scientific data

Check for updates

OPEN

DATA DESCRIPTOR

De novo transcriptome assembly and annotation of *Penaeus monodon* hemocytes under WSSV infection and STAT knockdown

Thapanan Jatuyosporn[1,2], Pasunee Laohawutthichai[1,2], Fernanda Cornejo-Granados[3], Luigui Gallardo-Becerra[3], Anchalee Tassanakajon[2], Juan Manuel Hurtado-Ramírez[3], Adrian Ochoa-Leyva[3] ✉ & Kuakarun Krusong[1] ✉

Black Tiger shrimp (*Penaeus monodon*) underpins global aquaculture. However, it is periodically impacted by White Spot Syndrome Virus (WSSV). Because shrimp lack adaptive immunity, antiviral defense relies on innate pathways, including Toll, IMD, and JAK/STAT signaling. We provide a *de novo* hemocyte transcriptome assembly and RNA-seq dataset to interrogate JAK/STAT function during WSSV infection in *P. monodon*. Hemocytes were collected at 12 h post-infection under three conditions: (i) WSSV-infected (WSSV), (ii) *Pm*STAT knockdown (*Pm*STAT dsRNA), and (iii) *Pm*STAT-knockdown + WSSV-infected (*Pm*STAT dsRNA + WSSV). Twelve libraries produced 179.4 million paired-end sequencing reads, with ~99% passed quality filtering. *De novo* assembly produced 778,680 transcripts; selecting the longest isoform per gene yielded a 586,928-isoform reference set and 28,423 predicted proteins. BUSCO assesment indicated a transcriptome completeness around 90% (Arthropoda). We functionally annotatted 15,966 proteins using BLAST, eggNOG, InterPro, and UniProt. qRT–PCR confirmed *Pm*STAT knockdown, WSSV infection, and RNA-seq trends for nine differentially expresed genes (DEGs). This validated resource supports future studies of JAK/STAT-dependent antiviral programs and host–WSSV interactions in *P. monodon*.

Background & Summary

Black Tiger shrimp (*Penaeus monodon*) underpins global aquaculture, with production rising from ~5.6 to ~5.9 million metric tons between 2023 and 2024 (~4.3% growth)[1]. However, its production is periodically disrupted by pathogen outbreaks, notably White Spot Syndrome Virus (WSSV)[2–4]. First identified in Taiwan in 1992, WSSV rapidly spread across Asia and beyond[5,6]. It can cause mortality within 2–7 days[7] and remains one of the most damaging pathogens in shrimp aquaculture. Because shrimp lack adaptive immunity, antiviral defense relies on innate immune pathways, with prominent activation of the Toll, IMD, and JAK/STAT signaling cascades[8].

The JAK/STAT pathway, centered on the Domeless (DOME) receptor, Janus kinase (JAK), and STAT transcription factors, transduces extracellular cytokine-like signals into transcriptional programs that regulate hematopoiesis, tissue repair, proliferation, apoptosis, and immune responses[9–14]; Upon ligand binding, DOME receptor dimerizes and activates the associated JAK, which undergoes autophosphorylation and phosphorylates tyrosine residues on the receptor's cytoplasmic tail. These phosphotyrosines then recruit STAT proteins[15] through their Src homology 2 (SH2) domains, then STAT is phosphorylated by JAK, dimerizes, and translocate into the nucleus where it binds promoter elements to regulate target gene expression[2]. In penaeid shrimp, STAT

[1]Center of Excellence in Structural and Computational Biology, Department of Biochemistry, Faculty of Science, Chulalongkorn University, Bangkok, 10330, Thailand. [2]Center of Excellence for Molecular Biology and Genomics of Shrimp, Department of Biochemistry, Faculty of Science, Chulalongkorn University, Bangkok, 10330, Thailand. [3]Departamento de Microbiología Molecular, Instituto de Biotecnología (IBT), Universidad Nacional Autónoma de México (UNAM), Av. Universidad 2001, Col. Chamilpa, Cuernavaca, Morelos, 62210, Mexico. ✉e-mail: adrian.ochoa@ibt.unam.mx; kuakarun.k@chula.ac.th

has been implicated in antiviral responses but may also be targeted by WSSV; for example, shrimp STAT has been reported to activate the WSSV immediate-early gene *ie1*[3]. Consistently, STAT silencing in *P. monodon* hemocytes has shown to reduce WSSV copy number[4,16,17]. Transcriptomic studies across hemocytes, hepatopancreas, and gills have begun to map these responses[18,19]. Despite this accumulating evidence, publicly available datasets that combine WSSV infection with controlled STAT perturbation—allowing separation of infection-driven and STAT-dependent transcriptional programs—remain limited.

We provide a *de novo* hemocyte transcriptome assembly and associated RNA-seq dataset for *P. monodon* under WSSV infection, providing a resource for reproducible analyses of crustacean host–pathogen interactions. The dataset was generated using a targeted experimental design to examine transcriptional responses related to JAK/STAT signaling during WSSV challenge. Hemocyte samples were collected from three experimental conditions: (i) WSSV-infected shrimp (WSSV), (ii) knockout *Pm*STAT dsRNA-injected shrimp (*Pm*STAT dsRNA), and (iii) *Pm*STAT dsRNA-injected shrimp subsequently challenged with WSSV (*Pm*STAT dsRNA + WSSV).

Across 12 libraries, we generated 179.4 million paired-end reads (2 × 75 bp), of which ~99% were retained after quality filtering. *De novo* assembly with Trinity produced a transcriptome comprising 778,680 transcripts and 586,928 unigenes. Assembly statistics calculated based on the longest isoform of each unigene (586,928) showed a contig N50 of 417 bp, and a median contig length of 325 bp. Benchmarking Universal Single-Copy Orthologs (BUSCO) analysis indicated an overall completeness of approximately 90%. From the 586,928 transcript isoforms, protein prediction identified 28,423 proteins considering the longest isoforms. Functional annotation by BLAST, Gene Ontology (GO), eggNOG, Kyoto Encyclopedia of Genes and Genomes (KEGG), and InterProScan was obtained for 15,966 proteins (56.2%). Quantification (Bowtie2/RSEM) and differential expression (DESeq. 2) analyses revealed 675 upregulated genes in WSSV versus *Pm*STAT dsRNA + WSSV and 138 upregulated in *Pm*STAT dsRNA + WSSV versus WSSV.

We released full raw sequencing data, the assembled transcriptome, functional annotations, and reproducible analysis code. This dataset's three-condition design (WSSV, PmSTAT dsRNA, and PmSTAT dsRNA + WSSV) enables separation of transcriptional responses to viral infection, *PmSTAT* knockdown, and their combined effects in shrimp hemocytes. In addition to host immune responses, it may facilitate identification of candidate *PmSTAT*-dependent WSSV transcripts, supporting future mechanistic studies of antiviral immunity.

This resource is intended to support comparative immune-transcriptomics across crustaceans, benchmarking of assembly/annotation pipelines and downstream discovery of candidate genes and transcriptional modules associated with WSSV infection and STAT-dependent regulation. By combining viral challenge with host-factor knockdown in a standardized design, this dataset enables direct inference of STAT-dependent host responses during WSSV infection. Overall, this transcriptomic dataset provides valuable resources for comparative analyses of antiviral pathway activity across crustaceans; inference of transcriptional modules associated with WSSV infected and host-factor perturbation; selection of candidate genes for qRT-PCR validation or functional assays; training and evaluation of machine-learning models for infection-state classification; and meta-analyses of hemocyte transcriptional programs.

Methods

Animals and Ethics statement. Healthy *Penaeus monodon* (≈3 g) organisms were obtained from a shrimp farm in Chachoengsao Province, Thailand. Animals were acclimated for at least one week in a recirculating seawater system (20 ppt, aerated) and fed a commercial diet twice daily. All procedures followed ethical guidelines under CU-ACUC protocol 1923021 (Chulalongkorn University Animal Care and Use Committee). Biosafety practices were approved by the Institutional Biosafety Committee of Chulalongkorn University (SC-CU-IBC-004-2018).

