*Malaysian Cocoa Journal*
*Volume 13(1)/2021*

# TRANSCRIPTOME-WIDE ANALYSIS OF COCOA POD BORER (*CONOPORMORPHA CRAMERELLA*) REVEALED POTENTIAL TARGET GENES FOR CONTROL OF THE INSECT

**Tan C. L.[1], Rosmin K.[1], Leong W. M.[2]**

[1]Malaysian Cocoa Board, Wisma SEDCO, Kota Kinabalu, Sabah, Malaysia. [2]Neoscience Sdn. Bhd., Kelana Square, Kelana Jaya, Selangor, Malaysia.
*Corresponding author: tancl@koko.gov.my*

Malaysian Cocoa J. (2021) 13(1): 53-59

**ABSTRACT** - *The cocoa pod borer, Conopomorpha cramerella (Snellen) is a serious pest in cocoa plantations in Southeast Asia. It causes significant losses in the crop. Unfortunately, genetic resources for this insect are extremely scarce. To improve these resources, we sequenced the transcriptome of C. cramerella representing the three stages of development, larva, pupa and adult moth using Illumina NovaSeq6000. We have identified a number of genes that are involved in reproduction and development such as genes involved in general function processes in the insect. Genes found to be involved in reproduction such as porin, dsx, bol and fruitless were associated with sex determination, spermatogenesis and pheromone binding. This serves as a valuable genetic repository of the insect for potential RNA interference and genome editing.*

***Keywords**: Conopomorpha cramerella*, transcriptome, cocoa pod borer, RNA interference, genome editing

## INTRODUCTION

Cocoa pod borer (*Conopomorpha cramerella* Snellen) is a Lepidopteran moth of the family *Gracillariidae* (Posada & Vega, 2005). It is known to be of south Asian origin (Bradley 1986). It is found mainly in Thailand, Brunei, Indonesia (Sumatra, Sulawesi, Papua, Papua Barat, Java, Kalimantan, Moluccas), Malaysia, Vietnam, Australia, Philippines, Samoa, the Solomon Islands, Sri Lanka, Taiwan and Vanuatu. Economic losses due to this insect can be up to 80% in some geographical regions (Posada *et al.*, 2011). Control of this notorious pest is achieved mainly by chemical pesticides. However, overuse of pesticides leads to environmental and food safety issues. Therefore, alternate pest control strategies for CPB are highly desirable and need to be developed.

To develop RNAi-mediated pest control methods, it is critical to find suitable target genes. Target genes should not only have insecticidal effects on the target pests, but should also be safe to non-target organisms. Unfortunately, genetic resources for CPB insects are extremely scarce and therefore additional resources are required for effective screening of target genes. In this study, we present the results from the sequencing and assembly of the transcriptome of *Conopomorpha cramerella* Snellen at different developmental stages (larvae to pupa and adult) using Illumina NovaSeq6000 technology. Genes involved in metabolic processes, general development and reproduction were identified and functionally annotated. A great number of differentially expressed genes were obtained and some of these genes have been cloned using PCR for further downstream studies. The transcriptome study is undoubtedly valuable for molecular studies of the underlying mechanism on the development and reproduction of the insect. It also serves as a useful resource for target genes for RNA interference studies and the development of effective and environmental-friendly strategies for pest control.

## MATERIALS AND METHODS

### Insects

Cocoa pods that were infected with cocoa pod borer (CPB) were obtained from cocoa farms in Keningau, Sabah, Malaysia. They were wrapped in papers and kept in the dark for two weeks. During the period, they were constantly checked for CPB larvae, pupae and moth.

