Data in Brief 36 (2021) 107053

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

# RNA-Seq dataset of thoracic ganglia transcriptome across four ovarian development stages in *Fenneropenaeus merguiensis*

Prasert Yodsawat[a,c], Jiratchaya Nuanpirom[a,c], Ponsit Sathapondecha[a,b,c], Unitsa Sangket[a,b,c,*]

[a] *Department of Molecular Biotechnology and Bioinformatics, Faculty of Science, Prince of Songkla University, Hat Yai, Songkhla, Thailand*
[b] *Center for Genomics and Bioinformatics Research, Faculty of Science, Prince of Songkla University, Hat Yai, Songkhla, Thailand*
[c] *Division of Biological Science, Faculty of Science, Prince of Songkla University, Hat Yai, Songkhla, Thailand*

ARTICLE INFO

*Article history:*
Received 13 January 2021
Revised 27 February 2021
Accepted 7 April 2021
Available online 20 April 2021

*Keywords:*
Banana shrimp
RNA sequencing
Transcriptome
Vitellogenesis
Crustacean
Sexual reproduction
Thoracic ganglion
*De novo* assembly

ABSTRACT

Banana shrimp (*Fenneropenaeus merguiensis*) is an economically important shrimp in marine aquaculture. Although there is plenty of transcriptome research for this species, the molecular mechanisms in thoracic ganglia of banana shrimp during ovarian maturation have not yet been investigated. Here we report the transcriptomic data of female banana shrimp obtained from thoracic ganglia during ovarian developmental stages. The samples were collected from four stages of ovarian development with two individual shrimps per stage. Total RNA was extracted and used to prepare the sequencing library. Approximately 188 million pair-end raw reads, ranging from 21 to 31 million reads for each library, were generated using an Illumina HiSeq 2500 platform. Quality control was applied to the raw reads before the assembly process. After *de novo* assembly, the final transcript assembly was generated by vector decontamination, coding regions prediction, redundancy reduction, and foreign sequence depletion. A total of 77,681 transcripts, ranging between 255 and 23,016 bp with an N50 value of 1,167 were obtained to

* Corresponding author.
*E-mail address:* unitsa.s@psu.ac.th (U. Sangket).

https://doi.org/10.1016/j.dib.2021.107053

the final assembly. Finally, the final transcripts assembly was evaluated by calculated assembly completeness with Arthropoda orthologous genes dataset. A total of 92.1% of Arthropoda orthologous genes were found in our final assembly. These data might provide benefits for gene discovery, gene annotation, transcript profiling, and other research topics in the context of banana shrimp.



## Specifications Table

| | |
|---|---|
| Subject | Biochemistry, Genetics and Molecular Biology (General) |
| Specific subject area | Aquaculture, Bioinformatics, Crustacean reproductive neuroendocrinology, Transcriptomics |
| Type of data | Raw sequencing data<br>Transcriptome assembly<br>HTML<br>Table<br>Figure |
| How data were acquired | RNA extraction, Oligo-dT beads selection, Illumina HiSeq 2500 system, FastQC software version 0.11.8, Cutadapt software version 2.7, MultiQC software version 1.8, Trinity software version 2.8.5, VecScreen standalone software, BLAST+ software version 2.10.1, SeqKit software version 0.14.0, TransDecoder software version 5.5.0, CD-HIT software version 4.8.1, BUSCO software version 4.1.4 |
| Data format | Raw RNA sequencing data in FASTQ format<br>Final transcriptome assembly sequences in FASTA format<br>Quality control statistics of raw RNA sequencing data in HTML format |
| Parameters for data collection | The adult females of *F. merguiensis* were separated into four groups by ovarian maturation stages: previtellogenic, early vitellogenic, vitellogenic, and late vitellogenic. Thoracic ganglia samples from two individual shrimps per stage were collected. The samples were stored in RNAlater before RNA extraction. |
| Description of data collection | Total RNA from each sample was extracted using RNeasy Micro Kit (Qiagen, USA). Oligo(dT) beads were used to select the mRNA. The sequencing library was prepared by mRNA fragmentation, cDNA synthesis, adapter ligation, PCR, and size selection. Eight sequencing libraries were sequenced using Illumina HiSeq 2500 system. Paired-end raw sequencing reads were quality controlled by FASTQC, Cutadapt, and MultiQC software. Clean reads were subjected to *de novo* transcriptome assembly using Trinity software. Transcripts that have vector or adapter contamination were removed using VecScreen and SeqKit software. Coding regions were identified using TransDecoder software. CD-HIT software was used to reduce the redundant transcripts. Foreign sequences were excluded from final transcripts using SeqKit software. The final transcripts were evaluated using BUSCO software. Assembly statistics was obtained using SeqKit software. |
| Data source location | Institution: Department of Molecular Biotechnology and Bioinformatics, Division of Biological Science, Faculty of Science, Prince of Songkla University<br>City/Town/Region: Hat Yai, Songkhla<br>Country: Thailand |
| Data accessibility | RNA-Seq data in this paper was published at NCBI BioProject with accession number PRJNA611903 (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA611903). The published data contains raw sequence reads available at Sequence Read Archive (SRA) database with accession number SRA1157726 (https://www.ncbi.nlm.nih.gov/sra/?term=SRA1157726). The final transcript assembly has been deposited at DDBJ/EMBL/GenBank as Transcriptome Shotgun Assembly (TSA) under the accession number GIXQ00000000. The version described in this paper is the first version, GIXQ01000000 |

