, 2016). The final non-redundant pool from Trinity assembled transcripts were merged into a Super-Transcript using Trinity's internal module, by collapsing common and unique regions, amongst the isoforms into a linear representation of the sequence. For completeness assessment of the assembled transcript, gVolante (Nishimura et al., 2019) was used against

Arthropod and Eukaryote database, using BUSCO v5 orthology pipeline.

Long Non-coding RNAs (lncRNAs) Prediction

Long non-coding RNAs (lncRNAs) of length ≥200 bp were identified using coding potential calculator (CPC, Kang et al., 2017, http://cpc.cbi.pku.edu.cn/) and CPC2 (http://cpc2.cbi.pku.edu.cn/), with the significant value of less than $e10^{-04}$. LncRNAs thus identified were removed from the assembled transcripts, to improve the transcripts dataset with protein-coding potential.

Functional Annotation and Characterization

For further study, assembled transcripts were functionally annotated and GO classified, using Omics Box v.2.0.36 (Gotz et al., 2008, Omics Box 2019), using BLASTx against Reference protein (RefSeq-protein v5) database, with selected species and e-value $\leq 10^{-5}$, similarity score >40%. The gene ontology (GO) terms associated with the annotated genes were carried out using DAVID v6.8 (https://david.ncifcrf.gov/).Molecular pathways were identified through KAAS, P<0.05 (KEGG Automatic Annotation Server, Moriya, et al., 2007) database.

Differential Gene Expression Estimation

The filtered transcripts were indexed using Salmon 19 tool (Patro et. al., 2017) in order to map the reads and to derive the respective transcript counts by using quasi-alignment approach. Subsequently, the transcript counts were summarized to a gene-level, imported into R using the bioconductor package tximport by creating a Transcript-gene map (Supplementary Figure 2). The counts are rounded off to the closest integers and were used for further analyses in DESeq2 (Love et al., 2014). The 'regularized log' transformation in DESeq2 was used for principal component and clustering analysis. The DESeq2 result files were filtered with cut off $\log_2$FC (foldchange) >1 and P<0.05. Prior to statistical analysis, normalization of counts were carried out by library size (DESeq2: estimateSizeFactors) and gene-wise dispersion (DESeq2: estimateDispersions). This normalized the gene counts by gene-wise geometric mean over samples.

Protein-Protein Interaction (PPI) Analysis for Differentially Expressed Genes

For identification of protein-protein interaction between the differentially expressed genes, STRING (Szklarczyk et al., 2019), a biological database for protein-protein interaction prediction study was used, with reference organism *Drosophila melanogaster*. The differentially expressed genes (DEGs) found in two different populations were further studied for functional identification of interacting proteins with interaction scores and medium confidence (0.040).

Simple Sequence Repeats (SSRs) Identification

The SSRs were identified using PERF (Avvaru et al., 2018) and MISA (Beier et al., 2017) from the super-transcripts. Compound SSRs were identified using MISA alone, whereas other SSRs were filtered based on concordance using a 90% reciprocal intersection between the two populations. Primer sequence designed for validation of SSR in differentially expressed genes-SSR in two different populations of *P. monodon* is given in Supplementary Table 1.

Pcr Amplification and Genotyping of Ssr

Multiple primers sets were designed and synthesised per transcript, for testing a particular SSR primer set for success of amplifications, initial amplifications were done using sample size of 2/per population, that is total 4 individuals.

Primers were designed for microsatellite repeats, with sufficient flanking sequences using Primer 3.0 tool (Untergasser et al., 2012) and predictive target sequence length between 100 to 200 bp were selected. Multiple primers sets were designed and synthesised per transcript, for testing a particular SSR primer set for success of amplifications, initial amplifications were done using sample size of 2/per population, that is total 4 individuals, using traditional PCR amplification. The PCR amplification was carried out in 25 µl reaction volume, contained approximate 50 ng of DNA, 1X PCR buffer, 25 mM dNTPs, 1.5 mM $MgCl_2$, 5 pM primers (2.5 pM each forward and reverse primer) and 1.5 U Taq polymerase (Genei, India). Amplification conditions were: denaturation at 95°C for 5 min, followed by 30 cycles of 95°C for 30 sec, primer-specific annealing temperature (T_a) for 30 sec, 72°C for 1.5 min, and final extension at 72°C for 10 min and final cooling at 4°C. PCR amplicons were resolved in 10% native polyacrylamide gels (10×10.5 cm, Wipro GE Healthcare Pvt. Ltd., India) using 1XTBE buffer for about 5 h at 150 V and the bands were visualised through silver staining. Molecular sizing of alleles was done on UVP Imaging System (Cambridge, UK), using MspI digested PBR-322 as a ladder.

The successful primers were then screened for polymorphism in twelve individuals from two different locations, ie., total 24 individuals were genotyped using automatic QIAxcel Capillary Electrophoresis System with QIAxcel DNA high-resolution kit (Qiagen, Hilden, Germany) with 25- 500 bp size marker and 15 bp 600 bp alignment marker in Qiaxcel screenGel software v1.6.0. OM500 was the method selected for running the samples. These amplified fragments were automatically scored using QIAxcel ScreenGel software.

Snp Variants Analyses

Reads were mapped for all the 11 samples individually using a dual-pass STAR alignment mode (Dobin et al., 2013), option --outSJfilterOverhangMin 12 12 12 to filter the output splice junctions and read alignments were visualised using Gviz (Hahne and, Ivanek, 2016). Variants were called using HaplotypeCaller 4.1.4.1 from GATK4 (Van 2020). These variants were subsequently filtered using a distance-based approach (35 bp between variants) to exclude mismappings or sequencing artifacts; using Fisher Strand score >30.0 to avoid strand bias; Qual by Depth (QD)>2.0 for a reliable call and finally homozygous variants were excluded. Passed variants were then screened amongst populations to be present in at least 2 samples. SNPs were annotated using the SuperTranscripts GTF (General Transfer Format) using SNPDat 1.0.4 (Doran and Creevey, 2013).

Results

Pre-processing and De Novo Assembly of Transcriptome

A total of 518.5 and 444.8 million Illumina sequencing raw reads were generated for population 1 and 2, respectively, of *P. monodon*, with a total 77.77 Gb (11.0 to 13.66 Gb) to 66.73 Gb (11.55-16.4 Gb) raw bases. After removal of adapters and rRNA, a total of 9.56 to 14.26 GB Q30 filtered bases were assembled into a total of 135,654 clustered transcripts (Supplementary Table 2). Further filtering resulted into final 130,684 super-transcripts with GC 39.18% and longest transcript length was 48,502 bp (Supplementary Table 3). NCBI accessions for SRA and transcriptome are given in the Data Availability section. The length distribution statistics of super-transcripts showed that the sequences of length ranging from 300-350 bp were maximum (21,621 sequences), followed 350-400 bp transcripts with 19,708 sequences (Supplementary Figure 2). Out of 13,0684 super-transcripts, after removal of 89,576 Long Noncoding RNAs (lncRNAs) sequences identified and 41,108 transcripts with coding potential were further analysed. BUSCO analysis revealed 88.55% and 92.94% transcriptome completeness with Arthropod and Eukaryota databases, respectively (Figure 1, Supplementary Table 4).

