Transcriptome analysis of developmental stages of cocoa pod borer, *Conopomorpha cramerella*: A polyphagous insect pest of economic importance in Southeast Asia.

(Short title: Transcriptome analysis of *Conopomorpha cramerella* development)

Chia Lock Tan ^{1*}, Rosmin Kasran¹, Wei Wei Lee², Wai Mun Leong²

¹Malaysian Cocoa Board, Wisma SEDCO, Kota Kinabalu, Sabah, Malaysia. ²Neoscience Sdn. Bhd., Kelana Square, Kelana Jaya, Selangor, Malaysia.

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 is 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. Transcriptome assembly was performed by Trinity for all the samples. A total number of 147,356,088 high quality reads were obtained. Of these, 285,882 contigs were assembled. The mean contig size was 374 bp. Protein coding sequence (CDS) was extracted from the reconstructed transcripts by TransDecoder. Subsequently, BlastX and InterProScan were applied for homology search to make a prediction of the function of CDS in unigene. Additionally, we 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*

*Corresponding author. Tel.: +60 88489101; Orcid ID: <https://orcid.org/0000-0002-5071-8788>
Email address : tancel@koko.gov.my

and *fruitless* were associated with sex determination, spermatogenesis and pheromone binding. Furthermore, transcriptome changes during development were analysed. There were 2,843 differentially expressed genes (DEG) detected between the larva and pupa samples. A total of 2,861 DEG were detected between adult and larva stage whereas between adult and pupa stage, 1,953 DEG were found. In conclusion, the transcriptomes could be a valuable genetic resource for identification of genes in *C. cramerella* and the study will provide putative targets for RNAi pest control.

Keywords: *Conopomorpha cramerella*, transcriptome analysis, insect developmental stages, cocoa pod borer, RNA interference

Introduction

Cocoa pod borer (*Conopomorpha cramerella* Snellen) is a Lepidopteran moth of the family *Gracillariidae* [1]. It is known to be of south Asian origin [2]. It is found mainly in Thailand, Brunei, Indonesia (Sumatra, Sulawesi, Papua New Guinea, Java, Kalimantan, Moluccas), Malaysia, Vietnam, Australia, Philippines, Samoa, the Solomon Islands, Sri Lanka, Taiwan and Vanuatu. Its primary hosts are plants native to the area such as Rambutan (*Nephelium lappaceum*); Pulasan (*Nephelium mutabile*); Kasai (*Pometia pinnata*); Cola (*Cola nitida*, *C. acuminata*); and Nam-nam, (*Cynometra cauliflora*). With the introduction of cocoa (*Theobroma cacao* L.) to this geographic region, cocoa pod borer (CPB) moved onto this crop and exploited *T. cacao* as its new host. Since 1986, CPB has become the most serious insect pest of cocoa in Southeast Asia (Indonesia, Philippines, Malaysia, and Papua New Guinea). Economic losses due to this insect can be up to 80% in some geographical regions [56]. 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 is highly desirable and need to be developed.

Double-stranded RNA (dsRNA)-mediated gene silencing, commonly referred to as RNA interference (RNAi), is becoming a widely used functional genomics tool in insects to ascertain the function of the many newly identified genes accumulating from genome sequencing projects [3, 4]. The basic components of the RNAi process, namely the endonuclease Dicer, which first chops long dsRNAs into short interfering RNAs (siRNAs), and the RNA-induced silencing complex (RISC), which facilitates the targeting and endonucleolytic attack on mRNAs with sequence identity to the dsRNA, are evolutionarily conserved across virtually all eukaryotic taxa [5], and consequently, RNAi could be readily applied to any insect species. This RNAi technique has been successfully applied to study gene functions in many insects, including *Drosophila melanogaster* [6], *Tribolium castaneum* [7], *Helicoverpa armigera* [8], *Gryllus bimaculatus* [9], *Schistocerca gregaria* [10], *Plutella xylostella* [11], *Nilaparvata lugens* [12], and *Epiphyas postvittana* [13]. There are two kinds of RNA delivery methods, oral intake or injection. Injection of siRNA or dsRNA is widely used in the laboratory at a small scale level, whereas oral intake is more feasible to be used for controlling pest in the field condition.

RNAi-mediated pest control is a novel and promising technique because interference with important insect genes using RNAi can lead to death of pests. Proof of principle for the application of RNAi in insect crop pest control comes from early studies conducted on the western corn rootworm (WCRW - *Diabrotica virgifera*), and cotton bollworm (CBW - *Helicoverpa armigera*) [14]. The researchers fed larval WCRW on 290 dsRNAs, from which they identified 14 genes that reduced larval performance, and one of these, vacuolar ATPase

subunit A (V-ATPase), was carried forward for detailed analysis. Low concentrations of orally-delivered dsRNA against V-ATPase in artificial diet suppressed the corresponding WCRW mRNA. Importantly, larvae reared on transformed corn plants that express V-ATPase dsRNA also displayed reduced expression of the V-ATPase gene and caused much reduced plant root damage. In the study on CBW, the target gene was a cytochrome P450, CYP6AE14, which is expressed in the larval midgut and detoxifies gossypol, a secondary metabolite common to cotton plants. When CBW was exposed to either *Arabidopsis thaliana* or *Nicotiana tabacum* expressing CYP6AE14 dsRNA, levels of this transcript in the insect midgut decreased, larval growth was retarded, and both effects were more dramatic in the presence of gossypol. Transgenic cotton plants expressing CYP6AE14 dsRNA also support drastically retarded growth of the CBW larvae, and suffered less CBW damage than control plants [15]. In another study, researchers used hairpin RNA expressed in both *Escherichia coli* and transgenic tobacco plants to decrease mRNA and protein levels of the *H. armigera*-derived molt-regulating transcription factor in larval *H. armigera*, which resulted in developmental deformity and larval lethality [16]. Another example is provided by nicotine, a neurotoxin made by species of tobacco. The tobacco hornworm *Manduca sexta* (Lepidoptera) can tolerate high nicotine concentrations. Larvae even exhale nicotine through their spiracles, deterring spider predation. Dietary nicotine induces the cytochrome P450 gene CYP6B46 in *M. sexta*. Tobacco plant transformed with a construct expressing dsRNA targeting 300 nt of the *M. sexta* gene for CYP6B46. Tobacco hornworm larvae consuming the transformed tobacco were more susceptible to spider predation because they exhaled less nicotine [17]. The success of these studies attests to the functionality of the RNAi in controlling insect pests.