### RNA isolation and cDNA construction

Total RNA from CPB larvae, pupae and moths were extracted using the GeneAll Hybrid-R™ kit (GeneAll Biotechnology, Seoul, Korea) according to the manufacturer's instructions. RNA Integrity Number (RIN) was determined using RNA Nano

6000 Assay Kit (Agilent Technologies, CA, USA) with the Agilent 2100 Bioanalyzer (Agilent Technologies).

The libraries were prepared for 150bp paired-end sequencing using TruSeq stranded mRNA Sample Preparation Kit (Illumina, CA, USA). Namely, mRNA molecules were purified and fragmented from 1μg of total RNA using oligo (dT) magnetic beads. The fragmented mRNAs were synthesized as single-stranded cDNAs through random hexamer priming. By applying this as a template for second strand synthesis, double-stranded cDNA was prepared. After a sequential process of end repair, A-tailing and adapter ligation, cDNA libraries were amplified with PCR (Polymerase Chain Reaction). Quality of these cDNA libraries was evaluated with the Agilent 2100 BioAnalyzer (Agilent, CA, USA). They were quantified with the KAPA library quantification kit (Kapa Biosystems, MA, USA) according to the manufacturer's library quantification protocol. Following cluster amplification of denatured templates, sequencing was progressed as paired-end (2×150bp) using Illumina NovaSeq6000 (Illumina, CA, USA).

**Bioinformatics Analysis of RNA-seq data**

**Transcriptome assembly & Unigene discovery**

A. Filtering
Prior to the assembly, filtering was proceeded to remove low quality reads and adapter sequence according to the following criteria; reads contain more than 10% of skipped bases (marked as 'N's), reads contain more than 40% of bases whose quality scores are less than 20 and reads of which average quality scores of each read is less than 20. Furthermore, bases of both ends less than Q20 of filtered reads were removed additionally. This process is to enhance the quality of reads due to mRNA degradation in both ends of it as time goes on (Martin and Wang, 2011). The whole filtering process was performed using the in-house scripts.

B. Assembly
Transcriptome assembly was performed by Trinity (Grabherr *et al*., 2011; Hass *et al*., 2013) program using data from all samples. Trinity is a representative RNA assembler based on the de Bruijin graph (DBG) algorithm for RNA-seq de novo assembly, and its assembly pipeline consists of three consecutive modules: Inchworm, Chrysalis, and Butterfly. First, the Inchworm module is to construct contigs according to the following steps; each 100bp read divides into 4 fragments (each fragment is 25bp). When overlapping 24bp of each fragment, the 24 overlapped region is merged for construction of contigs. The module requires a single high-memory server so that classification into subgroups after the construction was progressed for efficient usage of memory. Next, Chrysalis clusters related Inchworm contigs into components. And, the DBG is generated in each cluster. Finally, Butterfly reconstructs transcript sequences in a manner that indicates the original cDNA molecules. All options were set to default values.

C. Clustering
According to the previous publication (Yang and Smith, 2013), there are some problems as to when to perform the assembly by Trinity. At first, the assembled transcripts contained the overlapping sequence of the same region. This is due to the transcripts originated from transcripts containing isoforms and not genes. In addition to that, chimera transcripts are generated through the assembly process. To overcome these problems, grouping the assembled transcripts by TGICL (Pertea *et al*., 2003), a pipeline for transcriptome analysis in which the sequences are clustered based on pairwise sequence similarity, was carried out for removal of the overlapping and the chimera sequences. Subsequently, extraction of the representative sequence was carried out using CAP3 (Huang and Madan, 1999): a sequence assembly program. The criterion of sequence similarity for grouping was set to 0.94 value.

D. CDS prediction
Protein coding sequence (CDS) was extracted from the reconstructed transcripts by TransDecoder: a utility included with Trinity to assist in the identification of potential coding regions (Haas *et al*., 2013). The coding region is predicted according to following procedures; 1) search all possible CDSs of the transcripts, 2) verify the predicted CDSs by GeneID (Blanco *et al*., 2007) through selecting it for more than 0 value of log-likelihood score, and 3) choose the region which has the highest score among candidate sequences.

**Functional annotation of Unigenes**

Blast and InterProScan were applied for homology search to make a prediction of the function of CDS in unigene.