Functional Annotation of Transcriptome

Out of 41,108 transcripts, 97.24% (39974) had functional annotations, while 3.15% (1297) with significant coding potential. A total of 22,389 transcripts has protein annotation, with 18242 hit accessions, 9,373 GO terms and 35,123 InterPro terms (Table 1). Species distribution of the annotation of transcripts showed the maximum number of BLAST hits (10,156) were with *P. monodon,* followed by *P. vannamei* (3,278) (Supplementary Figure 3).

Gene Ontology terms identification through DAVID tool identified 125 GO terms, out of which 41 terms were under biological process with maximum number of 87 genes under translation (GO:0006412) followed by 55 genes under intracellular protein transport (GO:0006886), and 39 terms under cellular component, with maximum number of 898 genes nucleus (GO:0005634), followed by 698 genes under cytoplasm (GO:0005737), and 45 terms under molecular function, with maximum number of 840 genes under protein binding (GO:0005515), followed by 447 genes under ATP binding (GO:0005524) (Supplementary Figure 4). Molecular pathways analysis identified 6,569 genes involved in Organismal Systems, followed by Metabolism (4520), Environmental Information Processing (3328), Cellular Processes (3226) and Genetic Information Processing (2289).

Identification and Enrichment of Genes with Differential Expression

A total of 79 genes were found to have significant differential expression between populations out of which, 65 genes were found to be down-regulated in

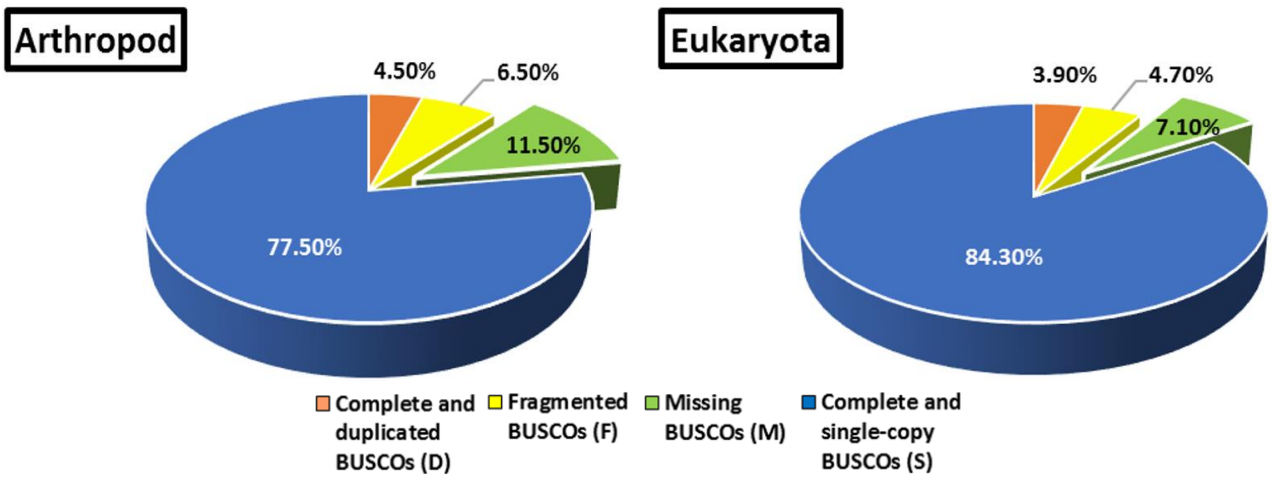

Figure 1. Completeness Assessment of assembled muscle transcriptome of *Penaeus monodon*.

population 2. Gene Ontology terms identification identified 12 GO terms out, of which 4 G0 terms were under biological process with maximum number of 5 genes under glycolytic process, followed by 4 genes under gluconeogenesis and 6 GO terms under cellular component, with maximum number of 6 genes under cytoplasm and cytosol followed by 4 genes under membrane and nucleus, and 1 GO term under molecular function with 3 genes in ATP binding (Figure 2, Supplementary Table 5).

Molecular pathways analysis for the differentially expressed genes identified 10 KEGG pathways (FDR<0.05), out of which the critical pathways were Glycolysis / Gluconeogenesis (10 genes, Figure 3) under Metabolism, Glucagon signalling pathway (5) under Organismal Systems and HIF-1 signalling pathway (4)

Table 1. Summary of functional annotation of assembled muscle transcriptome of *Penaeus monodon*

Descriptors	Number	Percent
Total super-transcripts assembled	130684	
Number of long noncoding transcripts	89576	
Number of transcripts (coding)	41108	
Number of transcripts with Functional annotation	39974	97.24%
· with Refseq database	22389	
· with GO Terms	9373	
· with Interpro Terms	35123	
· with KO terms	12618	
· Annotated with blast2go	843	
Unannotated transcripts with coding potential	1297	3.15%