Hemolymph collection and RNA/cDNA preparation. Hemolymph was collected from Black Tiger shrimps using a 26 G needle and a 1 mL syringe preloaded with 500 µL anticoagulant (pH 5.6: 0.82% w/v NaCl, 0.55% w/v citric acid, 1.98% w/v glucose, 0.88% w/v sodium citrate). Hemocytes were pelleted by centrifugation (800 × g, 10 min, 4 °C); the cell-free hemolymph was discarded. Total RNA was extracted from the hemocyte pellet using the FavorPrep Tissue Total RNA Mini Kit (Favorgen, Taiwan) following the manufacturer's instructions, quantified with a NanoDrop 2000c (Thermo Fisher), and stored at −80 °C. cDNA was synthesized from 500 ng total RNA with the RevertAid First Strand cDNA Synthesis Kit (Thermo Fisher) per the manufacturer's protocol and stored at −20 °C.

***Pm*STAT dsRNA synthesis.** Two *Pm*STAT primer sets (Table 1), as described previously[17], were used. Hemocyte cDNA was diluted 1:10 in ultrapure water, and 1 µL of the dilution served as template for *Pm*STAT dsRNA amplification. PCR conditions were: 94 °C for 2 min; 35 cycles of 98 °C for 10 s, 58 °C for 30 s, and 68 °C for 30 s; final extension at 68 °C for 7 min. *Pm*STAT dsRNA was synthesized *in vitro* using the T7 RiboMAX™ Express Large Scale RNA Production System (Promega) following the manufacturer's instructions. Product integrity was assessed by 2% agarose gel electrophoresis, and concentration was determined by absorbance at 260 nm dsRNA aliquots were stored at −80 °C until use.

Experimental design, sample collection and total RNA isolation. *P. monodon* organisms were allocated to three experimental groups: (i) WSSV-infected (WSSV), (ii) *Pm*STAT-silenced (*Pm*STAT dsRNA), and (iii) *Pm*STAT-silenced and challenged with WSSV (*Pm*STAT dsRNA + WSSV). Sample size (n = 10 shrimp per group) was determined with G*Power 3.1[20]. In the *Pm*STAT-silenced groups (*Pm*STAT-dsRNA and *Pm*STAT-dsRNA + WSSV), each shrimp received 10 µg *Pm*STAT dsRNA per g of body weight, followed by a second identical dose 24 h later. The WSSV-only group received phosphate-buffered saline (PBS) instead of dsRNA.

Primer name	Orientation	Sequence (5′ – 3′)	Experiment
dsSTAT	Forward	GCC AGT TGT AGT CAT TGT CC	double-stranded RNA synthesis
	Reverse	CAAAGCTGCCACTGGAAGGG	
dsSTAT-T7	Forward	GGA TCC TAA TAC GAC TCA CTA TAG G GCC AGT TGT AGT CAT TGT CC	double-stranded RNA synthesis
	Reverse	GGA TCC TAA TAC GAC TCA CTA TAG G CAA AGC TGC CAC TGG AAG GG	
EF1-α	Forward	GGT GCT GGA CAA GCT GAA GGC	qRT-PCR
	Reverse	CGT TCC GGT GAT CAT GTT CTT GAT G	
C-reactive protein 1.1-like	Forward	GCG ACC ACG TTT AAC TAT GC	qRT-PCR
	Reverse	CCA GAA TGC CGT TAA TGT GG	
legumain-like	Forward	CCG AGC TCA GCG AAA CCA TC	qRT-PCR
	Reverse	CAG GAG GTT GGA GAA CAT TG	
leukocyte elastase inhibitor A-like	Forward	AGC GTG GAT AAC AGG TAC AG	qRT-PCR
	Reverse	AAC CG TCT GTA GAC ACT CTC	
ankyrin repeat domain-containing protein 12-like	Forward	AGG GCG CAA GAA TCT AGC AC	qRT-PCR
	Reverse	AGG GCC TCG ATC CTC TTG TC	
lactoylglutathione lyase-like	Forward	GTG GAC AAG GCA TGT GAA AG	qRT-PCR
	Reverse	ATG TCG CAA TGG TTC CAG AG	
peptidyl-prolyl cis-trans isomerase H	Forward	ATG GCC CAG GCT TAT TAT CG	qRT-PCR
	Reverse	CGA TCT TTC GCA TCA CAA GG	
ProPO1	Forward	GGT CTT CCC CTC CCG CTT CG	qRT-PCR
	Reverse	GCC GCA GGT CCT TTG GCA GC	
ProPO2	Forward	GCC AAG GGG AAC GGG TGA TG	qRT-PCR
	Reverse	TCC CTC ATG GCG GTC GAG GT	
PPAE2	Forward	ATG CAC TAC CGG GTT CCC ACG ATC	qRT-PCR
	Reverse	CTA AGG TTT GAG ATT CTG CAC G	
*Pm*STAT	Forward	TAT ATC CGA ATG TGC CTA AG	qRT-PCR
	Reverse	ATA GTT TGT GGT GTG TTG GG	
WSSV-IE1	Forward	GCT AGG GAT GTG ACT TTC	qRT-PCR
	Reverse	TGC ACC TAC ACG CAT TAC	
VP28	Forward	TGA GGT TGG ATC AGG CTA CTTC	qRT-PCR
	Reverse	CCG CAT CTT CTT CCT TCA TCT G	

Table 1. Primer sequences used for double-stranded RNA (dsRNA) synthesis and qRT-PCR experiments.

For viral infection, the *Pm*STAT dsRNA + WSSV and WSSV groups were injected with ~6×10^6 copies of WSSV, while the *Pm*STAT dsRNA group received PBS. At 12 h post-infection, hemolymph was collected from each shrimp using a 26 G needle and 1 mL sterile syringe preloaded with 500 µL anticoagulant. Shrimp were then anesthetized by hypothermic shock (~5 min). Collected hemocytes were pelleted by centrifugation (800 × g, 10 min, 4 °C). Total RNA was isolated from hemocytes using the Tissue Total RNA Mini Kit (Favorgen, Taiwan) following the manufacturer's protocol and stored at −80 °C until use.

Library preparation and RNA sequencing. As mentioned above, all shrimps were allocated to groups of equal size (n = 10 per group), and hemocytes were collected using identical procedures across treatments. For RNA-seq library construction, however, we applied a stringent RNA quality criterion to ensure robust transcriptome profiling: only samples with RNA integrity number (RIN) > 7 were processed for library preparation. RNA concentration was quantified using Qubit™ (Thermo Fisher Scientific) and RNA integrity was assessed with a Bioanalyzer (Agilent Technologies). Consequently, only a subset of individuals met the quality threshold, resulting in unbalanced numbers of RNA-seq biological replicates across groups (WSSV n = 5; *Pm*STAT dsRNA n = 3; *Pm*STAT dsRNA + WSSV n = 4). Libraries were prepared with the TruSeq RNA Library Prep Kit (Illumina) following the manufacturer's protocol. Library size distributions were verified by Bioanalyzer, and sequencing was performed on an Illumina platform (2 × 75 bp paired-end) at the Instituto de Biotecnología, UNAM (Cuernavaca, México). Sequencing yielded 179,352,237 total paired-end reads (mean 14,946,020 reads per sample). Group means were 19,682,899 (WSSV), 9,977,391 (*Pm*STAT dsRNA), and 12,751,391 (*Pm*STAT dsRNA + WSSV). Raw reads were processed with Trimmomatic to remove adapters, ambiguous bases (N), and low-quality bases (Q > 20). Clean-read retention was 98.90–99.51% per sample (Table 2), providing high-quality input for downstream assembly and differential-expression analyses.