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 insect is extremely scarce and therefore additional resources are required for effective screening of target genes. Insect transcriptomes has been reported to be useful genetic resources for high-throughput screening of RNAi target genes [18]. The introduction of next-generation sequencing technologies has provided significant convenience for further studies of non-model organisms including insects [19, 20]. Next generation sequencing such as Illumina and PacBio have been widely used to identify genes involved in several developmental and physiological processes. These technologies has been used to identify candidate chemosensory genes of oligophagous insect, *Ophraella communa* (Coleoptera: Chrysomelidae). These genes plays a key role in insect survival, which mediates important behaviors like host search, mate choice, and oviposition site selection [21]. Using NextSeq500 (Illumina) sequencing, Singh *et al.* [22] studied *de novo* transcriptome assembly and analysis of RNAi in *Phenacoccus solenopsis* Tinsley (Hemiptera: Pseudococcidae), one of the major polyphagous crop pests in India. The study provides a base for future research on developing RNAi as a strategy for management of this pest. Gao *et al.* [23] used PacBio to profile full-length transcriptomes of insect *Erthesina fullo* Thunberg mitochondrial gene expression. However, even though CPB is an important pest to cocoa in South-east Asia, there is no published report on the genome or transcriptome of the insect. To the best of our knowledge, this is the first report on transcriptomic analysis of *C. cramerella*, covering the three developmental stages of the life cycle of the insect.

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 serve 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 farm 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. Approximately thirty of each larvae, pupae and moth were collected and kept in RNA Later[®] and maintained in -70°C freezer until later use. The samples were grind to fine powder in liquid nitrogen before RNA isolation.

RNA isolation and cDNA construction

Total RNA from CPB larvae, pupae and moth 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 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

[24]. The whole filtering process was performed using the in-house scripts.

B. Assembly

Transcriptome assembly was performed by Trinity [25, 26] program using data from all samples. Trinity is a representative RNA assembler based on the de Bruijn graph (DBG) algorithm for RNA-seq de novo assembly, and its assembly pipeline consists of three consecutive modules: Inchworm, Chrysalis, and Butterfly. First, Inchworm module is to construct contigs according to the following steps; each 100bp read divides into 4 fragments (each fragment is 25bp). When to overlap 24bp of the 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 [27], there are some problems as to when to perform the assembly by Trinity. At first, the assembled transcripts contained the overlapping sequence of 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 [28], 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 [29]: 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 [26]. The coding region is predicted according to following procedures; 1) search all possible CDSs of the transcripts, 2) verify the predicted CDSs by GeneID [30] 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 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

220 InterProScan v5 against ProDom, PfamA, Panther, SMART, SuperFamily and Gene3d
221 databases based on E-value < 1e-5.

222

223 **Gene expression estimation**

224

225 Gene expression level was measured with RSEM [31]. The RSEM is a tool to measure the
226 expression for transcripts without any information on reference, and Bowtie is applied to the
227 RSEM using directed graph model following reads alignment to the transcripts for the
228 expression.

229

230 **Differential Expressed gene (DEG) analysis**

231 TCC package was applied for DEG analysis through the interactive DEGES/DEseq method.
232 This method is based on DESeq [32] using Negative-binomial distribution. Normalization
233 was progressed three times to search meaningful DEGs between comparable samples [33].
234 The DEGs were identified based on the qvalue threshold less than 0.05.

235

236 **Data availability**

237 The datasets generated and analysed during the current study are available at NCBI Gene
238 Expression Omnibus (GEO) Accession Series GSE146610.

239

240 **qRT-PCR validation**

241 To verify the differential expression detected by Illumina RNA-Seq, qRT-PCR was
242 performed on the same samples that had been used previously. A set of seventeen genes was
243 chosen at random, the expression of each gene was evaluated for two life stages of cocoa pod

borer insect and compared with their observed FPKM. qRT-PCR was performed using Rotor-Gene™ 6000 Real-Time thermocycler (Corbett Research, Australia) with Brilliant SYBR® Green QPCR Master Mix (Stratagene, La Jolla, CA) following the manufacturer's instructions. The forward and reverse primers used for qRT-PCR are listed in Supplementary Table S6. The thermal cycling conditions were as follows: 95°C for 10 min, followed by 45 cycles of 95°C for 15 s and 55°C for 60°C. Gene expression was normalised with actin gene using primer pairs qActin-F and qActin-R (Supplementary Table S6). The data are presented as mean $\pm$ SE of three independently produced RT preparations used for PCR runs, each having at least three replicates. The relative expression levels were calculated using the delta-delta Ct method [34].

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%.

268 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%

269

270 Annotation of predicted sequences

271 To analyse which part of the assembled sequences had counterparts with other insect
272 species, orthologous genes shared between *C. cramerella* and other three insect species were
273 compared. These insect species chosen for comparison were Dipteran *Drosophila*
274 *melogaster*, Lepidopteran *Bombyx mori* and Lepidopteran *Helicoverpa armigera*. The results
275 showed a total number of 16,595 hits (Figure 2). There were 7,523 identifiable genes shared

between *D. melanogaster* and *C. cramerella*, indicating a good coverage of *C. cramerella* transcriptomes. Homologous genes shared between *Bombyx mori* and *C. cramerella* were 6,047 and between *Helicoverpa armigera* and *C. cramerella* were 5,036. There were more genes that *C. cramerella* shared with both *Bombyx mori* and *Helicoverpa armigera* (1,845) than *C. cramerella* with all the three insects combined (73). This is unsurprising as *C. cramerella*, *Bombyx mori* and *Helicoverpa armigera* are Lepidopteran insects whereas *Drosophila melanogaster* belongs to Diptera.