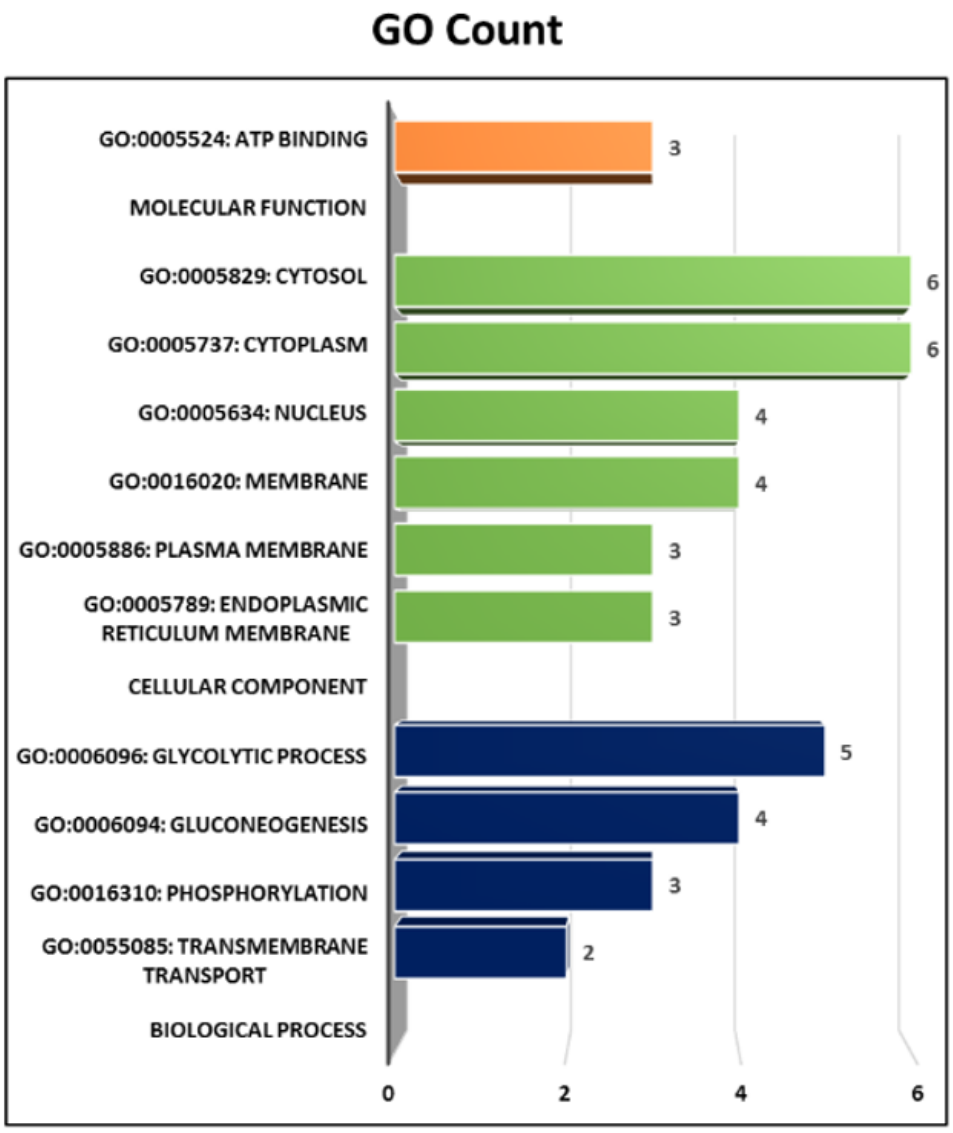

Figure 2. Gene Ontology terms identified in differentially expressed genes of two different populations of *Penaeus monodon*.

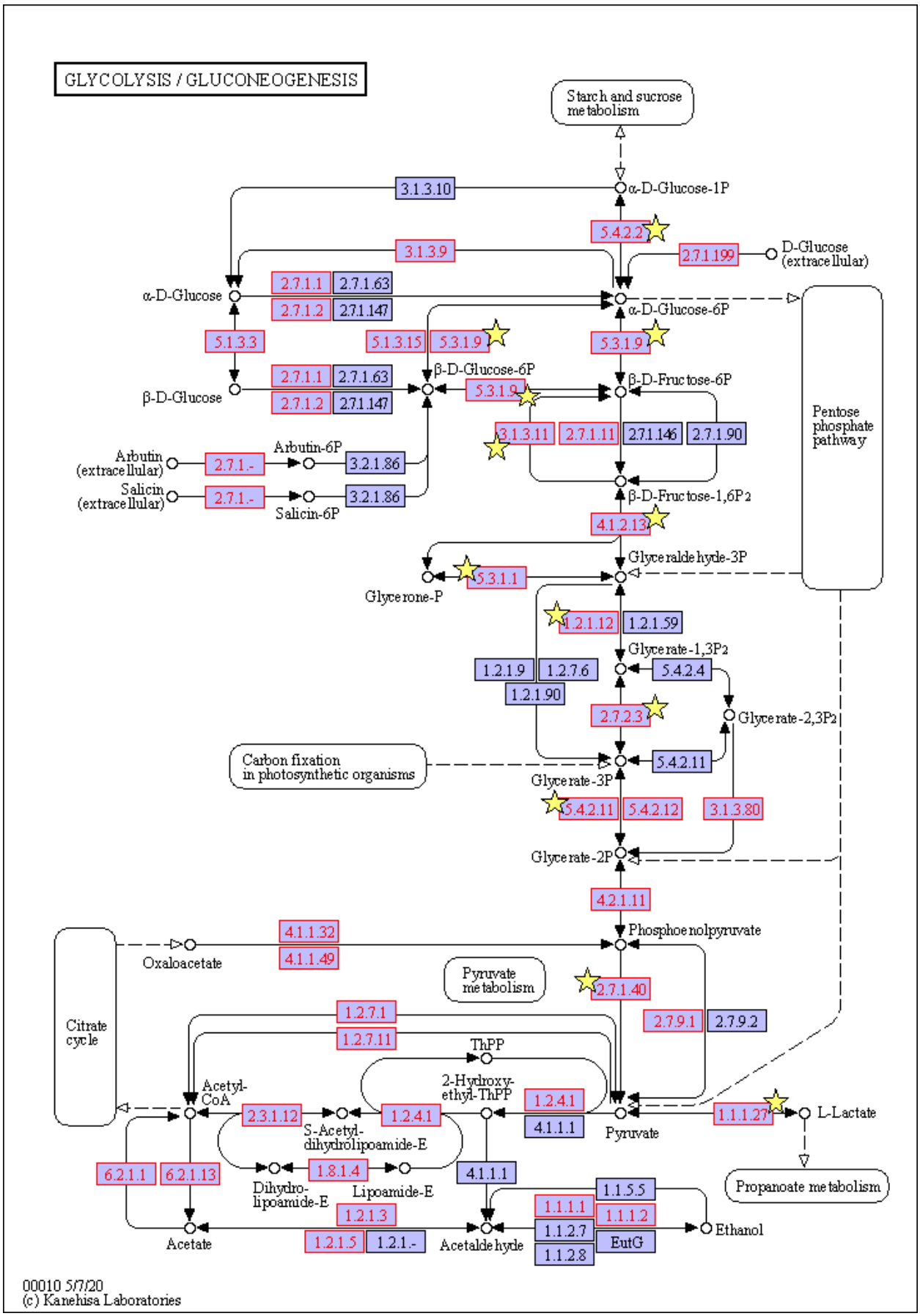

Figure 3. Glycolysis/Gluconeogenesis signalling pathway, depicting differentially expressed genes of two different populations of *Penaeus monodon*. Figure made from Kyoto Encyclopedia of Genes and Genomes, Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. & Tanabe, M. New approach for understanding genome variations in KEGG. Nucleic Acids Res. 47, D590–D595 (2019).

under Environmental Information Processing. Other significant pathways were cAMP signalling pathway, MAPK signalling pathway and Pentose phosphate pathway (Supplementary Table 6).

Protein-Protein Interaction Network Construction

The String protein-protein interaction (PPI) analysis revealed that out of 85 DEGs, 35 genes were mapped against *Drosophila melanogaster* database. The PPI network among these mapped genes showed 35 nodes and 53 edges (PPI enrichment $P<1.0e^{-16}$) and 106 interactions among 25 proteins. Protein-protein interactions showed maximum number of 8 interactions with *mp20, pgi, tpi* followed by 7 genes interactions with *act88f, ald1, ldh, pgk* and 6 interactions with *pgm1, pyk* (Figure 4, Supplementary Table 7). Gene ontology term identification of interacting genes showed 44 GO terms in biological processes with highest number of genes in Cellular process (31), 17 in Cellular component in Cellular anatomical entity (33) and 3 in molecular function with in Binding (26) (FDR<0.05).