A. Blastx with nucleotide sequence
NCBI Blast 2.2.29+ was applied for nucleotide sequence-based homology search. The function of CDS was predicted by Blastx to search all possible proteins matched with unigene sequence against the SwissProt db. The criterion regarding significance of the similarity was set to E-value < 1e-5.

B. InterProScan with protein sequence
InterProScan is another tool for homology search using protein sequence. The InterProScan is based on the Hidden Markov Model to predict the function

of CDS by similarity search using the protein domain: units of protein structure for function. The search was progressed by InterProScan v5 against ProDom, PfamA, Panther, SMART, SuperFamily and Gene3d databases based on E-value < 1e-5.

**Gene expression estimation**

Gene expression level was measured with RSEM (Li and Dewey, 2011). The RSEM is a tool to measure the expression for transcripts without any information on reference, and Bowtie is applied to the RSEM using directed graph model following reads alignment to the transcripts for the expression.

**Differential Expressed gene (DEG) analysis**

The TCC package was applied for DEG analysis through the interactive DEGES/DEseq method. This method is based on DESeq (Anders and Huber, 2010) using Negative-binomial distribution. Normalization was progressed three times to search meaningful DEGs between comparable samples (Kadota *et al*., 2012). The DEGs were identified based on the qvalue threshold less than 0.05.

## RESULTS AND DISCUSSION

**Generation and assembly of cocoa pod borer transcriptomes**

In order to obtain an overview of *Conopomorpha cramerella* gene expression profile, cDNA from three different developmental stages (larvae, pupae and adult moth) were prepared and sequenced on Illumina NovaSeq6000 machine. A total of 22,961,926,438 bp from 147,356,088 sequence reads with an average read length of 146 bp was obtained (Table 1).

These raw data were assembled into 285,882 contigs. The mean contig length is 374 bp with lengths ranging from 225 bp to 16,526 bp. The percentage and number of singletons for larvae were 2.88% (1,659,115), pupae: 2.55% (1,044,176) and moth: 2.69% (1,314,145). The GC percentage of the transcriptomes is 38%.

**Table 1 Summary of the *C. cramerella* transcriptome**

| | |
|---|---|
| Total base pair (bp) | 22,961,926,438 bp |
| Number of high-quality reads | 147,356,088 |
| Number of reads assembled in contigs | 147,356,088 |
| Average read length (bp): | 146 bp |
| Number of contigs | 285,882 |
| Average contig length (bp) | 374 bp |
| Range of contig length (bp): | 225~16,526bp |
| Number of singletons (based on mapped reads counts on the assembled unigene by using BWA software) | |
| o LARVA: | mapping 72%, singletons 2.88% (1,659,115) |
| o PUPAE | mapping 69.44%, singletons 2.55% (1,044,176) |
| o MOTH | mapping 70.54%, singletons 2.69% (1,314,145) |
| GC percentage | 38% |

**Gene ontology and cluster of orthologous groups classification**

Gene ontology (GO) assignment programs were utilised for functional categorisation of annotated genes. These sequences were categorised into 54 main functional groups belonging to 3 categories, including biological process, molecular function and cellular component. Among the biological processes (Figure 1A), the dominant GO terms were grouped into either metabolic processes (28%), biological regulation (18%) or cellular processes (16%). Within the molecular function category, there was a high percentage of genes with binding (45%) and catalytic activity (35%) (Figure 1B). For cellular components, those assignments were mostly given to the cell part (27%), organelle (21%), membrane part (14%) and membrane (12%) (Figure 1C). The three largest functional groups were binding, catalytic activity and metabolic process.