The identity distribution of *C. cramerella* transcriptomes were then analysed (Figure 3). Out of a total of 67,770 (41%) hits that has homology, 80.32% (54,434) were of plant origin. The second largest group were invertebrates, which include insects (11.16%, 7,565). The other groups like bacteria, primates, virus and vertebrates were less than 5% of homology. The high homology with plant genes could be due to the fact that *C. cramerella* is a phytophagous insect [35].

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 4A), the dominant GO terms were grouped into either metabolic process (28%), biological regulation (18%) or cellular process (16%) (Figure 3). Within the molecular function category, there was a high percentage of genes with binding (45%) and catalytic activity (35%) (Figure 4B). For cellular components, those assignments were mostly given to cell part (27%), organelle (21%), membrane part (14%) and membrane (12%) (Figure 4C). The three largest functional groups were binding, catalytic activity and metabolic process.

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. Overall, 3,269 were classified as involved in different metabolic process ([Fig. 1](#)). Among the 25 COG categories, the majority of the cluster were “General function prediction only” (358, 10.95%), “Posttranslational 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 ([Fig. 5](#)). 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 [36]. In *Helicoverpa armigera*, alcohol dehydrogenase gene (HaADH5) regulates the expression of CYP6B6, a gene involved in molting and metamorphosis [37]. The second largest group of genes in general function are GTPase. These genes are involved in metabolic pathways of insect [38]. In *Drasophila*, GTPase is found to be involved in endocytosis and vesicle trafficking in the insect renal system [39]. GTPase is also known to regulate diverse cellular and developmental events, by regulating the exocytotic and transcytotic events inside the cell [40]. 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 [41, 42].

Genes expression profile among the different developmental stages

To identify genes showing differential expression during development, the differentially expressed sequences between two samples were identified (Fig. 6). There were 2,843 differentially expressed genes detected between the larva and pupa samples, including 1,979 down-regulated genes ($P < 0.05$) and 864 up-regulated genes ($P < 0.05$). The large number of differentially expressed genes between these two samples may be attributed to the important molting and metamorphosis processes during transition from larva to pupa. A cascade of physiological processes occurs during molting and complicated physiological processes takes place during metamorphosis including histolysis of larval tissues, remodelling and formation of adult tissues, and a molting cascade similar to the larva molt [43]. In addition, a total of 2,861 differentially expressed genes were detected between adult and larva stage, with 1,646 down-regulated genes and 1,215 up-regulated genes (Figure 6). Between the adult and the pupa stage, 897 genes were down-regulated whereas 1,056 genes were up-regulated from a total of 1,953 differentially expressed genes (Fig. 6).

In larva, there is a total of 140,427 expressed genes (> 1.0 fpkm), of which 14,023 were known genes and 126,404 novel genes (Table S2). In pupae, a total of 124,368 expressed genes with 13,417 known genes and 110,951 novel genes. In adult moth, the total of expressed genes were 129,652, of which 13,536 were known genes whereas 116,116 were novel genes. The sheer number of novel genes as compared to known genes goes to show that there are many genes in *C. cramerella* that was yet to be discovered.

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 mechanism 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 [[44](#)]. 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 [[45](#)]. The gene, *dsx* is also found in the transcriptome analysis. This gene is involved in sex determination in insect [[46](#); [47](#)]. 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 [[48](#), [49](#)].

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	<i>porin</i>	273	3.00E-09	67.86	pheromone binding protein
TBIU040283	<i>bol</i>	1456	3.00E-07	40.68	spermatogenesis of insect male
TBIU000835	<i>dsx</i>	344	4.00E-09	29.91	sex determination
TBIU000002	<i>fruitless</i>	663	1.00E-84	92.7	sex determination

Verification of differentially expressed genes

In order to evaluate our DEG library, the expression level of seventeen genes involved in development were analysed by qRT-PCR. Results showed that real-time PCR revealed the same expression trend as in the DEG data, albeit with some quantitative differences in expression level ([Table 4](#), [Fig. 7](#)). The genes *atr* and *me31B* were highly expressed in the pupa stage. These genes are involved in cross-over patterning effect and transitioning [[50](#), [51](#)] in *Drosophila*, probably plays a crucial role in growth development from larva to pupa. Src64B protein are actively involved in modulating actin level in cell development [[52](#), [53](#)] are highly expressed in the larva. Setdb1 is involved in histone modifications and genome regulation [[54](#)] and is expressed higher in the moth compared to the larva stage ([Fig. 7](#)). The *pol* gene is almost entirely expressed in the larva and not in the moth. It has function in RNA

synthesis and as a growth effector of Ras/ERK signalling in *Drosophila* [55]. As control, actin is used as it is demonstrated to be almost equally expressed in all the three developmental stages (Fig. 7). These data will provide us with molecular targets to further study on the development of *Conopomorpha cramerella*.

Table 4 Comparisons of DEGs data and qRT-PCR results

Gene	Unigene ID	DEGs library	Fold by DEG	Fold by qPCR
atr	TBIU052540	larva vs. pupa	-10.7	-4.7
me31B	TBIU052640	larva vs. pupa	-5.46	-0.27
Src64B	TBIU052493	larva vs. pupa	7.66	5.47
Mtnd1	TBIU052872	larva vs. pupa	3.27	4.49
Mitd1	TBIU053114	larva vs. moth	-2.0	-3.57
Setdb1	TBIU053165	larva vs. moth	-5.59	-2.79
JMJD4	TBIU053571	larva vs. moth	-26.4	-4.15
let-268	TBIU052928	larva vs. moth	2.74	0.98
pol	TBIU052928	larva vs. moth	4.64	3.16

X- element\ORF2	TBIU053569	larva vs. moth	4.72	1.16
Hgsnat	TBIU053727	larva vs. moth	2.45	6.09
exba	TBIU053780	larva vs. moth	17.28	1.54
Slc2a13	TBIU053815	pupa vs. moth	-2.77	-8.42
SDCBP	TBIU045865	pupa vs. moth	1.35	9.06
Rpl12	TBIU055268	pupa vs. moth	4.95	3.20
Bap60	TBIU056833	pupa vs. moth	4.87	1.06
Prm	TBIU057029	pupa vs. moth	4.13	5.63

411

412

413 **Conclusions**

414 We have generated a comprehensive transcriptome of the *C. cramerella* development using
 415 Illumina NovaSeq6000 platform. The single run produced 285,882 contigs with a mean
 416 length of 374 bp. A large number of genes involved in reproduction, general function and
 417 development pathways are found in the transcriptome. In addition, genes differentially

expressed at different development stages were identified. To our knowledge, this is the first report of transcriptome sequencing in *C. cramerella*, a lepidopteran insect pest lacking a reference genome. These data make a substantial contribution to genetic resources of cocoa pod borer. It also provide 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 insect.