(*continued on next page*)

(https://www.ncbi.nlm.nih.gov/nuccore/GIXQ00000000.1). The supplementary data, analysis commands, and related files are available in the Zenodo repository (https://doi.org/10.5281/zenodo.4561158).

## Value of the Data

- The transcriptome of the thoracic ganglia of *F. merguiensis* during ovarian development is provided. This allows partial fulfillment of the public database of this species, including nucleotide sequences and annotations useful for sequence analysis.
- These datasets are potentially useful to researchers in the field of crustacean reproductive neuroendocrinology.
- These datasets could facilitate data-driven research in this field. For example, transcript isoform discovery, transcriptome expression profiling, and structural function determination of homolog genes.

## 1. Data Description

An overview of sample collection and data analysis pipeline is presented in Fig. 1. The transcriptome data of thoracic ganglia for female *F. merguiensis* was generated from four stages of ovarian development: previtellogenic, early vitellogenic, vitellogenic, and late vitellogenic. The thoracic ganglia samples were collected from two individual shrimps per stage. Then, the samples were used to isolate the total RNA and only mRNA was collected. The transcriptome sequencing library was constructed and sequenced using Illumina HiSeq 2500 system. A total of 188 million paired-end reads were generated in FASTQ format. Raw transcriptome sequence reads were deposited under the NCBI Sequence Read Archives (SRA) database (SRA1157726). A detailed summary of sequencing strategy is provided in Table 1. The quality control of raw transcriptome data was performed by removing low-quality bases and adapter sequences to produce clean reads. Summary of quality statistics of both raw reads and clean reads in HTML format were available in Zenodo repository (https://doi.org/10.5281/zenodo.4561158). Clean reads were *de novo* assembled to produce the raw transcripts. To generate the final transcripts, the transcripts were scanned for vector contamination. Then, the coding regions within transcripts were extracted, and redundant transcripts were also depleted. The transcripts were finally curated by excluding the foreign sequence contamination. The final transcript sequences have been deposited at DDBJ/EMBL/GenBank as Transcriptome Shotgun Assembly (TSA) under the accession GIXQ00000000. Table 2 demonstrates the general statistics of transcripts from the raw assembly compared to those from the final assembly.

Finally, the final transcripts were evaluated for completeness based on expectations of gene content against Arthropoda orthologous genes dataset. The completeness result is provided in Table 3. Additionally, the commands and related files were also deposited in Zenodo repository (https://doi.org/10.5281/zenodo.4561158).