Transcriptome assembly and functional annotation. Quality-filtered reads from all experimental samples were used for the *de novo* assembly with Trinity[21] producing 778,680 transcripts (total assembled length: 364,068,637 bp) grouped into 586,928 genes (Table 3). Transcript redundancy was conducted by selecting

Library ID	Experimental condition	Raw reads	Clean reads (%)
1_S1	*Pm*STAT dsRNA	16,202,030	16,097,983 (99.35)
2_S2	*Pm*STAT dsRNA	12,890,555	12,788,438 (99.20)
3_S3	*Pm*STAT dsRNA	839,588	830,354 (98.90
4_S4	WSSV	14,467,820	14,371,003 (99.33)
5_S5	WSSV	13,316,992	13,245,638 (99.46)
6_S6	WSSV	43,615,302	43,404,612 (99.51)
7_S7	WSSV	13,142,443	13,070,733 (99.45)
8_S8	WSSV	13,871,940	13,791,480 (99.41)
9_S9	*Pm*STAT dsRNA + WSSV	10,403,660	10,335,646 (99.34)
10_S10	*Pm*STAT dsRNA + WSSV	14,615,414	14,500,097 (99.21)
11_S11	*Pm*STAT dsRNA + WSSV	14,428,705	14,355,271 (99.49)
12_S12	*Pm*STAT dsRNA + WSSV	11,557,788	11,479,499 (99.32)

Table 2. Sequencing yield and quality-filtering efficiency for each hemocyte RNA-Seq library.

	Count
Total trinity 'genes'	586,928
Total trinity transcripts	778,680
Percent GC	37.7
Total proteins predicted with Transdecoder	58,678
Total assembled length (bp)	364,068,637

Table 3. Summary statistics of the *de novo* Trinity assembly of the *Penaeus monodon* hemocyte transcriptome.

	Length (bp)
Contig N50	417
Median contig length	325
Average contig	409.1
Total assembled bases	240,110,795
Total number of proteins considering the longest isoforms	28,423

Table 4. *De-novo* assembly statistics calculated using the longest isoform per Trinity "gene".

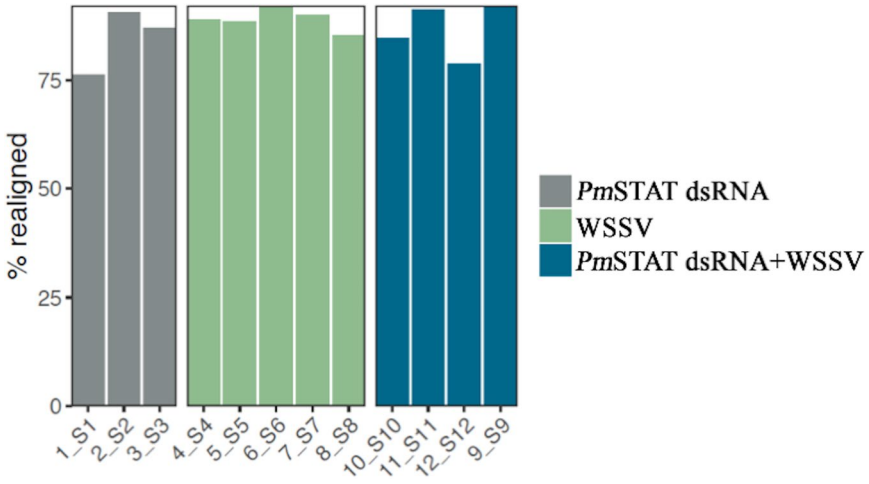

Fig. 1 Percentage of reads realigned to the transcriptome assembly. Bars represent the percentage of paired-end reads from each sample that successfully realigned to the *de novo* hemocyte transcriptome assembly. High mapping rates (>75% in all libraries) confirm the overall consistency of sequencing quality across treatments, providing a robust foundation for downstream differential expression analyses.

the longest and most supported isoform per gene, yielding a final reference set of 586,928 isoforms used for all downstream analyses. The mean isoform length was 409.1 bp (total assembled length: 240,110,795 bp). From these

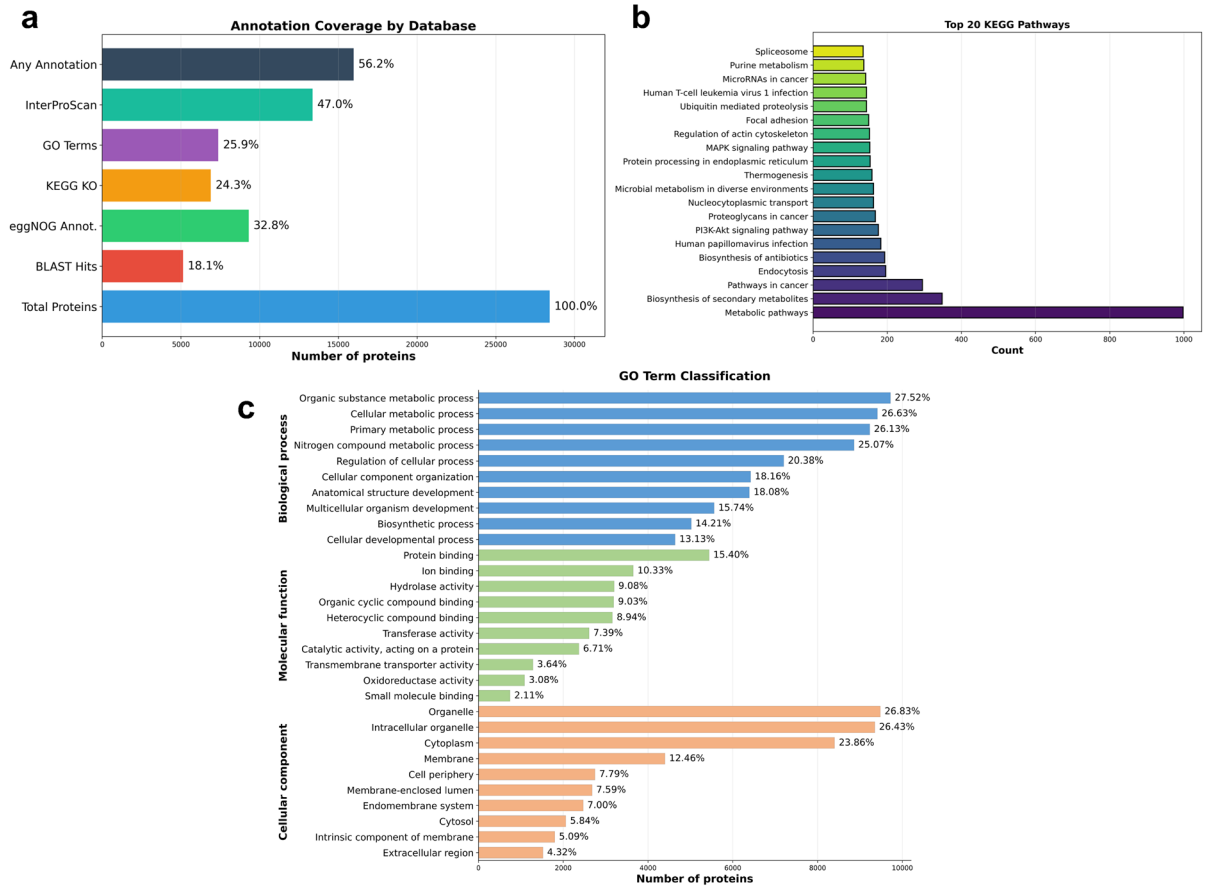

Fig. 2 Functional annotation summary of all predicted proteins generated with EnTAP. (**a**) Number of proteins annotated by each reference database used. (**b**) Top 20 KEGG pathways. (**c**) Top 10 Gene Ontology terms in each category: Molecular Function, Cellular Component, and Biological Process. The x-axis shows the number of proteins annotated to each database, term, or pathway.

Experimental condition	Number of DEGs genes	Number of genes supported by the reproducibility score
WSSV	675	349
*Pm*STAT dsRNA + WSSV	138	138

Table 5. Differentially expressed genes (DEGs) between WSSV and *Pm*STAT dsRNA + WSSV experimental conditions.

isoforms TransDecoder identified 28,423 putative proteins (Table 4). Bowtie2 realignment of reads to the transcriptome assembly indicated robust representation across libraries, with per-sample mapping rates of 76.34–92.13% (Fig. 1).