The critical KEGG pathways identified under interacting genes showed 4 molecular pathways with maximum number of 7 genes under Glycolysis / Gluconeogenesis followed by Carbon metabolism with 7 genes, Biosynthesis of amino acids with 4 genes and Pentose phosphate pathway with 3 genes.

Hub genes and sub-networks analysed based on genes from the PPI network showed 18 hub genes (Supplementary Figure 5, Supplementary Table 8) to be significant in the five algorithms (Mcc, Mnc, Degree, Epc, Eccentricity), out of which the function analysis of 13 hub genes revealed their categorization into 3 groups of GO's, and a summary of their respective GO and KEGG terms is given in Table 2. Major groups were Group 2: ADP metabolic process with 5 GO and 2 KEGG terms with 5 genes and Group 3: sarcomere with 8 GO and 7 genes (Figure 5 A, B, C).

Ssr İdentification from Transcriptome and Degs

A total of 41,108 transcripts examined, 19,637 SSR were identified, out of which 4,633 sequences contained more than 1 SSR (Table 3). Dinucleotide repeats were the most abundant motifs (51.50%) followed by trinucleotide repeat motifs (31.90%), and other categories are given in Supplementary Figure 6. In the dinucleotide repeats, AG/CT was the most abundant repeat motif, with a frequency of 22.13% (4,785 SSR) and CG/CG the least abundant motif with 0.148% (32 SSR) of total SSRs. The most frequent repeats among the trinucleotide repeat motifs were AAG/CCT (11.2% of total 2438 SSRs), in Quad nucleotide AAAG/CTTT was the most abundant repeat (2.61%) (Supplementary Table 9).

A total of 21 SSRs were identified in 19 differential expressed genes of metabolic pathways between two different populations, of which 7 SSR loci were of compound, 8 dinucleotide (p2) and 6 trinucleotide (p3) (Supplementary Table 10). Out of eight primer pairs tested for genes *wupA, FBP1, GJ16889, Arp3, Unc-89, aop, slbo,* LOC119571649, six loci resulted in amplification, out of which three were polymorphic and three monomorphic (Table 4; Supplementary Table 11, Supplementary Figure 7).

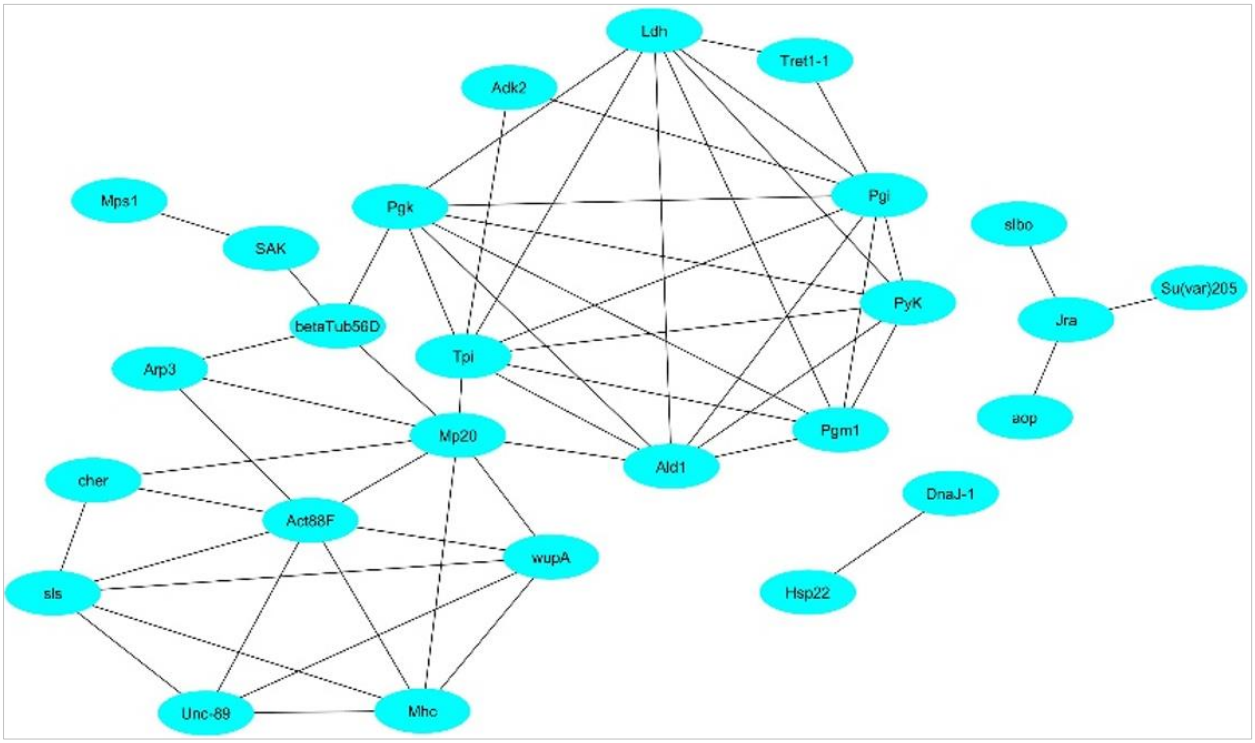

Figure 4. Protein-protein interaction network of differentially expressed genes of two different populations of *Penaeus monodon*.

Table 2. Gene Ontology terms grouping for Hub genes DEGs identified in two different populations of *Penaeus monodon*

Groups	Function	ID	Term	Nr. Genes	Associated Genes Found
Group1	myoblast fusion	GO:0007520	myoblast fusion	2	*arp3, mp20*
Group2	ADP metabolic process	GO:0046031	ADP metabolic process	6	*adk2, impl3, pgi, pgk, pyk, tpi*
		GO:0046034	ATP metabolic process	6	*adk2, impl3, pgi, pgk, pyk, tpi*
		KEGG:00010	Glycolysis / Gluconeogenesis	5	*impl3, pgi, pgk, pyk, tpi*
		GO:0006734	NADH metabolic process	5	*impl3, pgi, pgk, pyk, tpi*
		GO:0061621	canonical glycolysis	4	*pgi, pgk, pyk, tpi*
		KEGG:00620	Pyruvate metabolism	2	*impl3, pyk*
		GO:0042866	pyruvate biosynthetic process	2	*impl3, pyk*
Group3	Sarcomere	GO:0030017	sarcomere	7	*act88f, mhc, pgk, tpi, unc-89, cher, wupa*
		GO:0051146	striated muscle cell differentiation	7	*act88f, arp3, mhc, mp20, unc-89, cher, wupa*
		GO:0015629	actin cytoskeleton	6	*act88f, arp3, mhc, mp20, cher, wupa*
		GO:0030239	myofibril assembly	5	*act88f, mhc, unc-89, cher, wupa*
		GO:0031672	A band	4	*mhc, pgk, tpi, unc-89*
		GO:0030018	Z disc	4	*act88f, pgk, tpi, cher*
		GO:0036379	myofilament	3	*act88f, mhc, wupa*
		GO:0071689	muscle thin filament assembly	2	*act88f, mhc*

Snp Discovery

Uniquely mapped reads were found to be 66.69 to 70.93% (Supplementary Table 12 and supplementary Figure 8) and variant Summary in all the 11 samples individually of two different populations of *Penaeus monodon* muscle transcriptome is given in Supplementary Table 13 (Supplementary Figure 9).