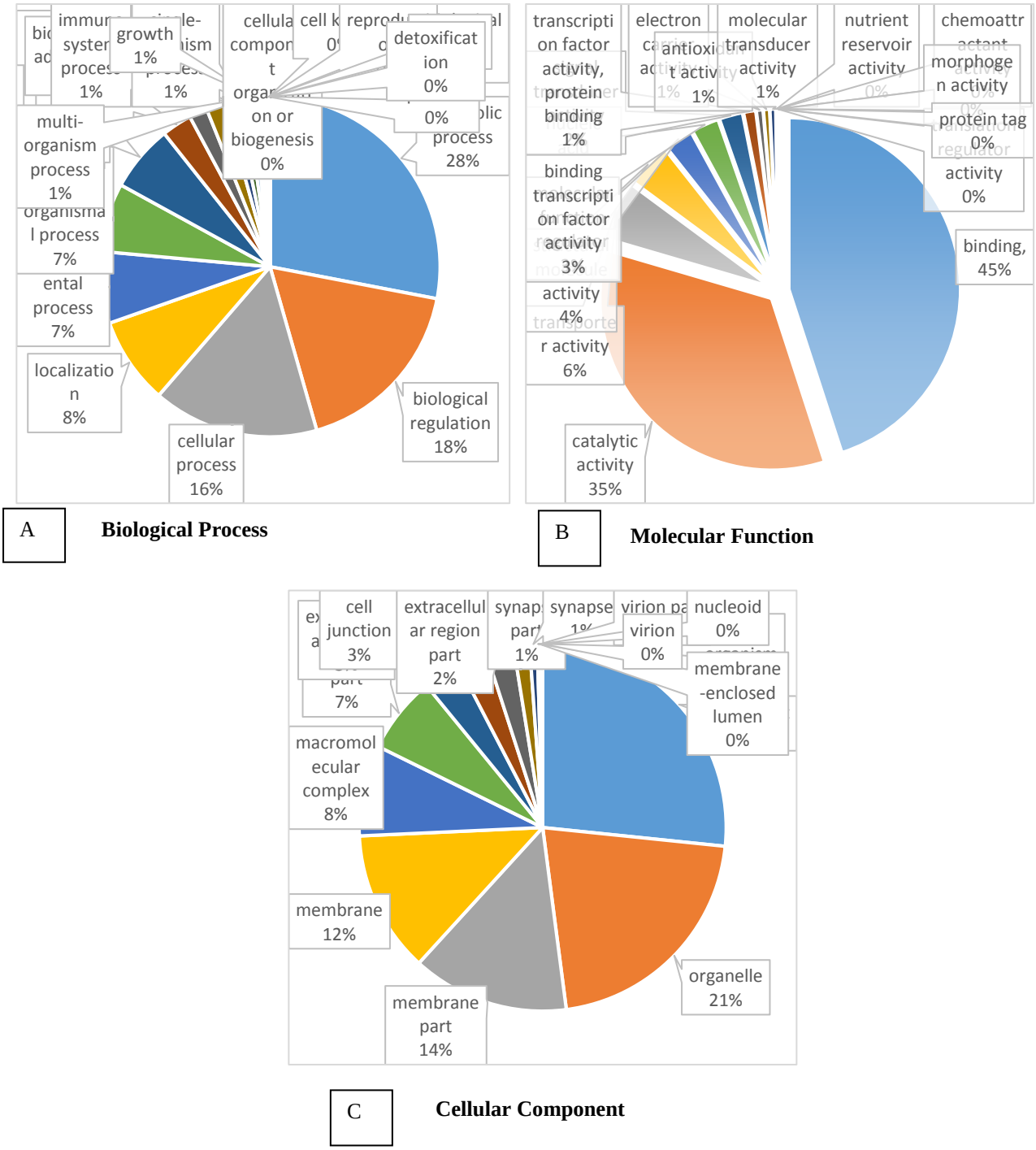


**Figure 1 GO analyses of *Conopomorpha cramerella* transcriptome data.**

To further evaluate the completeness of our transcriptomic library and the effectiveness of our annotation process, assignments of cluster of orthologous groups (COG) were used. Among the 25 COG categories, the majority of the cluster were "General function prediction only" (358, 10.95%), "Post Translational modification, protein turnover, chaperones" (344, 10.52%), "Translation, ribosomal structure and biogenesis" (306, 9.36%) and "Carbohydrate transport and metabolism" (296, 8.23%) whereas "RNA processing and modification" (1, 0.03%), "Chromatin structure and

dynamics" (14, 0.43%) and "Extracellular structures" (18, 0.55%) represented the smallest groups (Figure 5).