Functional annotation of the 28,423 predicted proteins (obtained from the longest isoforms) was performed using EnTAP using the default setting. In total, 5,138 proteins (18.1%) had non-redundant (nr) hits against the UniProt/Swiss-Prot reference database using an E-value threshold of 1×10^{-5} and a minimum query coverage of 50%; 7,372 (25.9%) received Gene Ontology (GO) terms; 9,313 (32.8%) received Clusters of Orthologous Groups (COG) using eggNOG annotations; 13,361 (47.0%) obtained InterProScan (domain-level functional information) matches; and 6,898 (24.3%) received KEGG orthology (KO) terms (Fig. 2). The comprehensive protein annotations for each protein can be found in https://github.com/LuiguiGallardo/p_monodon_transcriptome/tree/master/Additional_Tables.

Differential expression analysis. Transcript abundance was quantified with Bowtie2 and RSEM, and differential expression was assessed with DESeq2 using thresholds of log2 fold change (LFC) > 2 and the FDR-adjusted p-values (Benjamini–Hochberg FDR) < 0.05. No significant Differential Expressed Genes (DEGs) were detected between the *Pm*STAT dsRNA Vs. *Pm*STAT dsRNA + WSSV groups. In contrast, for the *Pm*STAT dsRNA + WSSV Vs. WSSV comparison, 138 genes were upregulated for *Pm*STAT dsRNA + WSSV, and 675 genes were upregulated for WSSV (Table 5 and Table S1). A Pearson-correlation heatmap of variance-stabilized counts shows high concordance among biological replicates and clear clustering by treatment (Fig. 3).

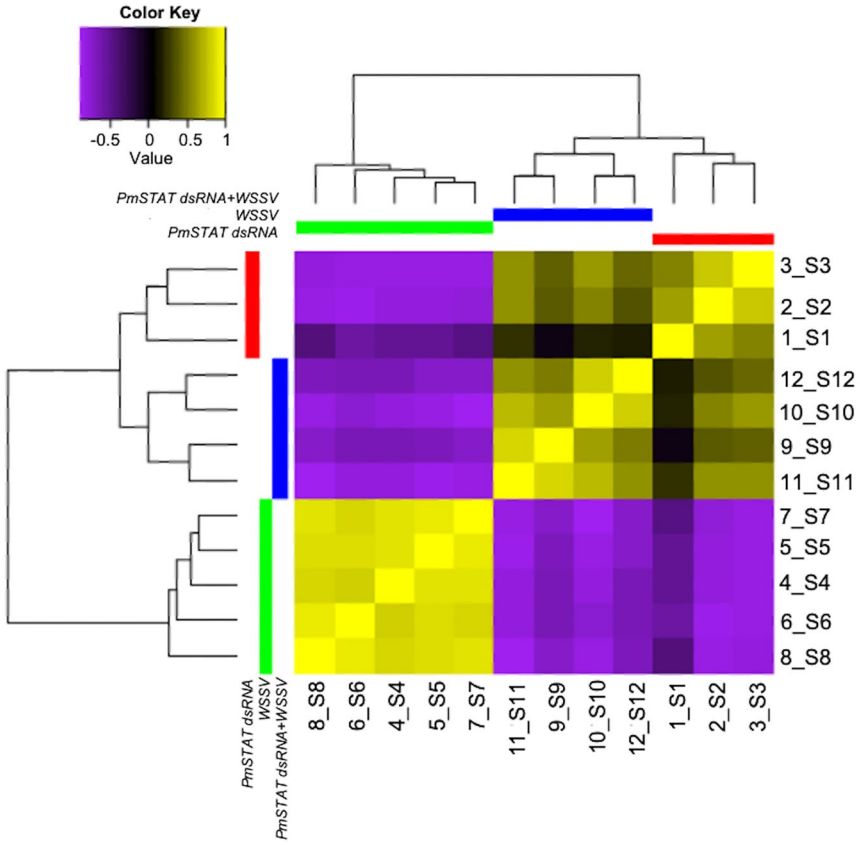

Fig. 3 Global similarity of hemocyte transcriptomes across treatments. A heat-map matrix displays pair-wise Pearson correlation coefficients (r) for all samples. Hierarchical clustering of both rows and columns groups biological replicates into three groups of samples.

Transcriptome completeness analysis. The completeness of the transcriptome assembly was evaluated using Benchmarking Universal Single-Copy Orthologs v5.7.1 (BUSCO)[22] in transcriptome mode (-m transcriptome). Two complementary lineage datasets from OrthoDB v10 were employed: 1) arthropoda_odb10 (n = 1,013 BUSCOs), the most phylogenetically specific lineage available for *P. monodon*, and 2) metazoa_odb10 (n = 954 BUSCOs), used for broader comparison. All other parameters were kept at default values. BUSCO categories reported are: Complete and single-copy (S), Complete and duplicated (D), Fragmented (F), and Missing (M) (Fig. 4 and Table 6).

Data Records

The raw reads of the 12 samples used for the transcriptome assembly have been deposited in the NCBI Sequence Read Archive (SRA) database under the accession number SRP499992[23]. The analysis pipelines and custom bioinformatics scripts are available at: https://github.com/LuiguiGallardo/p_monodon_transcriptome. Additional tables containing the comprehensive protein annotations for all the transcriptome and overexpressed genes are available at: https://github.com/LuiguiGallardo/p_monodon_transcriptome/tree/master/Additional_Tables.

Technical Validation

Experimental design and rationale for the sampling time point. Hemocytes were collected at 12 hours post-infection (hpi) to capture an infection window that is informative for both host and viral transcription while minimizing confounding effects from severe disease progression. This time point was selected for two main reasons. First, the WSSV dose used for the RNA-seq experiment matched the dose used in our mortality assays; under these conditions, sampling at later time points increases mortality risk and secondary stress effects that can bias transcriptomic profiles. Thus, 12 hpi represents a practical compromise that preserves sample integrity while maintaining a clear infection signal, consistent with our previous experimental framework[17]. Second, published evidence supports that this period corresponds to an active WSSV replication state, when viral transcription is readily detectable and informative for evaluating virus–host interactions[24]. Consistently, the WSSV group over-expressed genes of seven viral proteins supporting the time of infection (https://github.com/LuiguiGallardo/p_monodon_transcriptome/tree/master/Additional_Tables). Because we already had targeted expression profiles at 6 hpi and 24 hpi, RNA-seq at 12 hpi was chosen to provide an intermediate window to better resolve *Pm*STAT-dependent effects on host differentially expressed genes and to nominate candidate viral transcripts potentially influenced by reduced *Pm*STAT activity.

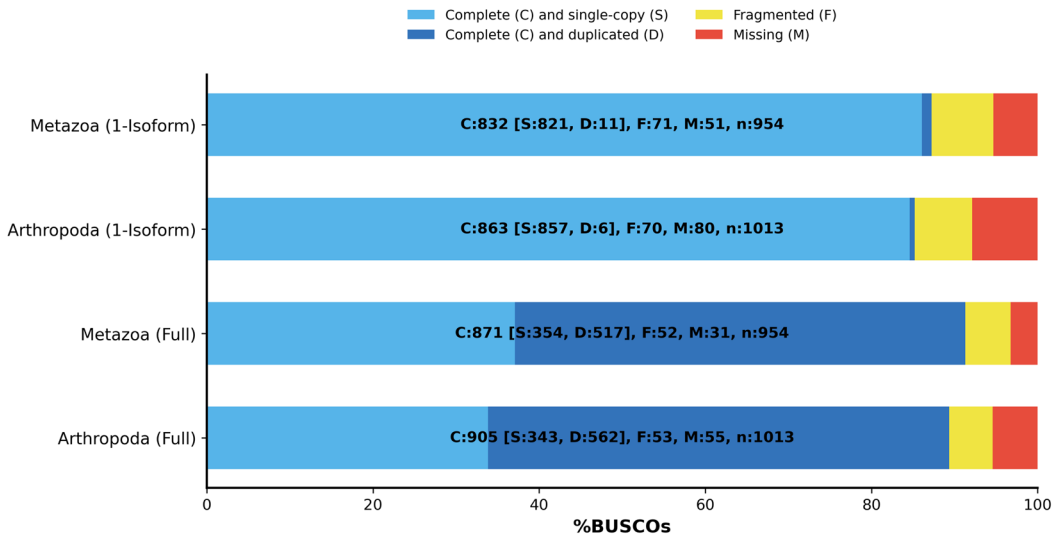

Fig. 4 BUSCO assessment of transcriptome completeness. Completeness of the *de novo P. monodon* hemocyte transcriptome evaluated with BUSCO using the Arthropoda and Metazoa lineage datasets. Analyses were performed on the full assembly including all isoforms (Full) and on a filtered set retaining the longest isoform per gene (1-isoform). For each dataset and lineage, we report the percentages of complete (C) BUSCOs, partitioned into single-copy (S) and duplicated (D), as well as fragmented (F) and missing (M) BUSCOs.