## 2. Experimental Design, Materials and Methods

### *2.1. Sample collection and ovarian development stage separation*

The adult *F. merguiensis* females were wild, caught from the Gulf of Thailand, Nakhon Si Thammarat, Thailand. The collected shrimps were then acclimatized in 30 ppt seawater for 3 days before sorting by ovarian maturation stages: stage 0 (previtellogenic), stage 1 (early vitellogenic), stage 2 (vitellogenic) and stage 3 (late vitellogenic) (Fig. 1). The banana shrimps were

**Table 1**
Summary of sequencing strategies.

| Sample name | Replicate number | Development stage | Stage abbreviation | Instrument platform | Instrument model | Library strategy | Library layout |
|---|---|---|---|---|---|---|---|
| TgS0_rep1 | 1 | Previtellogenic | Stage 0 | ILLUMINA | Illumina HiSeq 2500 | RNA-Seq | Paired |
| TgS0_rep2 | 2 | Previtellogenic | Stage 0 | ILLUMINA | Illumina HiSeq 2500 | RNA-Seq | Paired |
| TgS1_rep1 | 1 | Early vitellogenic | Stage 1 | ILLUMINA | Illumina HiSeq 2500 | RNA-Seq | Paired |
| TgS1_rep2 | 2 | Early vitellogenic | Stage 1 | ILLUMINA | Illumina HiSeq 2500 | RNA-Seq | Paired |
| TgS2_rep1 | 1 | Vitellogenic | Stage 2 | ILLUMINA | Illumina HiSeq 2500 | RNA-Seq | Paired |
| TgS2_rep2 | 2 | Vitellogenic | Stage 2 | ILLUMINA | Illumina HiSeq 2500 | RNA-Seq | Paired |
| TgS3_rep1 | 1 | Late vitellogenic | Stage 3 | ILLUMINA | Illumina HiSeq 2500 | RNA-Seq | Paired |
| TgS3_rep2 | 2 | Late vitellogenic | Stage 3 | ILLUMINA | Illumina HiSeq 2500 | RNA-Seq | Paired |

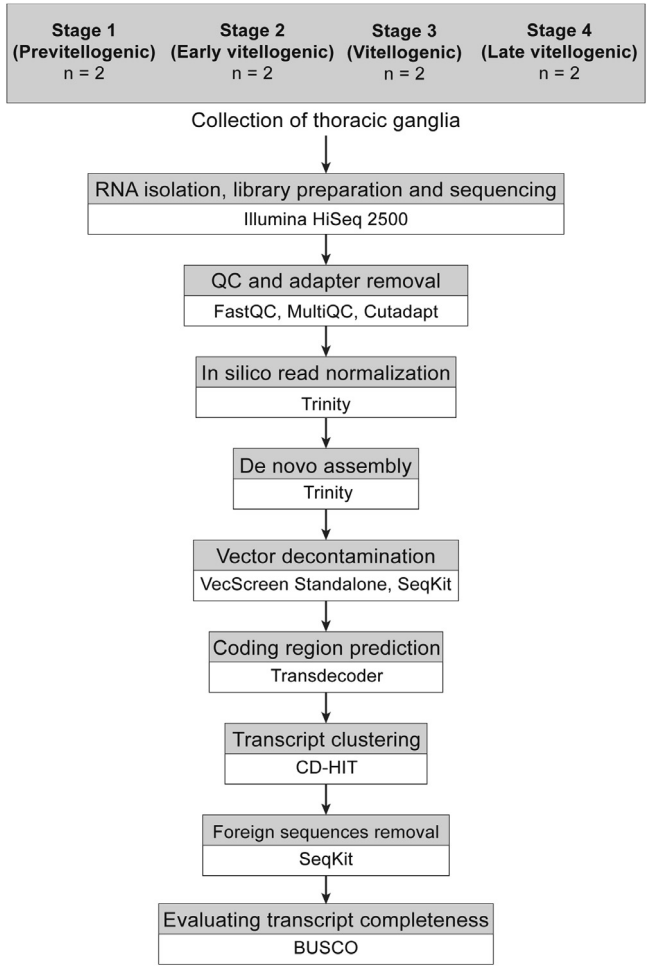


**Fig. 1.** An overview of the sample collection and data analysis pipeline.

**Table 2**
Assembly statistics.

| Metrics | Raw assembly | Final assembly |
|---|---|---|
| Number of transcripts | 428,299 | 77,681 |
| Number of bases (bp) | 384,387,798 | 65,962,293 |
| Average transcript (bp) | 897.5 | 849.1 |
| Longest transcript (bp) | 76,897 | 23,016 |
| Shortest transcript (bp) | 177 | 255 |
| N50 (bp) | 1862 | 1167 |

then dissected to collect the thoracic ganglia preserved in RNAlater™ Stabilization Solution (Ambion, USA) to prevent RNA degradation from RNase activity.