Acknowledgement

We would like to thank the Director-General of the Malaysia Cocoa Board for permission to publish this paper. We also like to thank the Director of Biotechnology for allowing funding from the 11th Malaysia Development Fund for this project. Lastly, we also thank Neoscience Sdn. Bhd., Malaysia and Theragen, South Korea for the sequencing work.

Author Contributions

Conceived and designed the experiment: CLT, RK, WWL, WML. Performed the experiment: CLT, WML. Analysed the data: CLT, WWL, WML. Wrote the paper: CLT, WML.

References

1. 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](https://doi.org/10.1080/15572536.2006.11832729). PMID: [16722213](https://pubmed.ncbi.nlm.nih.gov/16722213/)
2. 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. doi: [10.1017/S000748530001525X](https://doi.org/10.1017/S000748530001525X)
3. Garcia RA, Macedo LLP, Nascimento DC, Gillet FX, Moreira-Pinto CE, Faheem M, Basso AMM, Silva MCM, Grossi-de-Sa MF (2017) Nucleases as a barrier to gene silencing in the cotton boll weevil, *Anthonomus grandis*. *PLOS One* doi: [10.1371/journal.pone.0189600](https://doi.org/10.1371/journal.pone.0189600), pp1-22. PMID: [29261729](https://pubmed.ncbi.nlm.nih.gov/29261729/)
4. Nandety RS, Kuo YW, Nouri S, Falk BW (2015) Emerging strategies for RNA interference (RNAi) applications in insects. *Bioengineered*. 6(1):8-19. doi: [10.4161](https://doi.org/10.4161). PMID: [25424593](https://pubmed.ncbi.nlm.nih.gov/25424593/)
5. Lim ZX, Robinson KE, Jain RG, Chandra GS, Asokan R, Asgari S, Mitter N (2016) Diet-delivered RNAi in *Helicoverpa armigera* – Progresses and challenges. *Journal of Insect Physiology* 85: 86–93. doi: [10.1016/j.jinsphys.2015.11.005](https://doi.org/10.1016/j.jinsphys.2015.11.005). PMID: [26549127](https://pubmed.ncbi.nlm.nih.gov/26549127/)
6. Liao, J. F., Wu, C. P., Tang, C. K., Tsai, C. W., Rouhova, L., & Wu, Y. L. (2019). Identification of regulatory host genes involved in sigma virus replication using RNAi knockdown in *Drosophila*. *Insects*, 10(10). doi: [10.3390/insects10100339](https://doi.org/10.3390/insects10100339). PMID: [31614679](https://pubmed.ncbi.nlm.nih.gov/31614679/)
7. Bi, J., Feng, F., Li, J., Mao, J., Ning, M., Song, X., . Li, B. (2019). A C-type lectin with a single carbohydrate-recognition domain involved in the innate immune response of

- 462 *Tribolium castaneum*. Insect Mol Biol, 28(5), 649-661. doi: [10.1111/imb.12582](https://doi.org/10.1111/imb.12582). PMID:
463 [30843264](https://pubmed.ncbi.nlm.nih.gov/30843264/)
- 464 8. Israni, B., & Rajam, M. V. (2017). Silencing of ecdysone receptor, insect intestinal mucin
465 and sericotropin genes by bacterially produced double-stranded RNA affects larval growth
466 and development in *Plutella xylostella* and *Helicoverpa armigera*. Insect Mol Biol, 26(2),
467 164-180. doi: [10.1111/imb.12277](https://doi.org/10.1111/imb.12277). PMID: [27883266](https://pubmed.ncbi.nlm.nih.gov/27883266/)
- 468 9. Ishimaru, Y., Bando, T., Ohuchi, H., Noji, S., & Mito, T. (2018). Bone morphogenetic
469 protein signaling in distal patterning and intercalation during leg regeneration of the
470 cricket, *Gryllus bimaculatus*. Dev Growth Differ, 60(6), 377-386. doi: [10.1111/dgd.12560](https://doi.org/10.1111/dgd.12560).
471 PMID: [30043459](https://pubmed.ncbi.nlm.nih.gov/30043459/)
- 472 10. Boerjan, B., Tobback, J., Vandersmissen, H. P., Huybrechts, R., & Schoofs, L. (2012).
473 Fruitless RNAi knockdown in the desert locust, *Schistocerca gregaria*, influences male
474 fertility. J Insect Physiol, 58(2), 265-269. doi: [10.1016/j.jinsphys.2011.11.017](https://doi.org/10.1016/j.jinsphys.2011.11.017). PMID:
475 [22138053](https://pubmed.ncbi.nlm.nih.gov/22138053/)
- 476 11. Peng, L., Wang, L., Zou, M. M., Vasseur, L., Chu, L. N., Qin, Y. D., . . . You, M. S.
477 (2019). Identification of Halloween Genes and RNA Interference-Mediated Functional
478 Characterization of a Halloween Gene shadow in *Plutella xylostella*. Front Physiol, 10,
479 1120. doi: [10.3389/fphys.2019.01120](https://doi.org/10.3389/fphys.2019.01120). PMID: [31555150](https://pubmed.ncbi.nlm.nih.gov/31555150/)
- 480 12. Zeng, J. M., Ye, W. F., Noman, A., Machado, R. A. R., & Lou, Y. G. (2019). The
481 Desaturase Gene Family is crucially required for Fatty Acid Metabolism and Survival of
482 the Brown Planthopper, *Nilaparvata lugens*. Int J Mol Sci, 20(6). doi:
483 [10.3390/ijms20061369](https://doi.org/10.3390/ijms20061369). PMID: [30893760](https://pubmed.ncbi.nlm.nih.gov/30893760/)
- 484 13. Turner, C. T., Davy, M. W., MacDiarmid, R. M., Plummer, K. M., Birch, N. P., &
485 Newcomb, R. D. (2006). RNA interference in the light brown apple moth, *Epiphyas*