SNP mining resulted in identification of 69,028 SNPs in population 1 and 65,310 in population 2 with 26,785 and 26,612 indels. Transitions were found more frequent (41,314 and 38,659) than transversions (28,885 and 27564) among the identified SNPs in population 1 and population 2, respectively and the Ts:Tv ratio of population 1 and population 2 was 1.43 and 1.40 (Supplementary Table 14).

SNP analysis of annotated genes, having known function found a total of 504 nonsynonymous and 176 synonymous SNP. In case of SNPs in annotated transcripts, frequency of G<A (22.5%; 19.49%) was the highest in transition and the frequency of T<A (7.3%; 9.4%) and was the highest in transversion in population 1 and population 2, respectively (Supplementary Figure 10, Supplementary Table 15). Finally, a total of 31 SNPs was found to be significantly difference (P<0.05) in frequency in both populations (Table 5).

SNP annotated summary of two different populations of *Penaeus monodon* muscle transcriptome is given in Supplementary Table 15. Total number of SNPs annotated were higher in population 1 (70831) than 2 (66766). Similar trend was found in non-synonymous SNPS, as well as Ts/Tv (Supplementary Table 16). SNP substitution frequency of DEG of two different populations are given in Supplementary Table 17**),** which showed higher Ts: Tv for pop 2 than pop1.

SNP annotation of differential expression genes identified total of 62 synonymous SNPs in the exonic region, of which 7 SNP were significant (P<0.05) between two populations (Table 4). The frequency of G>A was highest, representing 17 SNP, followed by C>T (14) and T>C (13) (Figure 6).

Discussion

Transcriptomic studies have been extensively utilized in aquaculture species, to unearth the molecular mechanisms underlined to specific biological processes (Mohindra et al., 2023). Here, the present study, based on the large-scale transcriptomes generated from *P. monodon* of two habitats, represents a robust genomic database of a comprehensive expressed gene complement, which can be utilised for identification of underlying differentially expressed genes.

It was interesting to find from differentially expressed genes, the major pathways including Glycolysis/Gluconeogenesis and Glucagon signalling pathways, with a strong role in increasing the blood glucose, which may be a response to environmental disturbances (Malini et al., 2018). Another major group of genes fall under GO:0051146-striated muscle cell differentiation, which denotes a process in which a relatively unspecialized cell acquires specialized features of a muscle cell (uniport https://www.ebi.ac.uk/QuickGO/term/ GO:0042692). Under this GO, Unc-89 obscurin isoform X3, with 3 interactions in present study, has been reported to be involved in the development of a symmetrical sarcomere (Katzemich et al., 2012). Similarly, differentially expressed Adenylate kinase (AK) is a key enzyme in energy homeostasis (Dzeja et al., 2011) and has been reported to be regulated by a set of master transcription factors (Lane and Fan, 2015). It has also been reported to be stress response gene, functions to maintain cell survival (Jan et. al, 2019), especially to a hypoxic condition in cancer cells (Klepinin et. al. 2020).

In the present study, differentially expressed genes under the HIF-1 signalling pathway point out to the fact that these populations are facing a hypoxic condition under culture conditions. Under the hypoxic conditions, GAPDH acts as a sensor of oxidative stress and redox signals, leading to promotion of angiogenesis (https://www.ncbi.nlm.nih.gov/gene/ 2597). And another DEG, fructose-bisphosphate aldolase (*aldo*)

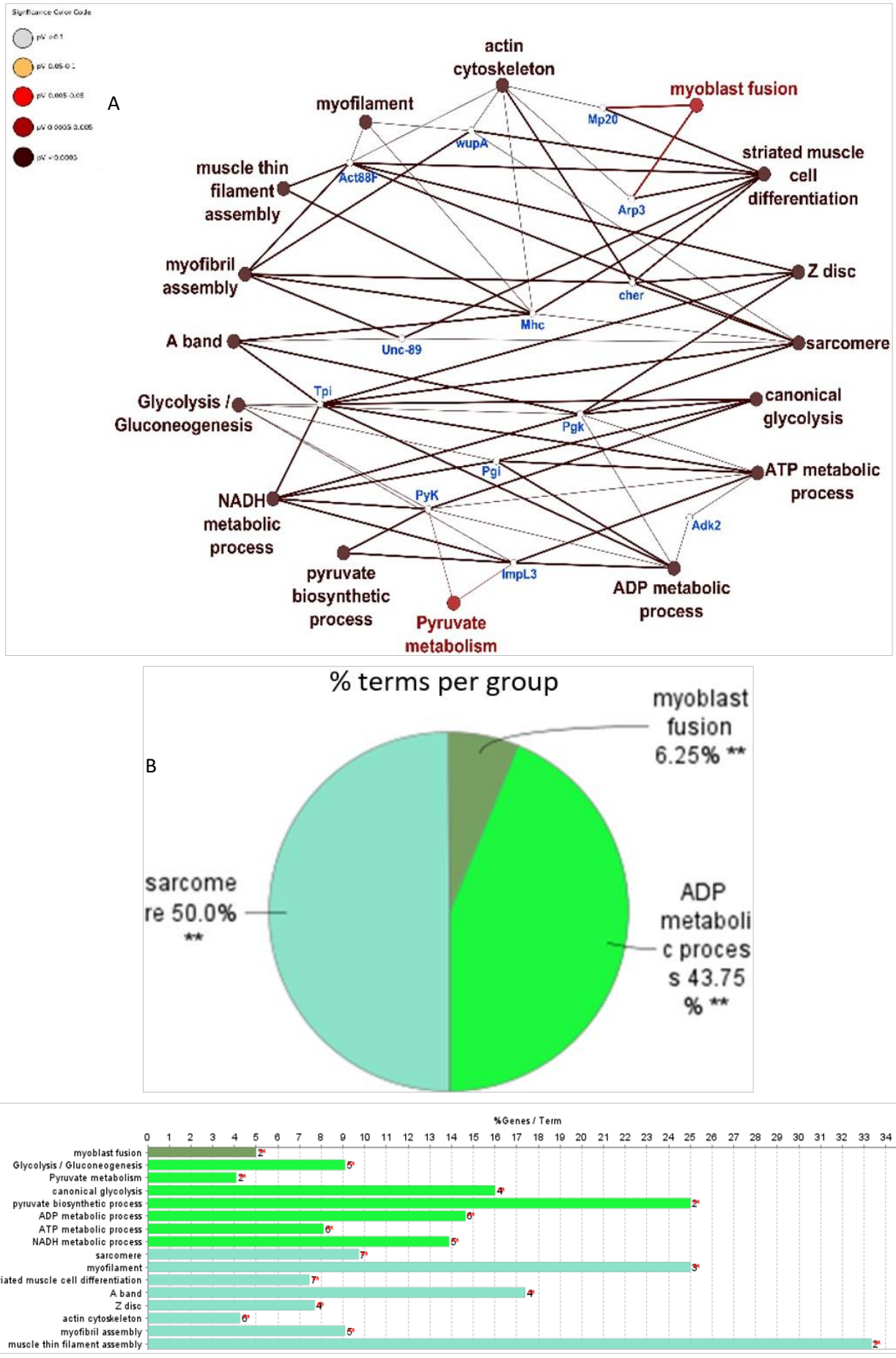

Figure 5. ClueGo network analysis of differential gene expression of P. monodon collected from two different environmental condition (A. ClueGo network showing the hub genes, B. %terms per group C. %genes/term).