Sample	Lineage	Total BUSCOs	Complete BUSCOs (%)	Single (%)	Duplicated (%)	Fragmented (%)	Missing (%)
Metazoa (1-Isoform)	Metazoa odb10	954	832 (87.2)	821 (86.1)	11 (1.2)	71 (7.4)	51 (5.3)
Arthropoda (1-Isoform)	Arthropoda odb10	1013	863 (85.2)	857 (84.6)	6 (0.6)	70 (6.9)	80 (7.9)
Metazoa (Full)	Metazoa odb10	954	871 (91.3)	354 (37.1)	517 (54.2)	52 (5.5)	31 (3.2)
Arthropoda (Full)	Arthropoda odb10	1013	905 (89.4)	343 (33.9)	562 (55.5)	53 (5.2)	55 (5.4)

Table 6. Summary metrics from the BUSCO completeness assessment of the *de novo* assembled transcriptome.

_Pm_STAT knockdown validation by qRT–PCR. The efficiency of *Pm*STAT silencing was validated by quantifying the *Pm*STAT transcript levels in hemocytes by qRT–PCR and calculated relative expression with the $2^{-\Delta\Delta Cq}$ method, using the EF1α as the internal reference gene. Knockdown was defined as the reduction in normalized *Pm*STAT expression in *Pm*STAT dsRNA–injected shrimp relative to the PBS-injected control. In addition, we confirmed *Pm*STAT knockdown in dsRNA-treated shrimp by observing consistently reduced *Pm*STAT transcript levels relative to the WSSV-only group (Fig. 5a). Because the magnitude of RNAi-mediated suppression varies across targets and experimental settings, we did not apply a universal fixed percentage threshold (e.g., 50–70%) to define success; instead, we aimed to reduce *Pm*STAT expression to the lowest reproducible level achievable under our optimized conditions, following the validation framework established in our previous shrimp STAT silencing study[17] and consistent with prior *Pm*STAT dsRNA work in crustaceans[25]. Assay quality control included verification of a single specific amplification product by melt-curve analysis (to exclude primer–dimer artifacts) and acceptance of reactions with Cq values within the reliable detection range (typically Cq 15–35). Collectively, these controls support that *Pm*STAT expression was effectively reduced by dsRNA treatment in this study.

WSSV infection validation by qRT-PCR. Infection was confirmed in individual shrimps used for transcriptome profiling by diagnostic qRT-PCR targeting VP28, a conserved structural gene widely used as a WSSV marker which plays a key role in the initial steps of the systemic WSSV infection[26]. Amplification was performed under standard conditions. Only challenged groups yielded quantifiable amplification (measurable Cq values), while non-infected controls showed no amplification (undetermined Cq), confirming that WSSV was present exclusively in infected treatments (Table 7). This pattern confirms that WSSV was present exclusively in the challenged groups and absent (below the assay detection limit) in the non-infected controls. Furthermore, the WSSV group over-expressed genes of seven viral proteins supporting the viral infection in this group (https://github.com/LuiguiGallardo/p_monodon_transcriptome/tree/master/Additional_Tables).

RNA quality for library construction. Total RNA integrity was verified by agarose gel electrophoresis and Agilent Bioanalyzer. Only samples with RIN > 7 were used for library construction.

Read filtering and assembly. Read quality was assessed with FastQC; low-quality bases ($Q \leq 20$), adapters, and ambiguous bases (N) were trimmed with Trimmomatic (parameters: SLIDINGWINDOW:6:20) (Table 2).

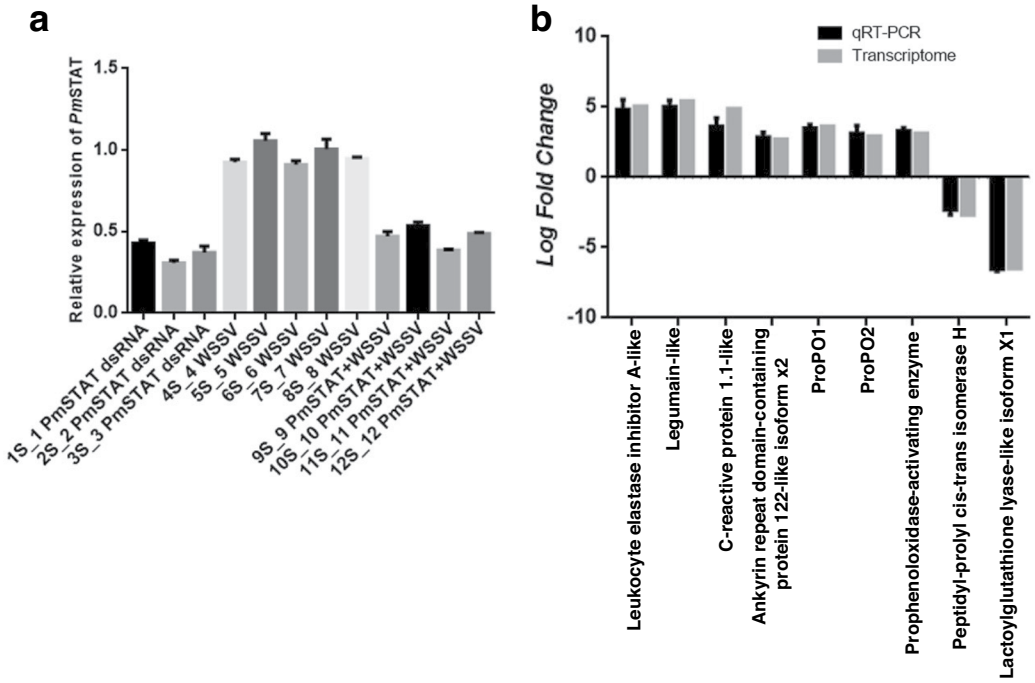

Fig. 5 *Pm*STAT knockdown and Differential Expression Genes validation experiments. (**a**) Relative *PmSTAT* expression in all RNA-seq samples, normalized to the PBS-injected control. Each sample was amplified in triplicate. (**b**) qRT–PCR validation of selected upregulated and downregulated genes differentially expressed between the WSSV and *Pm*STAT dsRNA + WSSV groups.

Library ID	Experimental condition	Cq value
1_S1	*Pm*STAT dsRNA	ND
2_S2	*Pm*STAT dsRNA	ND
3_S3	*Pm*STAT dsRNA	ND
4_S4	WSSV	24.58
5_S5	WSSV	27.26
6_S6	WSSV	23.08
7_S7	WSSV	25.42
8_S8	WSSV	24.59
9_S9	*Pm*STAT dsRNA + WSSV	36.38
10_S10	*Pm*STAT dsRNA + WSSV	31.55
11_S11	*Pm*STAT dsRNA + WSSV	30.18
12_S12	*Pm*STAT dsRNA + WSSV	35.73

Table 7. Cq values for VP28 qRT–PCR in all samples as validation of *Pm*STAT knockdown.