- postvittana* (Walker) induced by double-stranded RNA feeding. *Insect Mol Biol*, 15(3), 383-391. doi: [10.1111/j.1365-2583.2006.00656.x](https://doi.org/10.1111/j.1365-2583.2006.00656.x). PMID: [16756557](https://pubmed.ncbi.nlm.nih.gov/16756557/)
14. Scott JG, Michel K, Bartholomay LC, Siegfried BD, Hunter WB, Smagghe G, Zhu KY, Douglas AE (2013) Towards the elements of successful insect RNAi. *Journal of Insect Physiology* 59: 1212–1221. doi: [10.1016/j.jinsphys.2013.08.014](https://doi.org/10.1016/j.jinsphys.2013.08.014). PMID: [24041495](https://pubmed.ncbi.nlm.nih.gov/24041495/)
15. Zhang J, Khan SA, Heckel DG, Bock R (2017) Next-Generation Insect-Resistant Plants: RNAi-Mediated Crop Protection. *Trends in Biotechnology*, Vol. 35, No. 9:871-882. doi: [10.1016/j.tibtech.2017.04.009](https://doi.org/10.1016/j.tibtech.2017.04.009) PMID: [28822479](https://pubmed.ncbi.nlm.nih.gov/28822479/)
16. Kim YH, Soumaila Issa M, Cooper AM, Zhu KY (2015) RNA interference: Applications and advances in insect toxicology and insect pest management. *Pestic Biochem Physiol*. 120:109-17. doi: [10.1016](https://doi.org/10.1016). PMID: [25987228](https://pubmed.ncbi.nlm.nih.gov/25987228/)
17. Kumar P, Pandit SS, Steppuhn A, Baldwin IT. (2014) Natural history-driven, plant-mediated RNAi-based study reveals CYP6B46's role in a nicotine-mediated antipredator herbivore defense. *Proc. Natl. Acad. Sci. U.S.A.* 111: 1245–1252. doi: [10.1073/pnas.1314848111](https://doi.org/10.1073/pnas.1314848111). PMID: [24379363](https://pubmed.ncbi.nlm.nih.gov/24379363/)
18. Wang Y., Zhang H., Li H., Miao X. (2011) Second-generation sequencing supply an effective way to screen RNAi targets in large scale for potential application in insect pest control. *PloS One* 6:e16844. doi: [10.1371/journal.pone.0018644](https://doi.org/10.1371/journal.pone.0018644). PMID: [21494551](https://pubmed.ncbi.nlm.nih.gov/21494551/)
19. Schuster S.C. (2008) Next-generation sequencing transform today's biology. *Nat. Methods* 5:16-18. doi: [10.1038/nmeth1156](https://doi.org/10.1038/nmeth1156). PMID: [18165802](https://pubmed.ncbi.nlm.nih.gov/18165802/)
20. Ansorge W.J. (2009) Next-generation DNA sequencing techniques. *N. Biotechnol.* 25:195-203. doi: [10.1016/j.nbt.2008.12.009](https://doi.org/10.1016/j.nbt.2008.12.009). PMID: [19429539](https://pubmed.ncbi.nlm.nih.gov/19429539/)
21. Ma, C., Zhao, C., Cui, S., Zhang, Y., Chen, G., Chen, H., . . . Zhou, Z. (2019). Identification of candidate chemosensory genes of *Ophraella communa* LeSage

- (Coleoptera: Chrysomelidae) based on antennal transcriptome analysis. Sci Rep, 9(1), 15551. doi: [10.1038/s41598-019-52149-x](https://doi.org/10.1038/s41598-019-52149-x). PMID: [31664149](https://pubmed.ncbi.nlm.nih.gov/31664149/)
22. Singh, S., Gupta, M., Pandher, S., Kaur, G., Goel, N., & Rathore, P. (2019). Using de novo transcriptome assembly and analysis to study RNAi in *Phenacoccus solenopsis* Tinsley (Hemiptera: Pseudococcidae). Sci Rep, 9(1), 13710. doi: [10.1038/s41598-019-49997-y](https://doi.org/10.1038/s41598-019-49997-y). PMID: [31548628](https://pubmed.ncbi.nlm.nih.gov/31548628/)
23. Gao, S., Ren, Y., Sun, Y., Wu, Z., Ruan, J., He, B., . . . Bu, W. (2016). PacBio full-length transcriptome profiling of insect mitochondrial gene expression. RNA Biol, 13(9), 820-825. doi: [10.1080/15476286.2016.1197481](https://doi.org/10.1080/15476286.2016.1197481)
24. Martin J.A. and Wang Z. (2011) Next-generation transcriptome assembly. Nat Rev Genet. 12(10):671-82. doi: [10.1038/nrg3068](https://doi.org/10.1038/nrg3068). PMID: [27310614](https://pubmed.ncbi.nlm.nih.gov/27310614/)
25. Grabherr M.G. et al. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome, Nat Biotechnol. 15;29(7):644-52. doi: [10.1038/nbt.1883](https://doi.org/10.1038/nbt.1883)
26. Haas B.J. et al. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis Nat. Protoc. 8(8):1494-512. doi: [10.1038/nbt.1883](https://doi.org/10.1038/nbt.1883). PMID: [21572440](https://pubmed.ncbi.nlm.nih.gov/21572440/)
27. Yang Y. and Smith S.A. (2013) Optimizing de novo assembly of short-read RNA-seq data for phylogenomics, BMC Genomics. 14:328. doi: [10.1186/1471-2164-14-328](https://doi.org/10.1186/1471-2164-14-328). PMID: [23672450](https://pubmed.ncbi.nlm.nih.gov/23672450/)
28. Pertea G., Huang X., Liang F., Antonescu V., Sultana R., Karamycheva S., Lee Y., White J., Cheung F., Parvizi B., Tsai J., Quackenbush J.. (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets, Bioinformatics, 19(5):651-2. doi: [10.1093/bioinformatics/btg034](https://doi.org/10.1093/bioinformatics/btg034). PMID: [12651724](https://pubmed.ncbi.nlm.nih.gov/12651724/)
29. Huang X. and Madan A. (1999) CAP3: A DNA sequence assembly program, Genome Res. 9, 868-877. doi: [10.1101/gr.9.9.868](https://doi.org/10.1101/gr.9.9.868). PMID: [10508846](https://pubmed.ncbi.nlm.nih.gov/10508846/)