Table 3. SSR identified in muscle transcriptome of *Penaeus monodon*

Total number of sequences examined	41108
Total size of examined sequences (bp)	37859227
Total number of identified SSRs	19637
Number of SSR containing sequences	12252
Number of sequences containing more than 1 SSR	4633
Number of SSRs present in compound formation	4355

Table 4. Polymorphic SSR identified after testing from differentially expressed genes in two different populations of *Penaeus monodon*

ID	No. of alleles	Range of allele size	Annotation	Gene symbol (*Penaeus* species)	SSR type	SSR
DN2495_1	2	88,91	obscurin isoform X3	*c7m84_002126*	p3	(GAA)5
DN59441_1	2	99,102	CCAAT/enhancer-binding protein-like	*c7m84_016687*	p3	(GAG)7
DN971_2	4	173-182	uncharacterized protein LOC119571649	*loc119571649*	p3	(AGG)9

Table 5. Significant SNPs found, based on frequencies in two different populations of *Penaeus monodon* muscle tissue (*DGE)

S.no	Trinity	Base Location	Annotation	Ref	Alt	p-value
1	TRINITY_DN1022_c0_g1	658	Glucose-6-phosphate isomerase-like	T	C	0.025031
2	TRINITY_DN1586_c0_g1*	1579	Tubulin beta chain-like	G	T	0.004105
3	TRINITY_DN16193_c0_g1	576	Fructose-1,6-bisphosphatase 1-like isoform X2	G	A	0.025031
4	TRINITY_DN184_c0_g1*	202	Triosephosphate isomerase	G	A	0.025031
5	TRINITY_DN1844_c0_g1	1293	Alpha-crystallin A chain-like	C	T	0.025031
6	TRINITY_DN1844_c0_g1*	1340	Alpha-crystallin A chain-like	G	T	0.025031
7	TRINITY_DN24_c0_g1	3071	Protein TAR1-like	T	C	0.025031
8	TRINITY_DN2491_c0_g1	420	Calumenin	T	A	0.025031
9	TRINITY_DN2495_c0_g3*	362	Obscurin isoform X3	T	C	0.004105
10	TRINITY_DN27593_c0_g1*	635	Protein lethal (2) essential for life-like	T	A	0.025031
11	TRINITY_DN29728_c12_g1	297	Actin-like (LOC119597217), mrna	G	C	0.004105
12	TRINITY_DN3083_c0_g1	266	UDP-N-acetylglucosamine 1-carboxyvinyltransferase-like	C	T	0.025031
13	TRINITY_DN31_c0_g1	2816	Protein TIC 214-like	G	T	0.025031
14	TRINITY_DN31_c0_g1	3255	Protein TIC 214-like	A	T	0.025031
15	TRINITY_DN31_c0_g1	3377	Protein TIC 214-like	G	A	0.025031
16	TRINITY_DN311_c0_g2	616	Penaeidin-3	A	T	0.025031
17	TRINITY_DN46056_c0_g1*	1559	Adenylyl cyclase-associated protein 1-like isoform X5	G	A	0.025031
18	TRINITY_DN4940_c0_g1	1044	UPF0389 protein CG9231-like	T	G	0.025031
19	TRINITY_DN4940_c0_g1	1555	UPF0389 protein CG9231-like	G	C	0.025031
20	TRINITY_DN571_c0_g1	759	UDP-N-acetylglucosamine 1-carboxyvinyltransferase-like	A	T	0.025031
21	TRINITY_DN58214_c0_g1	279	Adenylate kinase isoenzyme 1-like	C	T	0.025031
22	TRINITY_DN6059_c0_g1	65	Phosphoglucomutase	C	A	0.025031
23	TRINITY_DN6858_c0_g1	2515	Probable phosphorylase b kinase regulatory subunit beta isoform X5	G	A	0.004105
24	TRINITY_DN84_c0_g1	1675	Facilitated trehalose transporter Tret1-like	T	A	0.025031
25	TRINITY_DN84_c0_g1	1694	Facilitated trehalose transporter Tret1-like	A	T	0.025031
26	TRINITY_DN84_c0_g1	818	Facilitated trehalose transporter Tret1-like	T	C	0.025031
27	TRINITY_DN84_c0_g1*	520	Facilitated trehalose transporter Tret1-like	C	T	0.025031
28	TRINITY_DN85_c0_g1	2570	UDP-N-acetylglucosamine 1-carboxyvinyltransferase-like	A	T	0.025031
29	TRINITY_DN85_c0_g1	3404	UDP-N-acetylglucosamine 1-carboxyvinyltransferase-like	T	A	0.004105
30	TRINITY_DN85_c0_g1	4910	UDP-N-acetylglucosamine 1-carboxyvinyltransferase-like	C	T	0.025031
31	TRINITY_DN858_c0_g2	249	Uncharacterized protein LOC119574245	G	A	0.004105

regulation effects increase in glycolysis, through interaction with fructose-bisphosphate, facilitating increase anaerobic production of ATP and lactate, under anoxic conditions (Dawson et. al., 2013). In GO group 3 of DEGs, under GO GO:0051146: striated muscle cell differentiation, Myophilin gene, which has highest eight interactions, is a smooth muscle protein (https://www.uniprot.org/uniprotkb/A0A6I8URR3/entry) and functions in regulation and modulation of smooth muscle contraction. Another gene with significant differential expression in this group, actin 2 (*act2*), a ubiquitous protein, has been reported to confer increased tolerance to oxidative stress (Kuběnová et al., 2021) and is also involved in the formation of filaments that are major components of the cytoskeleton (https://www.ebi.ac.uk/interpro/entry/InterPro/IPR004000/).