To evaluate sample representativeness in the assembly, reads were realigned to the transcriptome with Bowtie2 (v2.5.3) via Trinity's align_and_estimate_abundance.pl[27]. Mapping rates per sample are reported in Fig. 1. Expression estimates were generated with RSEM (v1.3.3), and count matrices were consolidated using Trinity's abundance_estimates_to_matrix.pl.

Completeness of the transcriptome assembly. The completeness of the *de novo* assembled transcriptome was evaluated using BUSCO with the Arthropoda lineage dataset and, for comparison, the Metazoa dataset (Fig. 4 and Table 6). Because BUSCO duplication metrics can be inflated in transcriptomes that retain multiple isoforms per gene, we applied BUSCO on two transcript sets: the unfiltered transcriptome including isoforms (Full = 778,680 transcripts) and the 1-isoform-per-gene filtered set (1-isoform = 586,928 transcripts). This follows BUSCO guidance that transcriptomes and protein sets that are not filtered for isoforms lead to a high proportion of duplicates and therefore should be filtered before analysis to enable proper interpretation of the "duplicated" category. In the full transcriptome set, BUSCO reported high completeness but a markedly larger duplicated fraction for arthropoda: complete (C) = 89.3% [single copy (S) = 33.9%, duplicated (D) = 55.5%], fragemented (F) = 5.2%, and missing (M) = 5.4% (Fig. 4 and Table 6). Contrary, in the 1-isoform set, BUSCO reported high completeness with minimal duplication for arthropoda: C = 85.2% [S = 84.6%, D = 0.6%], F = 6.9%, and M = 7.9%. Similar trends were observed for Metazoa set (Fig. 4 and Table 6). Across both datasets,

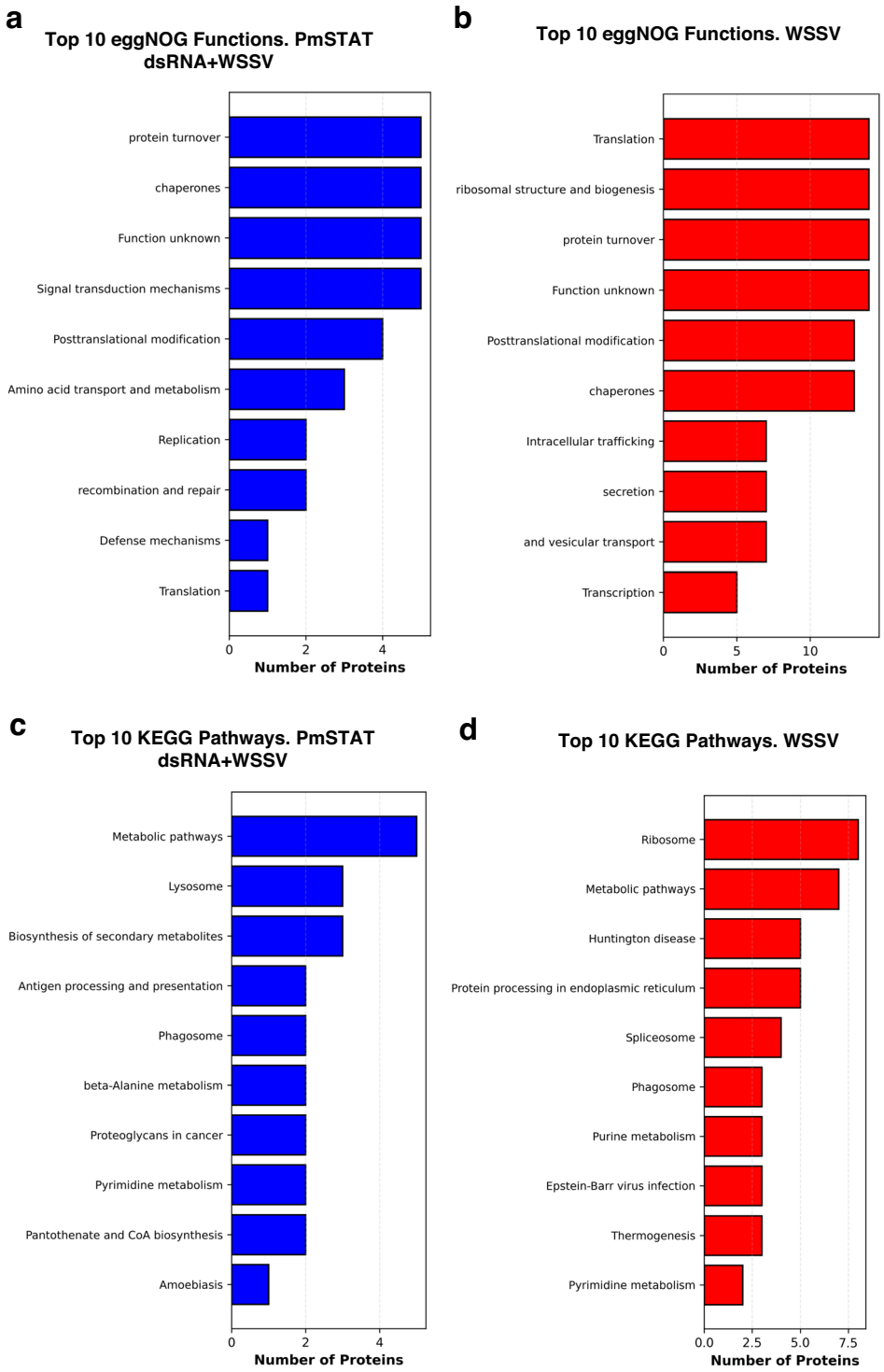

Fig. 6 EggNOG and KEGG annotations of DEGs. The top 10 eggNOG functional terms are shown for (**a**) *Pm*STAT dsRNA + WSSV and (**b**) WSSV conditions. The top 10 KEGG pathways are shown for (**c**) *Pm*STAT dsRNA + WSSV and (**d**) WSSV conditions. The x-axis indicates the number of proteins annotated to each term/pathway.

the proportions of fragmented and missing BUSCOs were comparatively low, supporting that the assembly captures most conserved arthropod/metazoan gene content. Notably, elevated duplicated BUSCO fractions comparable to those observed in our full set have been reported for other *P. monodon* transcriptomes[28], although these studies do not specify whether BUSCO was run on isoform-filtered or unfiltered transcript sets, underscoring the importance of reporting both for transparency. The low duplication observed after isoform filtering is consistent with the *P. monodon* genome context, which has been reported to contain relatively low levels of gene