30. Blanco E. *et al.* (2007) Using geneid to identify genes, Curr Protoc Bioinformatics. Jun;Chapter 4:Unit 4.3. doi: [10.1002/0471250953.bi0403s18](https://doi.org/10.1002/0471250953.bi0403s18). PMID: [18428791](https://pubmed.ncbi.nlm.nih.gov/18428791/)
31. 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. doi: [10.1186/1471-2105-12-323](https://doi.org/10.1186/1471-2105-12-323). PMID: [21816040](https://pubmed.ncbi.nlm.nih.gov/21816040/)
32. Anders S, and Huber W. (2010) Differential expression analysis for sequence count data, Genome Biol. 11(10):R106. doi: [10.1038/npre.2010.4282.1](https://doi.org/10.1038/npre.2010.4282.1). PMID: [20979621](https://pubmed.ncbi.nlm.nih.gov/20979621/)
33. Kadota K et al. (2012) A normalization strategy for comparing tag count data, Algorithms Mol Biol. 7(1):5. doi: [10.1186/1748-7188-7-5](https://doi.org/10.1186/1748-7188-7-5). PMID: [22475125](https://pubmed.ncbi.nlm.nih.gov/22475125/)
34. Schmittgen TD, & Livak KJ (2008) Analyzing real-time PCR data by the comparative C(T) method. Nat Protoc 3: 1101-1108, doi:[10.340/f.5500956.5467055](https://doi.org/10.340/f.5500956.5467055). PMID: [18546601](https://pubmed.ncbi.nlm.nih.gov/18546601/)
35. Maffei, M. E., Mithofer, A., & Boland, W. (2007). Before gene expression: early events in plant-insect interaction. Trends Plant Sci, 12(7), 310-316. doi: [10.1016/j.tplants.2007.06.001](https://doi.org/10.1016/j.tplants.2007.06.001). PMID: [17596996](https://pubmed.ncbi.nlm.nih.gov/17596996/)
36. 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](https://doi.org/10.1007/s00239-003-2568-5). PMID: [15170253](https://pubmed.ncbi.nlm.nih.gov/15170253/)
37. 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](https://doi.org/10.1111/1744-7917.12720). PMID: [31454147](https://pubmed.ncbi.nlm.nih.gov/31454147/)
38. 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](https://doi.org/10.1002/ps.5117). PMID: [29888850](https://pubmed.ncbi.nlm.nih.gov/29888850/)

- 560 **39.** Fu, Y., Zhu, J. Y., Zhang, F., Richman, A., Zhao, Z., & Han, Z. (2017). Comprehensive
561 functional analysis of Rab GTPases in *Drosophila* nephrocytes. *Cell Tissue Res*, 368(3),
562 615-627. doi:[10.1007/s00441-017-2575-2](https://doi.org/10.1007/s00441-017-2575-2). PMID: [28180992](https://pubmed.ncbi.nlm.nih.gov/28180992/)
- 563 **40.** Singh, D., & Kumar Roy, J. (2013). Rab11 plays an indispensable role in the
564 differentiation and development of the indirect flight muscles in *Drosophila*. *PLoS One*,
565 8(9), e73305. doi: [10.1371/journal.pone.0073305](https://doi.org/10.1371/journal.pone.0073305). PMID: [24023858](https://pubmed.ncbi.nlm.nih.gov/24023858/)
- 566 **41.** He, S., Tong, X., Han, M., Hu, H., & Dai, F. (2018). Genome-Wide Identification and
567 Characterization of WD40 Protein Genes in the Silkworm, *Bombyx mori*. *Int J Mol Sci*,
568 19(2). doi: [10.3390/ijms19020527](https://doi.org/10.3390/ijms19020527). PMID: [29425159](https://pubmed.ncbi.nlm.nih.gov/29425159/)
- 569 **42.** Orville Singh, C., Xin, H. H., Chen, R. T., Wang, M. X., Liang, S., Lu, Y., Cai, Z. Z;
570 Miao, Y. G. (2016). BmPLA2 containing conserved domain WD40 affects the metabolic
571 functions of fat body tissue in silkworm, *Bombyx mori*. *Insect Sci*, 23(1), 28-36. doi:
572 [10.1111/1744-7917.12189](https://doi.org/10.1111/1744-7917.12189). PMID: [25409652](https://pubmed.ncbi.nlm.nih.gov/25409652/)
- 573 **43.** Zheng, W., Peng, T., He, W., & Zhang, H. (2012). High-Throughput Sequencing to
574 Reveal Genes Involved in Reproduction and Development in *Bactrocera dorsalis*
575 (Diptera: Tephritidae). *PLoS ONE*, 7(5), e36463. doi:[10.1371/journal.pone.0036463](https://doi.org/10.1371/journal.pone.0036463).
576 PMID: [22570719](https://pubmed.ncbi.nlm.nih.gov/22570719/)
- 577 **44.** Jordan, M. D., Stanley, D., Marshall, S. D., De Silva, D., Crowhurst, R. N., Gleave, A.
578 P., . . . Newcomb, R. D. (2008). Expressed sequence tags and proteomics of antennae
579 from the tortricid moth, *Epiphyas postvittana*. *Insect Mol Biol*, 17(4), 361-373. doi:
580 [10.1111/j.1365-2583.2008.00812.x](https://doi.org/10.1111/j.1365-2583.2008.00812.x). PMID: [18651918](https://pubmed.ncbi.nlm.nih.gov/18651918/)
- 581 **45.** Sekine, K., Furusawa, T., & Hatakeyama, M. (2015). The boule gene is essential for
582 spermatogenesis of haploid insect male. *Dev Biol*, 399(1), 154-163. doi:
583 [10.1016/j.ydbio.2014.12.027](https://doi.org/10.1016/j.ydbio.2014.12.027). PMID: [25592223](https://pubmed.ncbi.nlm.nih.gov/25592223/)