The large-scale SSRs and SNP associated the expressed genes as well as the differentially expressed genes, generated in the present study, can be a useful reference resource for converting them into molecular markers for aquaculture research. Most of the SNPs belong to the genes in Glycolysis / Gluconeogenesis, Pentose phosphate pathway and Fructose and mannose metabolism, which can be as a marker to response to environmental disturbances (Malini et al., 2018). Notably, the identified DEGs and their associated molecular markers, including SNPs and SSRs, presents a

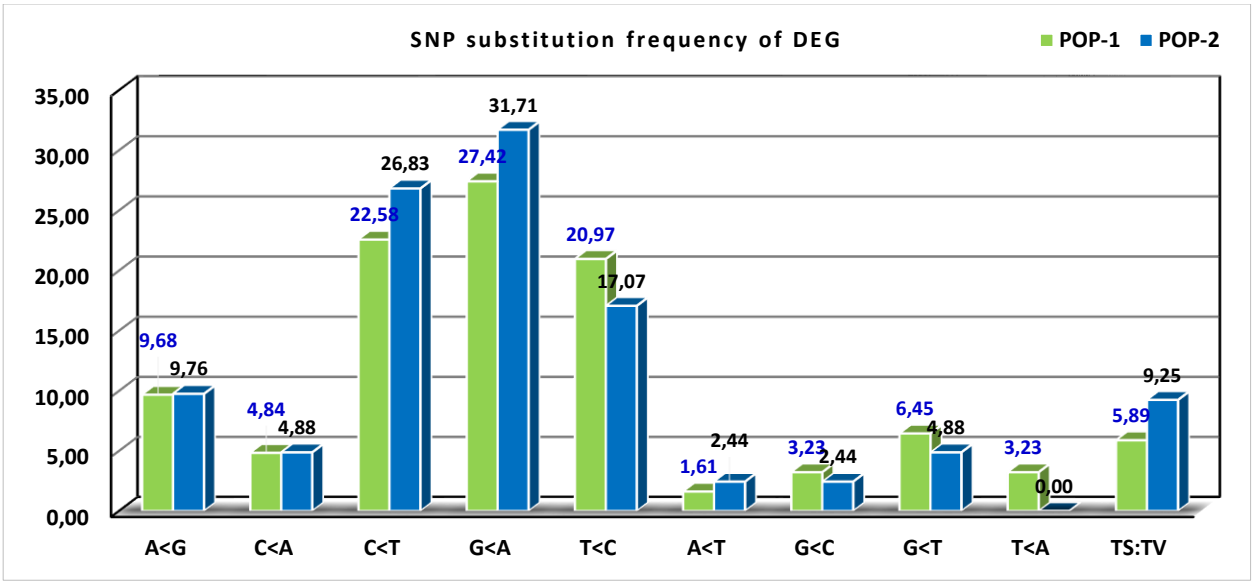

Figure 6. SNP substitution frequency of two different populations of *Penaeus monodon* differentially expressed genes in muscle transcriptome.

valuable genomic resource for future aquaculture research, through their utilization for selective breeding programmes.

Conclusion

This study provides significant insights into crucial gene expression patterns in *P. monodon*. Current study generated a robust transcriptome infotmation as well as their associated molecular markers, including SNPs and SSRs, presents a valuable genomic resource for future aquaculture research, through their utilization for selective breeding programmes. Thus, the cSSRs and cSNP associated with coding genes as well as differentially expressed genes identified in present study could be a part of the marker set, which would be useful as marker loci or candidate genes for identification of QTL effects, to be use in construction of smaller SNP arrays for pedigree analysis and/or genomic selections.

Ethical Statement

The study was reviewed and approved by the Institute Animal Ethics Committee of ICAR-National Bureau of Fish Genetic Resources, Lucknow (Registration No. 909/GO/Re/S/05/CPCSEA) vide No. G/IAEC/2022/6. All the methods were performed in accordance with the relevant guidelines and regulations.

Funding Information

This work was supported by grants from Indian Council of Agricultural Research-Consortium Research project on Genomics (ICAR-CRP Genomics), New Delhi, India, vide Sanction letter no. Fy/9/15/2017-IA.VI dated 08th August, 2017.

Author Contribution

VM and JKJ designed and supervised the study. VSB and SK collected the samples. LMC, NC, NS and VM performed transcriptome annotation and computational analysis. VM, LMC and NC wrote the manuscript. All edited the manuscript. All authors read and approved the final version of the manuscript.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Acknowledgements

The authors are grateful to Director, ICAR-NBFGR, Lucknow, India for providing facilities for this work. This work was carried out under ICAR-Consortium Research Platform on Genomics (ICAR-CRP Genomics), New Delhi, India and financial assistance by ICAR-CRP Genomics is duly acknowledged. Dr. Rajeev K. Singh and Dr. B. Kushwaha are acknowledged for their support.

References

Angthong P, Uengwetwanit T, Arayamethakorn S, Rungrassamee W. Transcriptomic analysis of the black tiger shrimp (Penaeus monodon) reveals insights into immune development in their early life stages. Sci Rep. 2021 Jul 6;11(1):13881.

Avvaru AK, Sowpati DT, Mishra RK. (2018) PERF: an exhaustive algorithm for ultra-fast and efficient identification of microsatellites from large DNA sequences. Bioinformatics. 34(6):943-8.

Beier S, Thiel T, Münch T, Scholz U, Mascher M. (2017) MISA-web: a web server for microsatellite prediction. Bioinformatics. 33(16):2583-5.

Burgos-Aceves MA, Abo-Al-Ela HG, Faggio C (2021) Physiological and metabolic approach of plastic additive effects: Immune cells responses. Journal of hazardous materials. 404:124114.

Chen S, Zhou Y, Chen Y, Gu J. (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 34(17): i884-90.

Dawson NJ, Biggar KK, Storey KB. (2013) Characterization of fructose-1, 6-bisphosphate aldolase during anoxia in the tolerant turtle, *Trachemys scripta elegans*: an assessment of enzyme activity, expression and structure. PloS one. 8(7): e68830.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 29(1):15-21.

Dong S, Nie H, Li D, Cai Z, Sun X, Huo Z, Yan X. (2020) Molecular cloning and characterization of Y-box gene (Rpybx) from Manila clam and its expression analysis in different strains under low-temperature stress. Animal Genetics. 51(3):430-8.

Doran, A.G., Creevey, C.J. (2013). Snpdat: Easy and rapid annotation of results from *de novo* snp discovery projects for model and non-model organisms. *BMC Bioinformatics* 14, 45.

Dzeja PP, Hoyer K, Tian R, Zhang S, Nemutlu E, Spindler M, Ingwall JS. (2011) Rearrangement of energetic and substrate utilization networks compensate for chronic myocardial creatine kinase deficiency. The Journal of physiology. 589(21):5193-211.

Firth JD, Ebert BL, Ratcliffe PJ. (1995) Hypoxic regulation of lactate dehydrogenase a: interaction between hypoxia-inducible factor 1 and camp response elements (∗). Journal of Biological Chemistry. 270(36):21021-7.

Flanagan SP, Jones AG. (2019) The future of parentage analysis: From microsatellites to SNPs and beyond. Molecular ecology. 28(3):544-67.

Gilbert D. (2016) Accurate and complete gene construction with EvidentialGene. InGalaxy Community Conference (Vol. 5, p. 1567).

Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic acids research. 36(10):3420-35.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology. 29(7):644-52.

