Data Article

The first dataset of de novo transcriptome assembly of *Heterotrigona itama* (Apidae, Meliponinae) queen larva

Amin Asyraf Tamizi*, Nazrul Hisham Nazaruddin, Wee Chien Yeong, Muhammad Faris Mohd Radzi, Mohd Azwan Jaafar, Rogayah Sekeli

Agri-Omics and Bioinformatics Programme, Biotechnology and Nanotechnology Research Centre, Malaysian Agricultural Research and Development Institute (MARDI), 43400 Serdang, Selangor, Malaysia

**ABSTRACT**

*Heterotrigona itama* is a species of stingless bee recently domesticated (or reared) for honey production in a few Southeast Asian countries namely Malaysia and Indonesia. Being categorized in the clade Corbiculata together with the honeybees (*Apis* spp.) and bumble bees (*Bombus* spp.), the stingless bees are highly social in which the colony members are subjected to labor division where a queen functions as the reproductive caste. In this data article, we provide a resource encompassing a transcriptome profile (de novo assembled) from *H. itama* queen larva — the first report of transcriptome assembly for this species. The generated data is pivotal for the characterization of important genes and biological pathways in order to further improve our understanding on the developmental biology, behavior, social structure and ecological needs of this eusocial hymenopteran insect from the molecular aspect. The raw RNA sequencing data is available at NCBI Sequence Read Archive (SAR) under the accession number SRP230250 and the assembled reads are deposited at DDBJ/EMBL/Genbank as Transcriptome Shotgun Assembly (TSA) under the accession GIIH00000000.

© 2020 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

* Corresponding author.
  E-mail address: aminasyraf@mardi.gov.my (A.A. Tamizi).

https://doi.org/10.1016/j.dib.2020.105235
2352-3409/© 2020 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
1. Data

The dataset contains raw sequencing data obtained through the transcriptome sequencing of a female H. itama larva from the queen caste. A queen larva at feeding stage (3rd instar) was collected from a mother colony and snap-frozen in liquid nitrogen (N2), and high quality total RNA was extracted before sent for sequencing through paired-end Illumina sequencing technology. De novo assembly was performed using Trinity and contigs were annotated against seven databases using multiple software. An overview of the data and sequencing assembly of H. itama data is presented in Table 1 and Table 2. The raw data file (reads in FASTQ format) was deposited at NCBI SRA database under accession no. SRP230250 and the transcriptome assembly data was deposited at NCBI TSA with accession number GIIH00000000.

2. Experimental design, material and methods

2.1. Insect material and RNA extraction

The queen larva, whose species had been validated through DNA barcode (cytochrome oxidase subunit I), was picked from a very healthy colony that had been placed at a dedicated research plot adjacent to a secondary forest for more than 6 months. Larva instar 3 (feeding stage) was selected for RNA sequencing since this stage was reported to be the ‘deciding stage’ for caste differentiation in female stingless bee larvae [1]. Total RNA was extracted from whole larva using an optimized protocol and sent for sequencing.

2.2. RNA preparation and sequencing

Prior to Complementary DNA (cDNA) libraries preparation and sequencing, quality check (QC) of the total RNA was done as follows: preliminary quantitation (Nanodrop), degradation and contamination
Table 1
Data production of RNA-seq.

| Sample                      | Raw Reads | Clean reads | Clean bases | Error (%) | Q20 (%) | Q30 (%) | GC (%) |
|-----------------------------|-----------|-------------|-------------|-----------|---------|---------|--------|
| Queen larva (3rd instar)    | 102300462 | 99370410    | 14.9G       | 0.01      | 97.48   | 93.50   | 42.18  |

Q20: percentages of bases whose correct base recognition rates are greater than 99% in total bases.
Q30: percentages of bases whose correct base recognition rates are greater than 99.9% in total bases.