46. 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](https://doi.org/10.3390/ijms20092361). PMID: [31086020](https://pubmed.ncbi.nlm.nih.gov/31086020/)
47. 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](https://doi.org/10.1186/s13071-019-3437-4). PMID: [30992032](https://pubmed.ncbi.nlm.nih.gov/30992032/)
48. 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](https://doi.org/10.1111/imb.12590). PMID: [31066110](https://pubmed.ncbi.nlm.nih.gov/31066110/)
49. 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](https://doi.org/10.1126/science.aaa2850). PMID: [25999371](https://pubmed.ncbi.nlm.nih.gov/25999371/)
50. McCambridge, A., Solanki, D., Olchawa, N., Govani, N., Trinidad, J. C., & Gao, M. (2020). Comparative Proteomics Reveal Me31B's Interactome Dynamics, Expression Regulation, and Assembly Mechanism into Germ Granules during Drosophila Germline Development. Sci Rep, 10(1), 564. doi: [10.1038/s41598-020-57492-y](https://doi.org/10.1038/s41598-020-57492-y). PMID: [31953495](https://pubmed.ncbi.nlm.nih.gov/31953495/)
51. Brady, M. M., McMahan, S., & Sekelsky, J. (2018). Loss of Drosophila Mei-41/ATR Alters Meiotic Crossover Patterning. Genetics, 208(2):579-588. doi: [10.1534/genetics.117.300634](https://doi.org/10.1534/genetics.117.300634). PMID: [29247012](https://pubmed.ncbi.nlm.nih.gov/29247012/)
52. Carter, T. Y., Gadwala, S., Chougule, A. B., Bui, A. P. N., Sanders, A. C., Chaerkady, R., Cormier, N.; Cole, R. N.; Thomas, J. H. (2019). Actomyosin contraction during

cellularization is regulated in part by Src64 control of Actin 5C protein levels. *Genesis*, 57(6), e23297. doi: [10.1002/dvg.23297](https://doi.org/10.1002/dvg.23297). PMID: [30974046](https://pubmed.ncbi.nlm.nih.gov/30974046/)

53. Eikenes, A. H., Malerod, L., Lie-Jensen, A., Sem Wegner, C., Brech, A., Liestol, K., Stenmark, H.; Haglund, K. (2015). Src64 controls a novel actin network required for proper ring canal formation in the *Drosophila* male germline. *Development*, 142(23), 4107-4118. doi: [10.1242/dev.124370](https://doi.org/10.1242/dev.124370). PMID: [26628094](https://pubmed.ncbi.nlm.nih.gov/26628094/)

54. Maksimov, D. A., Laktionov, P. P., Posukh, O. V., Belyakin, S. N., & Koryakov, D. E. (2018). Genome-wide analysis of SU(VAR)3-9 distribution in chromosomes of *Drosophila melanogaster*. *Chromosoma*, 127(1), 85-102. doi: [10.1007/s00412-017-0647-4](https://doi.org/10.1007/s00412-017-0647-4). PMID: [28975408](https://pubmed.ncbi.nlm.nih.gov/28975408/)

55. Sriskanthadevan-Pirahas, S., Lee, J., & Grewal, S. S. (2018). The EGF/Ras pathway controls growth in *Drosophila* via ribosomal RNA synthesis. *Dev Biol*, 439(1), 19-29. doi: [10.1016/j.ydbio.2018.04.006](https://doi.org/10.1016/j.ydbio.2018.04.006). PMID: [29660312](https://pubmed.ncbi.nlm.nih.gov/29660312/)

56. 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](https://doi.org/10.1673/031.011.5201). PMID: [21861656](https://pubmed.ncbi.nlm.nih.gov/21861656/)

Figure 1 Assembled unigene length distribution of *C. cramerella* transcriptome.

The x-axis indicates unigene size and the y-axis indicates the number of unigenes of each size.

Figure 2 Orthologous gene groups shared between *Conopomorpha*, *Drosophila*, *Bombyx* and *Helicoverpa*. Venn diagram of the distribution of the orthologous gene groups among the mentioned species.

Figure 3 Identity distribution of the top BLAST hits for each sequence of total 67,770 that has homology.

A

Biological Process

B

Molecular Function

C

Cellular Component

Figure 4 GO analyses of *Conopomorpha cramerella* transcriptome data. GO analysis of *Conopomorpha* sequences corresponding to a total of 285,882 contigs that are predicted to be involved in the biological processes (A) and molecular functions (B) and cellular component (C). Classified gene objects are depicted as percentages of the total number of gene objects with GO assignments.

Figure 5 Histogram of clusters of orthologous groups (COG) classification. A total of 3,296 predicted proteins have a COG classification among the 25 categories.

Figure 6 Differentially gene expression profile at different developmental stages. The number of up-regulated and down-regulated genes between larvae and pupae, between adults and pupae, and between adults and larvae are summarized here.

Figure 7 QRT-PCR validation of the differentially expressed genes between each of the two stages of growth (larva vs. pupa, pupa vs. moth and larva vs. moth). Relative transcript levels are calculated by real-time PCR using Actin gene as reference standard. Three biological replicates were performed, and the data shown are typical results.