**Genes involved in general function**

Genes involved in general function were listed in Table 2. The results showed that "General function prediction only" constitutes the majority of the cluster within the metabolism pathway classification of the *C. cramerella* transcriptome. This includes choline dehydrogenase or related flavour protein, GTPase SAR1 family domain, NAD(P)-dependent dehydrogenase, short-chain alcohol dehydrogenase family, pimeloyl-ACP methyl ester carboxylesterase, short-chain dehydrogenase, tetratricopeptide (TPR) repeat and WD40 repeat (Table 2).

**Table 2 Genes involved in general function**

| COG annotation | No. of genes |
|---|---|
| Choline dehydrogenase or related flavoprotein | 21 |
| GTPase SAR1 family domain | 49 |
| NAD(P)-dependent dehydrogenase, short-chain alcohol dehydrogenase family | 60 |
| Pimeloyl-ACP methyl ester carboxylesterase | 17 |
| Short-chain dehydrogenase | 13 |
| Tetratricopeptide (TPR) repeat | 11 |
| WD40 repeat | 27 |

Among the genes involved in general function, NAD(P)-dependent dehydrogenase, short-chain alcohol dehydrogenase family has the largest number of genes. Alcohol dehydrogenase is considered a very important enzyme in insect metabolism because it is involved in the catalysis of the reversible conversion of various alcohols in larval feeding sites to their corresponding aldehydes and ketones, thus contributing to detoxification and metabolic purposes (Eliopoulos *et al*., 2004). In *Helicoverpa armigera*, alcohol dehydrogenase gene (HaADH5) regulates the expression of CYP6B6, a gene involved in molting and metamorphosis (Zhao *et al.*, 2019). The second largest group of genes in general function are GTPase. These genes are involved in metabolic pathways of insects (Lee *et* al., 2019). In *Drasophila*, GTPase is found to be involved in endocytosis and vesicle trafficking in the insect renal system (Fu *et al.*, 2017). GTPase is also known to regulate diverse cellular and developmental events, by regulating the exocytotic and transcytotic events inside the cell (Singh & Kumar, 2013). The third largest group of general function genes are WD40 repeat genes. WD40 proteins are scaffolding molecules in protein-protein interactions and play crucial roles in fundamental biological processes such as the metabolic activities of the insect (He *et al*., 2018; Orville Singh *et al.*, 2016).

**Genes involved in reproduction**

In insects, sexual reproduction is a very important physiological process and is critical to the maintenance of a population. Therefore, identification of genes involved in reproduction is important and would be helpful for pest control purposes. In addition, it will also be useful to evaluate molecular mechanisms for higher order insect's species.

Several reproductive-related genes have been identified (Table 3) in the transcriptome libraries. Among them is the *porin* gene, a male-biased pheromone binding protein, a short chain dehydrogenase/reductase, and a member of the *takeout* gene family (Jordan *et al.*, 2008). Another reproductive-related genes is the boule (*bol*) gene. This gene is a member of the *Deleted* in *Azoospermia* (DAZ) gene family and plays an important role in meiosis (reductional maturation divisions) in a spermatogenesis of insect male (Sekine *et al*., 2015). The gene, *dsx* is also found in the transcriptome analysis. This gene is involved in sex determination in insect (Wang *et al*., 2019; Taracena *et al*., 2019). Another sex-determination gene that is found in *C. cramerella* is the *fruitless* gene. In *Drosophila melanogaster*, the *fruitless* gene produces sex-specific gene products under the control of the sex-specific splicing cascade and contributes to the formation of the sexually dimorphic circuits (Watanabe, 2019, Hall *et al.*, 2015).