Table 2
Overview of transcripts and unigenes assembled using Trinity (version r2014-04-13p1).

| Attributes       | Transcripts | Unigenes |
|------------------|-------------|----------|
| Min. length      | 201         | 201      |
| Max length       | 58291       | 58291    |
| Mean length      | 1837        | 1837     |
| Median length    | 904         | 904      |
| N50              | 3588        | 3588     |
| N90              | 717         | 718      |
| Total nucleotides| 101280555   | 101277047|
| **Total number** | **55148**   | **55135**|

Fig. 1. GO classification distribution of annotated genes from *H. itama* queen larva.
The RNA was then processed following these steps: Enrichment of mRNA using oligo(dT) beads, removal of rRNA using a specialized kit and fragmentation of mRNA. Afterward, the cDNA was synthesized from the mRNA fragments using random hexamers and reverse transcriptase. Following the first-strand synthesis, a custom second-strand synthesis buffer (Illumina) is added together with dNTPs, RNase H and *Escherichia coli* polymerase I to generate the second strand by nick-translation followed by two rounds of cDNA purification using AMPure XP beads. The cDNA was then proceeded for terminal repair, A-tailing, ligation of sequencing adapters, size selection and PCR enrichment. For library quality assessment, the cDNA library concentration was determined using Qubit 2.0 fluorometer (Life Technologies), and the insert size was checked on Agilent 2100 and quantified to greater accuracy by quantitative PCR (qPCR) (library activity >2 nM). Finally, the prepared cDNA libraries were fed into Illumina machines according to activity and expected data volume.

### 2.3. RNA-seq data analysis (raw reads handling and de novo assembly) and gene annotation

The raw data from Illumina was transformed to Sequence Reads by base calling and recorded in a FASTQ file. Raw reads were cleaned/filtered as follows: (1) removing reads with adaptor contamination, (2) removing reads when uncertain nucleotides constitute more than 10% of either read (N > 10%) and (3) removing reads when low quality nucleotides (base quality less than 20) constitute more than 50% of the read. De novo transcriptome reconstruction was carried out using Trinity (version r2014-04-13p1) with a minimum read length of 200 and k-mer = 25. The Trinity workflow followed the Inchworm, Chrysalis and Butterfly modules [2]. The summary of sequencing and assembly data are tabulated in Tables 1 and 2. The contigs were then clustered with Corset [3] to remove redundancy.

Gene functional annotations were carried out using Diamond (v0.8.22), KAAS (r14 0224), NCBI Blast (v2.2.28+), hmmscan (HMMER 3) and blast2go (b2g4 pipe_v2.5) software. A total of seven databases including Nr, Nt, KO, Swiss-Prot, Pfam, GO (Fig. 1) and KOG were used to annotate the contigs and 55,135 of unigenes had been successfully annotated (Table 3).

### Acknowledgements

This project was funded by National Conservation Trust Fund (NCTF), Ministry of Water, Land and Natural Resources (KATS) of Malaysia [KAT(S)600-2/1(48/2J; RB6011NC10]. We would like to extend our appreciation to Apical Scientific Sdn. Bhd. in aiding us for RNA sequencing and QC checks.

### Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.
Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dib.2020.105235.

References

[1] C.A.M. Cardoso-Junior, R.P. Silva, N.A. Borges, W.J. de Carvalho, S.L. Walter, Z.L.P. Simões, et al., Methyl farnesoate epoxidase (mfe) gene expression and juvenile hormone titers in the life cycle of a highly eusocial stingless bee, Melipona scutellaris, J. Insect Physiol. 101 (2017) 185–194, https://doi.org/10.1016/j.jinsphys.2017.08.001.

[2] 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. Haibocen, A. Gnrke, 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, https://doi.org/10.1038/nbt.1883.

[3] N.M. Davidson, A. Oshlack, Corset: enabling differential gene expression analysis for de novo assembled transcriptomes, Genome Biol. 15 (2014) 1–14, https://doi.org/10.1186/s13059-014-0410-6.