**Table 3**
Transcript completeness from BUSCO analysis.

| BUSCO evaluation metrics | Result |
|---|---|
| Complete and single-copy | 743 (73.3%) |
| Complete and duplicated | 190 (18.8%) |
| Fragmented | 10 (1.0%) |
| Missing | 70 (6.9%) |
| Total BUSCO groups searched | 1013 (100%) |

### *2.2. RNA extraction, library preparation, and sequencing*

Total RNAs from thoracic ganglia were extracted using RNeasy Micro Kit (Qiagen, USA) following the manufacturer's instructions. The preliminary quality measurement of the total RNA sample was performed by agarose gel electrophoresis to detect RNA degradation and potential contamination. The sample quantity and purity were evaluated by the Nanodrop spectrophotometer (Thermo Scientific, USA). The sample integrity was evaluated by Bioanalyzer 2100 (Agilent Technologies, USA). The mRNA was selected using oligo(dT) beads. To prepare the transcriptome sequencing library, mRNA was randomly fragmented by fragmentation buffer. The cDNA was then synthesized using random hexamer primers, a second-strand synthesis buffer (Illumina, USA), dNTPs, RNase H, and DNA polymerase I to initiate the reaction. The Illumina sequencing adapters were then ligated to the terminals of the cDNA strand, and the library preparation was completed by PCR and size selection. After that, eight cDNA libraries were sequenced using Illumina HiSeq 2500 obtaining around 6 Gbases per library.

### *2.3. RNA-Seq data analysis*

In this work, we mainly used Bioconda software [1] to manage the software installation and environment for each part of the analysis. Raw RNA-Seq reads were first quality checked using FastQC software (v. 0.11.8) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and then MultiQC software (v. 1.8) [2] was used to summarize the quality results into a single HTML file. Next, Cutadapt software (v. 2.7) [3] was used to remove the sequencing adapters, low-quality sequences (–quality-cutoff 20,20), and short sequences (–minimum-length 25) to produce clean reads. The quality of clean reads was checked again using FastQC software and summarized with MultiQC software. Before start of assembly, clean RNA-Seq reads were normalized based on sequencing depth using "insilico_read_normalization.pl" script file included in Trinity software (v. 2.8.5) [4]. The maximum targeted coverage was set to 30 (–max_cov 30) with a default k-mer size ($k = 25$). Each RNA-Seq read was selected based on its median k-mer coverage and the targeted maximum coverage. Both forward and reverse reads were altogether selected, and the singletons were removed using parameter –pairs_together. After normalizing clean RNA-Seq reads, Trinity software was used to *de novo* assemble, and all parameters were set to default without in silico read normalization process (–no_normalize_reads). The assembled transcripts from Trinity software were considered raw transcripts. After assembly, VecScreen standalone software released on 2012–03–23 (https://ftp.ncbi.nlm.nih.gov/blast/demo/vecscreen) was used to search for any vector or adapter contamination in raw transcripts against UniVec database released on 2017–03–20 (https://ftp.ncbi.nlm.nih.gov/pub/UniVec/UniVec) (retrieved on 10 December 2020). VecScreen used blastn from NCBI BLAST+ software [5] with pre-set parameters for optimal detection of vector contamination (-task blastn -reward 1 -penalty −5 -gapopen 3 -gapextend 3 -dust yes -soft_masking true -evalue 700 -searchsp 1,750,000,000,000). After searching for contamination, a script "VSlistTo1HitPerLine.awk" released on 2006–06–09 (https://ftp.ncbi.nlm.nih.gov/pub/kitts/VSlistTo1HitPerLine.awk) was used to categorize the VecScreen results based on the location and strength of the vector matches. Transcripts with the "Strong" and the "Moderate" hits categories were filtered out using SeqKit software (v. 0.14.0)