**Table 3 Cocoa pod borer assembled sequences with best-hit matches to insect genes involved in reproductive behaviors**

| Gene ID | Insect Gene | Length (bp) | E-value | Protein identity (%) | Function |
|---|---|---|---|---|---|
| TBIU002860 | *porin* | 273 | 3.00E-09 | 67.86 | pheromone binding protein |
| TBIU040283 | *bol* | 1456 | 3.00E-07 | 40.68 | spermatogenesis of insect male |
| TBIU000835 | *dsx* | 344 | 4.00E-09 | 29.91 | sex determination |
| TBIU000002 | *fruitless* | 663 | 1.00E-84 | 92.7 | sex determination |

## CONCLUSIONS

We have generated a comprehensive transcriptome of the *C. cramerella* development using Illumina NovaSeq6000 platform. A large number of genes involved in reproduction, general function and development pathways are found in the transcriptome. In addition, genes differentially expressed at different development stages were identified. These data make a substantial contribution to genetic resources of cocoa pod borer. It also provides potential molecular targets for the control of *C. cramerella* using RNAi. Finally, the study may also aid in the understanding of the molecular basis of development and reproduction in cocoa pod borer insects.

## ACKNOWLEDGEMENT

We would also like to thank the Director of Biotechnology for providing funding under the 11th Malaysian Development Fund to complete the project. Lastly, we like to thank the Director-General and Deputy Director-General (Research) for their support and permission to publish this paper.

## REFERENCES

Anders, S, and Huber W. (2010) Differential expression analysis for sequence count data, Genome Biol. 11(10):R106.

Blanco, E. *et al*. (2007) Using geneid to identify genes, Curr Protoc Bioinformatics. Jun;Chapter 4:Unit 4.3.

Bradley, JD. (1986). Identity of the South East Asian cocoa moth, *Conopomorpha cramerella* (Snellen) (Lepidoptera: Gracillariidae), with descriptions of three allied new species. Bulletin Entomological Res 76(1): 41–51.

Eliopoulos, E., Goulielmos, G. N., & Loukas, M. (2004). Functional constraints of alcohol dehydrogenase (ADH) of tephritidae and relationships with other Dipteran species. J Mol Evol, 58(5), 493-505. doi:10.1007/s00239-003-2568-5.

Fu, Y., Zhu, J. Y., Zhang, F., Richman, A., Zhao, Z., & Han, Z. (2017). Comprehensive functional analysis of Rab GTPases in Drosophila nephrocytes. *Cell Tissue Res, 368*(3), 615-627. doi:10.1007/s00441-017-2575-2.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., ... & Regev, A. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome, Nat Biotechnol. 15;29(7):644-52.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., ... & Regev, A. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis, Nat. Protoc. 8(8):1494-512.

Hall, A. B., Basu, S., Jiang, X., Qi, Y., Timoshevskiy, V. A., Biedler, J. K., Tu, Z. (2015). SEX DETERMINATION. A male-determining factor in the mosquito *Aedes aegypti*. Science, 348(6240), 1268-1270. doi:10.1126/science.aaa2850.

He, S., Tong, X., Han, M., Hu, H., & Dai, F. (2018). Genome-Wide Identification and Characterization of WD40 Protein Genes in the Silkworm, *Bombyx mori*. Int J Mol Sci, 19(2). doi:10.3390/ijms19020527

Huang X. and Madan A. (1999) CAP3: A DNA sequence assembly program, Genome Res. 9, 868-877.