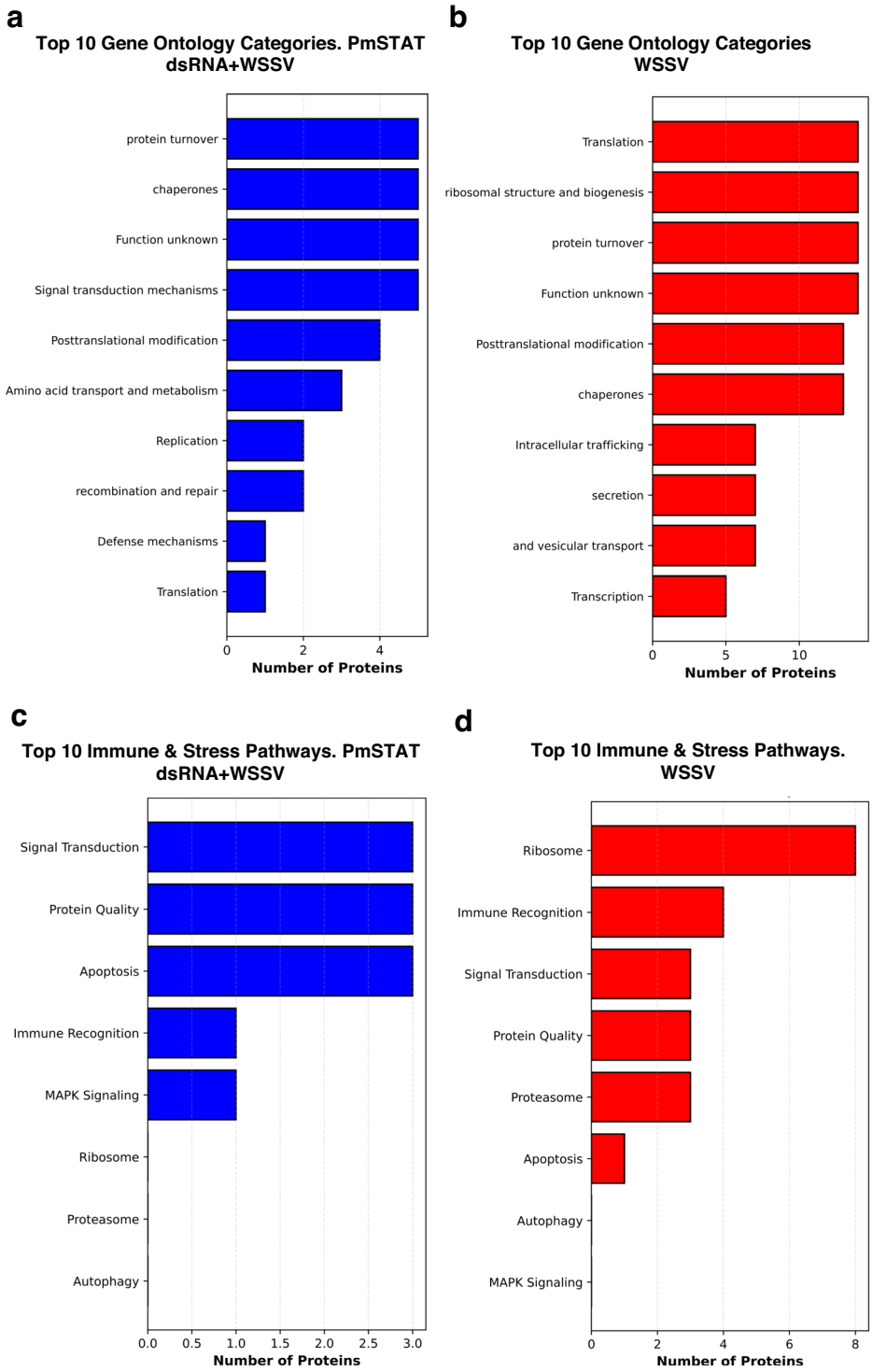

Fig. 7 Gene Ontology (GO) annotations and Immune & Stress related Pathways of DEGs. The top 10 GO functional terms are shown for (**a**) *Pm*STAT dsRNA + WSSV and (**b**) WSSV conditions. The top 10 Immune & Stress related Pathways are shown for (**c**) *Pm*STAT dsRNA + WSSV and (**d**) WSSV conditions. The x-axis indicates the number of proteins annotated to each term/pathway.

repetitiveness (~1–3%). Together, these BUSCO results support that the assembly is highly complete and that isoform filtering is critical for correctly interpreting duplication metrics in transcriptome-based BUSCO analyses.

Robustness of DEGs detection and sample clustering under equal resampling. To assess whether the original DEG detection could be influenced by unequal number of samples across groups (WSSV, n = 5; *Pm*STAT dsRNA, n = 3; *Pm*STAT dsRNA + WSSV, n = 4), we performed a sensitivity analysis based on systematic

independent resampling to the smallest group size (n = 3). Thus, we generated all possible within-group triplets of samples and analyzed every independent resampling (5C3 × 3C3 × 4C3 = 40 resampling sets). For each resampling set, we conducted the differential-expression analysis to obtain DEGs lists under equal replicate numbers for each group using DESeq2. Across all 40 resampling sets, hierarchical clustering using normalized expression values of the DEGs was fully consistent: in every resampling test, the three selected samples from each group always clustered together within their respective treatment (Table S2).

Additionally, the differential-expression results were highly stable across resamplings: no significant DEGs were detected between the *Pm*STAT dsRNA Vs. *Pm*STAT dsRNA + WSSV groups in any of the 40 resamplings, reinforcing that this result is robust and independent of sample size. On the other hand, the comparison of *Pm*STAT dsRNA + WSSV Vs. WSSV consistently yielded differentially expressed genes across all 40 resamplings. We further quantified the robustness of the original DEGs results by constructing a gene-level "confidence" matrix using the DEGs identified in the complete analysis as a reference. For each reference DEG, we computed a reproducibility score defined as the proportion of the 40 resampling sets in which it remained significantly differentially expressed with the same direction of change (Table S3). All 138 genes originally classified as upregulated for *Pm*STAT dsRNA + WSSV in the full dataset (Table S1) were recovered as significant in at least one of the 40 down-sampled combinations (Table S3), whereas for the 675 upregulated genes in WSSV (Table S1), 349 were recovered in at least one subsampling set (Table S3). This resampling strategy provides a quantitative measure of DEG stability (reproducibility score) under equal matched group sizes (Table S3). Overall, these results indicated that the upregulated response in *Pm*STAT dsRNA + WSSV relative to WSSV is highly robust in spite of sample imbalance, while a subset of upregulated signals in WSSV appears more sensitive to the imbalance of samples.

Of the 138 genes upregulated in the *Pm*STAT dsRNA + WSSV group, only 47 transcripts encoded proteins, whereas among the 349 genes upregulated in the WSSV group, 88 encoded proteins (https://github.com/LuiguiGallardo/p_monodon_transcriptome/tree/master/Additional_Tables). Figures 6, 7, and Table S4 show the summary of annotations for both sets of DEGs coding for proteins using eggNOG, KEGG, and GO terms/pathways.

Technical validation of DEGs by qRT–PCR. To support the robustness of the RNA-seq expression estimates, we performed qRT–PCR validation using hemocyte-derived cDNA from the same samples used for RNAseq. A targeted set of nine immune-related DEGs spanning multiple functional categories was selected: leukocyte elastase inhibitor A-like, legumain-like, C-reactive protein 1-like, ankyrin repeat domain-containing protein 12-like, peptidyl-prolyl cis–trans isomerase A, lactoylglutathione lyase-like, prophenoloxidase 1 (ProPO1), prophenoloxidase 2 (ProPO2), and Prophenoloxidase-activating enzyme (PPAE2). Relative expression was quantified using the $2^{-\Delta\Delta Cq}$ method using the housekeeping elongation factor 1α (EF1α) gene as internal control. Primer sequences for all genes were reported in Table 1. Overall, qRT–PCR results were concordant with RNA-seq in the direction of change and showed consistent trends across comparisons (Fig. 5b), supporting the robustness of the dataset for downstream analyses and reuse. The quantitative measure of DEG stability (reproducibility score) under equal matched group sizes for these nine genes is in Table S3.

Data availability

Any additional data are available from the corresponding author upon reasonable request.

Code availability

All software and parameters used for preprocessing, *de novo* assembly, quantification, annotation, and differential expression are listed below (versions in parentheses). Non-default settings are indicated.

- Quality control & filtering: FastQC (v0.12; Q > 20 threshold) and Trimmomatic (v0.36; SLIDINGWINDOW:6:20)[29,30].

- *De novo* assembly: Trinity (v2.13.1)[21].

- Read realignment & quantification: Bowtie2 (v2.5.3) via Trinity's align_and_estimate_abundance.pl[27]; RSEM (v1.3.3) to estimate transcript/gene abundance[29]; matrices via abundance_estimates_to_matrix.pl.

- Differential expression: DESeq2 (v1.32.0) using Trinity's analyze_diff_expr.pl; thresholds: |log2 fold change| > 2 and FDR < 0.05[31].

- Coding sequence prediction: TransDecoder (v5.5.0), retaining ORFs ≥ 100 aa[32].

All analysis pipelines and custom scripts are available at: https://github.com/LuiguiGallardo/p_monodon_transcriptome.