[6]. Next, TransDecoder software (v. 5.5.0) (https://github.com/TransDecoder/TransDecoder) was used to identify the coding regions in transcripts. TransDecoder extracted the long open reading frames (ORF) with at least 100 amino acid in length and selected a single best ORF per transcript (–single_best_only). To reduce the transcripts' redundancy, CD-HIT software (v. 4.8.1) [7] was used to cluster and remove the redundant transcripts with 95% local sequence identity threshold and 100% alignment coverage for the shorter sequence (-c 0.95 -G 0 -aS 1.00). The transcripts were submitted to NCBI TSA submission portal to screen for foreign sequence contamination. The final transcripts were generated by excluding the transcripts marked as contamination during TSA submission using SeqKit software. To assess the final transcripts' completeness, BUSCO (Benchmarking Universal Single-Copy Orthologs) software (v. 4.1.4) [8] was used to evaluate the completeness based on evolutionarily informed expectations of gene content by comparing the final transcripts to Arthropoda orthologous genes dataset (–lineage_dataset arthropoda_odb10). The sequence statistics of transcripts was obtained using SeqKit software. All analysis commands and related files used in this work are publicly available in Zenodo repository (https://doi.org/10.5281/zenodo.4561158).

## Ethics Statement

All animal samples and trials were approved by the Institutional Animal Care and Use Committee, Prince of Songkla University (2561–01–082), and the animal use procedure and ethic were under the regulation of animals for scientific purposes Act, BE 2558.

## CRediT Author Statement

**Prasert Yodsawat:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization, Project administration; **Jiratchaya Nuanpirom:** Conceptualization, Investigation, Resources, Writing - Original Draft, Writing - Review & Editing, Visualization; **Ponsit Sathapondecha:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Writing - Review & Editing, Funding acquisition; **Unitsa Sangket:** Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Data Curation, Writing - Review & Editing, Project administration, Funding acquisition.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

## Acknowledgments

This work was supported by the Faculty of Science Research Fund, Prince of Songkla University, Contract no. 1-2561-02-007; and the Research Fund, Research and Development Office, Prince of Songkla University, Contract no. SCI6202049b and SCI6202049M. The authors would like to thank Assoc. Prof. Seppo Karrila and the Research and Development Office, Prince of Songkla University for polishing the written English.

## References

[1] B. Grüning, R. Dale, A. Sjödin, B.A. Chapman, J. Rowe, C.H. Tomkins-Tinch, R. Valieris, J. Köster, Bioconda: sustainable and comprehensive software distribution for the life sciences, Nat. Methods 15 (2018) 475–476, doi:10.1038/s41592-018-0046-7.

[2] P. Ewels, M. Magnusson, S. Lundin, M. Käller, MultiQC: summarize analysis results for multiple tools and samples in a single report, Bioinformatics 32 (2016) 3047–3048, doi:10.1093/bioinformatics/btw354.

[3] M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads, EMBnet. J 17 (2011) 10–12, doi:10.14806/ej.17.1.200.

[4] M.G. Grabherr, B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B.W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Full-length transcriptome assembly from RNA-Seq data without a reference genome, Nat. Biotechnol. 29 (2011) 644–652, doi:10.1038/nbt.1883.

[5] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T.L. Madden, BLAST+: architecture and applications, BMC Bioinformatics 10 (2009) 421, doi:10.1186/1471-2105-10-421.

[6] W. Shen, S. Le, Y. Li, F. Hu, SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation, PLoS ONE 11 (2016) e0163962, doi:10.1371/journal.pone.0163962.

[7] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for clustering the next-generation sequencing data, Bioinformatics 28 (2012) 3150–3152, doi:10.1093/bioinformatics/bts565.

[8] M. Seppey, M. Manni, E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness, in: M. Kollmar (Ed.), Gene Prediction: Methods and Protocols, Springer New York, New York, NY, 2019, pp. 227–245, doi:10.1007/978-1-4939-9173-0_14.