Jordan, M. D., Stanley, D., Marshall, S. D., De Silva, D., Crowhurst, R. N., Gleave, A. P., Newcomb, R. D. (2008). Expressed sequence tags and proteomics of antennae from the

tortricid moth, *Epiphyas postvittana*. Insect Mol Biol, 17(4), 361-373. doi:10.1111/j.1365-2583.2008.00812.x

Kadota, K., Nishiyama, T., & Shimizu, K. (2012) A normalization strategy for comparing tag count data, Algorithms Mol Biol. 7(1):5.

Lee, S. J., Yang, Y. T., Kim, S., Lee, M. R., Kim, J. C., Park, S. E., Kim, J. S. (2019). Transcriptional response of bean bug (*Riptortus pedestris*) upon infection with entomopathogenic fungus, *Beauveria bassiana* JEF-007. Pest Manag Sci, 75(2), 333-345. doi:10.1002/ps.5117.

Li B. and Dewey C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome, BMC Bioinformatics, 4;12:323.

Martin J.A. and Wang Z. (2011) Next-generation transcriptome assembly, Nat Rev Genet. 12(10):671-82.

Orville Singh, C., Xin, H. H., Chen, R. T., Wang, M. X., Liang, S., Lu, Y., . . . Miao, Y. G. (2016). BmPLA2 containing conserved domain WD40 affects the metabolic functions of fat body tissue in silkworm, *Bombyx mori*. Insect Sci, 23(1), 28-36. doi:10.1111/1744-7917.12189

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., & Quackenbush, J. (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets, Bioinformatics, 19(5):651-2.

Posada, F. J., Virdiana, I., Navies, M., Pava-Ripoll, M., & Hebbar, P. (2011). Sexual dimorphism of pupae and adults of the cocoa pod borer, *Conopomorpha cramerella*. J Insect Sci, 11, 52. doi:10.1673/031.011.5201

Posada, F., & Vega, F. E. (2005). Establishment of the fungal entomopathogen *Beauveria bassiana* (Ascomycota: Hypocreales) as an endophyte in cocoa seedlings (*Theobroma cacao*). *Mycologia, 97(6), 1195–1200.* doi:10.1080/15572536.2006.11832729.

Sekine, K., Furusawa, T., & Hatakeyama, M. (2015). The boule gene is essential for spermatogenesis of haploid insect male. Dev Biol, 399(1), 154-163. doi:10.1016/j.ydbio.2014.12.027

Singh, D., & Kumar Roy, J. (2013). Rab11 plays an indispensable role in the differentiation and development of the indirect flight muscles in Drosophila. *PLoS One*, *8*(9), e73305. doi:10.1371/journal.pone.0073305.

Taracena, M. L., Hunt, C. M., Benedict, M. Q., Pennington, P. M., & Dotson, E. M. (2019). Downregulation of female doublesex expression by oral-mediated RNA interference reduces number and fitness of *Anopheles gambiae* adult females. Parasit Vectors, 12(1), 170. doi:10.1186/s13071-019-3437-4

Wang, Y., Zhao, Q., Wan, Q. X., Wang, K. X., & Zha, X. F. (2019). P-element Somatic Inhibitor Protein Binding a Target Sequence in dsx Pre-mRNA Conserved in Bombyx mori and *Spodoptera litura*. Int J Mol Sci, 20(9). doi:10.3390/ijms20092361.

Watanabe, T. (2019). Evolution of the neural sex-determination system in insects: does *fruitless* homologue regulate neural sexual dimorphism in basal insects? Insect Mol Biol, 28(6), 807-827. doi:10.1111/imb.12590

Yang Y. and Smith S.A. (2013) Optimizing de novo assembly of short-read RNA-seq data for phylogenomics, BMC Genomics. 14:328.

Zhao, J., Wei, Q., Gu, X. R., Ren, S. W., & Liu, X. N. (2019). Alcohol dehydrogenase 5 of *Helicoverpa armigera* interacts with the CYP6B6 promoter in response to 2-tridecanone. *Insect Sci*. doi:10.1111/1744-7917.12720