Huang Y, Niu B, Gao Y, Fu L, Li W. (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics. 26(5):680-2.

Jan YH, Lai TC, Yang CJ, Lin YF, Huang MS, Hsiao M. (2019) Adenylate kinase 4 modulates oxidative stress and stabilizes HIF-1α to drive lung adenocarcinoma metastasis. Journal of Hematology & Oncology. 12:1-4. Ji J, Wang Q, Li S, Chen Y, Zhang J, Yu H, Xu J, Li M, Zheng R, Lin N and Zhang Z (2024) Transcriptomic analysis of *Penaeus monodon* in response to acute and chronic hypotonic stress. *Front. Vet. Sci*. 11:1464291.

Johnston IA, Kent MP, Boudinot P, Looseley M, Bargelloni L, Faggion S, Merino GA, Ilsley GR, Bobe J, Tsigenopoulos CS, Robertson J. (2024) Advancing fish breeding in aquaculture through genome functional annotation. Aquaculture. 740589.

Kang YJ, Yang DC, Kong L, Hou M, Meng YQ, Wei L, Gao G. (2017) CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic acids research. 45(W1): W12-6.

Katzemich A, Kreisköther N, Alexandrovich A, Elliott C, Schöck F, Leonard K, Sparrow J, Bullard B. (2012) The function of the M-line protein obscurin in controlling the symmetry of the sarcomere in the flight muscle of Drosophila. Journal of cell science. 125(14):3367-79.

Klepinin A, Zhang S, Klepinina L, Rebane-Klemm E, Terzic A, Kaambre T, Dzeja P. (2020) Adenylate kinase and metabolic signaling in cancer cells. Frontiers in oncology. 10:660.

Kuběnová L, Takáč T, Šamaj J, Ovečka M. (2021) Single amino acid exchange in ACTIN2 confers increased tolerance to oxidative stress in Arabidopsis der1-3 mutants. International journal of molecular sciences. 22(4):1879.

Lane AN, Fan TW. (2015) Regulation of mammalian nucleotide metabolism and biosynthesis. Nucleic acids research. 43(4): 2466-85.

Li D, Nie H, Jahan K, Yan X. (2020) Expression analyses of C-type lectins (CTLs) in Manila clam under cold stress provide insights for its potential function in cold resistance of *Ruditapes philippinarum*. Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology. 230:108708.

Li, Y., Chen, J., Jiang, S. et al. (2024). A comprehensive study on nutritional quality, physiological enzyme activity and genetic diversity in six populations of *Penaeus monodon*. *Aquacult Int* 32, 10141–10157

Love MI, Huber W, Anders S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology. 15:1-21.

Malini DM, Apriliandri AF, Arista S. (2018) Increased blood glucose level on pelagic fish as response to environmental disturbances at east coast, Pangandaran, West Java. InIOP Conference Series: Earth and Environmental Science (Vol. 166, No. 1, p. 012011).

Martin RM, Gasser RB, Jones MK, Lightowlers MW. (1995) Identification and characterization of myophilin, a muscle-specific antigen of *Echinococcus granulosus*. Molecular and biochemical parasitology. 70(1-2):139-48.

Mohindra V, Chowdhury LM, Chauhan N, Paul A, Singh RK, Kushwaha B, Maurya RK, Lal KK, Jena JK. (2023) Transcriptome analysis revealed osmoregulation related regulatory networks and hub genes in the gills of Hilsa shad, *Tenualosa ilisha*, during the migratory osmotic stress. Marine Biotechnology. 25(1):161-73.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic acids research. 35(suppl_2): W182-5.

Nie H, Liu L, Huo Z, Chen P, Ding J, Yang F, Yan X. (2017) The HSP70 gene expression responses to thermal and salinity stress in wild and cultivated Manila clam *Ruditapes philippinarum*. Aquaculture.470:149-56.

Nishimura O, Hara Y, Kuraku S. (2019) Evaluating genome assemblies and gene models using gVolante. Gene prediction: methods and protocols. 247-56.

Nguyen Thanh Minh, Tran N M, Tran YTHi, Le T T H, Tran L M, Nguyen P, Le T, Elizur A, Ventura T, Vo TTM, Tuan Viet Nguyen (2025). Transcriptomic comparison between

wild-caught and domesticated black tiger shrimp (*Penaeus monodon*) in early and late-vitellogenic broodstock females. Aquaculture Reports, 41: 102675.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. (2017) Salmon provides fast and bias-aware quantification of transcript expression. Nature methods. 14(4):417-9.

Reverter A, Hudson NJ, McWilliam S, Alexandre PA, Li Y, Barlow R, Welti N, Daetwyler H, Porto-Neto LR, Dominik S. (2020) A low-density SNP genotyping panel for the accurate prediction of cattle breeds. Journal of Animal Science. 98(11): 337.

Rossi N, Grosso C, Delerue-Matos C. (2024) Shrimp Waste Upcycling: Unveiling the Potential of Polysaccharides, Proteins, Carotenoids, and Fatty Acids with Emphasis on Extraction Techniques and Bioactive Properties. Marine Drugs. 22(4):153.

Soo, T.C.C., Devadas, S., Mohamed Din, M.S. et al. Differential transcriptome analysis of the disease tolerant Madagascar–Malaysia crossbred black tiger shrimp, *Penaeus monodon* hepatopancreas in response to acute hepatopancreatic necrosis disease (AHPND) infection: inference on immune gene response and interaction. *Gut Pathog* 11, 39 (2019).

Sukhavachana S, Tongyoo P, Luengnaruemitchai A, Poompuang S. (2021) Optimizing genomic prediction using low-density marker panels for streptococcosis resistance in red tilapia (Oreochromis spp.). Animal genetics. 52(5):667-74.

Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ. (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic acids research. 47(D1): D607-13.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. (2012) Primer3-new capabilities and interfaces. Nucleic acids research. 40(15): e115.

Van der Auwera GA, O'Connor BD. (2020) Genomics in the cloud: using Docker, GATK, and WDL in Terra. O'Reilly Media.

Vu NTT, Zenger KR, Guppy JL et al (2020) Fine-scale population structure and evidence for local adaptation in Australian giant black tiger shrimp (Penaeus monodon) using SNP analysis. BMC Genomics 21:669.

Wong LL, Chun LC, Deris ZM et al (2021) Genetic diversity and population structure of wild and domesticated black tiger shrimp (*Penaeus monodon*) broodstocks in the Indo-Pacific regions using consolidated mtDNA and microsatellite markers. Gene Rep 23:101047.

Xiang R, Breen EJ, Prowse-Wilkins CP, Chamberlain AJ, Goddard ME. (2021) Bayesian genome-wide analysis of cattle traits using variants with functional and evolutionary significance. Animal Production Science.

Zenger KR, Khatkar MS, Jones DB, Khalilisamani N, Jerry DR, Raadsma HW (2019) Genomic selection in aquaculture: application, limitations and opportunities with special reference to marine shrimp and pearl oysters. Frontiers in genetics. 9:693.