Received: 30 September 2025; Accepted: 27 March 2026;
Published online: 07 April 2026

References

1. Food and Agriculture Organization of the United Nations (FAO). The State of World Fisheries and Aquaculture - Blue Transformation in Action. Rome.
2. Tran, L. *et al.* Determination of the infectious nature of the agent of acute hepatopancreatic necrosis syndrome affecting penaeid shrimp. *Diseases of Aquatic Organisms* **105**, 45–55, https://doi.org/10.3354/dao02621 (2013).
3. Chaijarasphong, T. *et al.* The shrimp microsporidian Enterocytozoon hepatopenaei (EHP): Biology, pathology, diagnostics and control. *Journal of Invertebrate Pathology* **186**, 107458, https://doi.org/10.1016/j.jip.2020.107458 (2021).
4. Chen, W. Y. *et al.* WSSV infection activates STAT in shrimp. *Developenntal and Comparative Immunology* **32**, 1142–1150, https://doi.org/10.1016/j.dci.2008.03.003 (2008).
5. Sánchez-Paz, A. White spot syndrome virus: an overview on an emergent concern. *Veterinary research* **41**, 43 (2010).
6. Chou, H. Y. *et al.* Pathologenicity of a baculovirus infection causing white spot syndrome in cultured shrimp in Taiwan. *Aquaculture* **23**, 161–173 (1995).
7. Wang, C. H. *et al.* Purification and genomic analysis of baculovirus associated with white spot syndrome (WSBV) of Penaeus monodon. *Diseases of Aquatic Organisms* **23**, 239–242 (1995).

8. Jiravanichpaisal, P. *et al.* Antibacterial peptides in hemocytes and hematopoietic tissue from freshwater crayfish Pacifastacus leniusculus: characterization and expression pattern. *Developental and Comparative Immunology* **31**, 441–455 (2007).
9. Arbouzova, N. I. & Zeidler, M. P. JAK/STAT signalling in Drosophila: insights into conserved regulatory and cellular functions. *Development* **133**, 2605–2616, https://doi.org/10.1242/dev.02411 (2006).
10. Kiu, H. & Nicholson, S. E. Biology and significance of the JAK/STAT signalling pathways. *Growth Factors* **30**, 88–106 (2012).
11. Brown, S., Hu, N. & Hombría, J. C. G. Identification of the first invertebrate interleukin JAK/STAT receptor, the Drosophila gene domeless. *Current Biology* **11**, 1700–1705, https://doi.org/10.1016/s0960-9822(01)00524-3 (2001).
12. Hou, X. S., Melnick, M. B. & Perrimon, N. marelle Acts Downstream of the Drosophila HOP/JAK Kinase and Encodes a Protein Similar to the Mammalian STATs. *Cell* **84**, 411–419, https://doi.org/10.1016/s0092-8674(00)81286-6 (1996).
13. Makki, R. *et al.* A Short Receptor Downregulates JAK/STAT Signalling to Control the Drosophila Cellular Immune Response. *PLoS Biology* **8**, https://doi.org/10.1371/journal.pbio.1000441 (2010).
14. Yan, R. *et al.* Identification of a Stat Gene That Functions in Drosophila Development. *Cell* **84**, 421–430, https://doi.org/10.1016/s0092-8674(00)81287-8 (1996).
15. Agaisse, H. & Perrimon, N. The roles of JAK/STAT signaling in Drosophila immune responses. *Immunological Reviews* **198**, 72–82 (2004).
16. Liu, W. J. *et al.* White spot syndrome virus annexes a shrimp STAT to enhance expression of the immediate-early gene ie1. *Journal of Virology* **81**, 1461–1471, https://doi.org/10.1128/jvi.01880-06 (2007).
17. Laohawutthichai, P. *et al.* Effects of PmDOME and PmSTAT knockdown on white spot syndrome virus infection in Penaeus monodon. *Scientific Reports* **13**, 9852, https://doi.org/10.1038/s41598-023-37085-1 (2023).
18. Wu, Z., Chu, K. H. & Ma, K. Y. Transcriptome Analysis of Multiple Tissues in the Shrimp Penaeus vannamei Reveals the Typical Physiological Response to Three Pathogens. *Journal of Marine Science and Engineering* **11**, 389 (2023).
19. Wang, F. *et al.* Transcriptome analysis reveals the activation of neuroendocrine-immune system in shrimp hemocytes at the early stage of WSSV infection. *BMC Genomics* **20**, 247, https://doi.org/10.1186/s12864-019-5614-4 (2019).
20. Faul, F. *et al.* Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behavior Research Methods* **41**, 1149–1160, https://doi.org/10.3758/BRM.41.4.1149 (2009).
21. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**, 644–652, https://doi.org/10.1038/nbt.1883 (2011).
22. Manni, M. *et al.* BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution* **38**(10), 4647–4654, https://doi.org/10.1093/molbev/msab199 (2021). **27**.
23. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP499992 (2024).
24. Chen, I.-T. *et al.* White Spot Syndrome Virus Induces Metabolic Changes Resembling the Warburg Effect in Shrimp Hemocytes in the Early Stage of Infection. *Journal of Virology* **85**(24), 12919–28, https://doi.org/10.1128/JVI.05385-11 (2011).
25. Wen, R. *et al.* Function of shrimp STAT during WSSV infection. *Fish & Shellfish Immunology* **38**(2), 354–60, https://doi.org/10.1016/j.fsi.2014.04.002 (2014).
26. Van Hulten, M. C. *et al.* White spot syndrome virus envelope protein VP28 is involved in the systemic infection of shrimp. *Virology* **5**;285(2):228-33. https://doi.org/10.1006/viro.2001.0928 (2001).
27. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357–359, https://doi.org/10.1038/nmeth.1923 (2012).
28. Huerlimann, R. *et al.* De novo assembly, characterization, functional annotation and expression patterns of the black tiger shrimp (Penaeus monodon) transcriptome. *Scientific Reports*. **8**, 13553, https://doi.org/10.1038/s41598-018-31148-4 (2018).
29. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)* **30**, 2114–2120, https://doi.org/10.1093/bioinformatics/btu170 (2014).
30. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **12**, 323, https://doi.org/10.1186/1471-2105-12-323 (2011).
31. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 *Genome Biology* **15**, 550, https://doi.org/10.1186/s13059-014-0550-8 (2014).
32. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols* **8**, 1494–1512, https://doi.org/10.1038/nprot.2013.084 (2013).

Acknowledgements

We acknowledge Chulalongkorn University for supporting T.J. and P.L. through the Ratchadapisek Somphot Fund for Postdoctoral Fellowships. Also Luigui Gallardo-Becerra (CVU 778192) would like to thank CONACyT for the doctoral fellowship, and the postdoctoral support by Estancias Posdoctorales por México 2022 program for Fernanda Cornejo-Granados (CVU 443238). We thank the Unidad Universitaria de Secuenciación Masiva y Bioinformática of the Laboratorio Nacional de Apoyo Tecnológico a las Ciencias Genómicas, UNAM, especially to Ricardo Alfredo Grande Cano and Lizeth A. Matías Valdez for the technical sequencing support. This research was funded by Thailand Science Research and Innovation Fund (Chulalongkorn University, FF_67), Center of Excellence in Structural and Computational Biology (to K.K) and Center of Excellence for Molecular Biology and Genomics of Shrimp (to A.T.). This research was also funded by CONAHCYT Fronteras de la Ciencia CF2019-G-263986 and DGAPA PAPIIT UNAM (IN219723 and IN222526).

Author contributions

T.J. and P.L. conducted gene silencing and qRT PCR experiments. T.J. analyzed the transcriptome data with help from L.G.B., F.C.G. and J.M.H.R. A.T. provided resources and equipment. A.O.L. supervised T.J. and analyzed the data. K.K. conceived and supervised the project and contribute to data analysis. T.J., P.L., K.K. and F.C.G. wrote the first draft. A.O.L. and K.K. edited the manuscript. All authors reviewed the results and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-026-07153-y.

Correspondence and requests for materials should be addressed to A.O.-L. or K.K.