825 **Figure caption** (All Figures: color on web; black and white in the print)

826

Figure 1	Assembled unigene length distribution of <i>C. cramerella</i> transcriptome. The x-axis indicates unigene size and the y-axis indicates the number of unigenes of each size.
Figure 2	Orthologous gene groups shared between <i>Conopomorpha</i>, <i>Drosophila</i>, <i>Bombyx</i> and <i>Helicoverpa</i>. Venn diagram of the distribution of the orthologous gene groups among the mentioned species.
Figure 3	Identity distribution of the top BLAST hits for each sequence of total 67,770 that has homology.
Figure 4	GO analyses of <i>Conopomorpha cramerella</i> transcriptome data. GO analysis of <i>Conopomorpha</i> sequences corresponding to a total of 285,882 contigs that are predicted to be involved in the biological processes (A) and molecular functions (B) and cellular component (C). Classified gene objects are depicted as percentages of the total number of gene objects with GO assignments.
Figure 5	Histogram of clusters of orthologous groups (COG) classification. A total of 3,296 predicted proteins have a COG classification among the 25 categories.
Figure 6	Differentially gene expression profile at different developmental stages. The number of up-regulated and down-regulated genes between larvae and pupae, between adults and pupae, and between adults and larvae are summarized here.
Figure 7	QRT-PCR validation of the differentially expressed genes between each of the two stages of growth (larva vs. pupa, pupa vs. moth and larva vs. moth). Relative transcript levels are calculated by real-time PCR using

	Actin gene as reference standard. Three biological replicates were performed, and the data shown are typical results.
--	---

827

828

Supporting information

Figure S1 Comparison of sequence expression between the larvae and pupae (A), moth and larvae (B), as well as moth and pupae (C). The abundance of each gene was normalised as Fragments Per Kilobase per Million (FPKM). The differentially expressed genes are shown in red and blue, while the other genes that are not differentially expressed (not DEGs) are shown in black.

alcohol

dehydrogenase

family

3-oxoacyl-[acyl-carrier-protein] reductase	TBIU004803	887	XP_004931809.1	<i>Bombyx mori</i>	3.00E-50	53.1
Carbonyl reductase [NADPH] 1-like	TBIU007337	1038	XP_004930987.1	<i>Bombyx mori</i>	4.00E-153	79.2
Alcohol dehydrogenase	TBIU013195	793	EHJ65258.1	<i>Danaus plexippus</i>	1.00E-100	58.8

Pimeloyl-ACP

methyl ester

carboxylesterase

Fatty alcohol acetyltransferase	TBIU001060	274	AIN34709.1	<i>Agrotis segetum</i>	8.00E-35	64.8
Juvenile hormone epoxide hydrolase-like protein 3	TBIU006811	232	NP_001159619.1	<i>Bombyx mori</i>	2.00E-33	76.6
Probable serine hydrolase-like	TBIU025807	1574	XP_004924867.1	<i>Bombyx mori</i>	6.00E-127	60.7

Tetratricopeptide

(TPR) repeat

Putative Heparan sulfate glucosamine 3-O-sulfotransferase 5	TBIU007544	1519	EHJ67645.1	<i>Danaus plexippus</i>	0	87.1
Hypothetical protein KGM_17730	TBIU014091	1210	EHJ68014.1	<i>Danaus plexippus</i>	1.00E-145	76.8

Regulator of	TBIU018173	982	XP_0049302	<i>Bombyx</i>	9.00E-124	70.2
microtubule			25.1	<i>mori</i>		
dynamics protein 1-						
like isoform X1						
WD40 repeat						
Hypothetical	TBIU002376	273	EHJ69771.1	<i>Danaus</i>	8.00E-46	90.1
protein				<i>plexippus</i>		
KGM_06966						
Guanine	TBIU005208	1450	EHJ71933.1	<i>Danaus</i>	0	88.4
nucleotide-binding				<i>plexippus</i>		
protein beta 2						
POC1 centriolar	TBIU011348	2486	XP_0049280	<i>Bombyx</i>	4.00E-166	80.8
protein homolog A-			14.1	<i>mori</i>		
like						

855

856

857

858

859

860

861

862

863

Table S2 Expression of genes in three different development stages of *Conopomorpha cramerella*.

Name	Gene				Gene (> FPKM 1.0)			
	Expressed	Known	Novel	Unexpressed	Expressed	Known	Novel	Unexpressed
Larva	143,065	14,236	128,829	142,817	140,427	14,023	126,404	74,966
Pupae	125,651	13,518	112,133	160,231	124,368	13,417	110,951	91,025
Moth	131,120	13,689	117,431	154,762	129,652	13,536	116,116	85,741

Table S3 Selected up-regulated and down regulated genes between samples (larva vs. pupa).

Gene name	Unigene ID	Gene description	Source	Expression – Larva (FPKM)	Expression – Pupa (FPKM)
Up-regulated genes					
atr	TBIU052540	Serine/threonine-protein kinase atr	SWISS;ACC :Q9DE14	0	10.7
me31B	TBIU052640	Putative ATP-dependent RNA helicase me31b	SWISS;ACC :P23128	0	5.46
dtwd1	TBIU052700	DTW domain-containing protein 1	SWISS;ACC :Q6DDV1	2.67	3.52
Trappc2	TBIU052984	Trafficking protein particle complex subunit 2	SWISS;ACC :D3ZVF4	0	1.64
AGBL1	TBIU052987	Cytosolic carboxypeptidase 4	SWISS;ACC :Q96MI9	0	5.25
Down-regulated genes					
Ubr5	TBIU052459	E3 ubiquitin-protein ligase	SWISS;ACC :Q62671	14.93	6.65
Src64B	TBIU052493	Tyrosine-protein kinase Src64B	SWISS;ACC :P00528	7.66	0
Dnah2	TBIU052596	Dynein heavy chain 2, axonemal	SWISS;ACC :P0C6F1	2.16	0

pycrl	TBIU046555	Pyrroline-5-carboxylate	SWISS;ACC	5.18	2.49
		reductase 3	:Q5SPD7		
Mtnd1	TBIU052872	NADH-ubiquinone	SWISS;ACC	3.27	0
		oxidoreductase chain 1	:P03888		

Table S4 Selected up-regulated and down regulated genes between samples (larva vs. moth).

Gene name	Unigene ID	Gene description	Source	Expression – Larva (FPKM)	Expression – Moth (FPKM)
Up-regulated					
genes					
Mitd1	TBIU053114	MIT domain- containing protein 1	SWISS;ACC :Q5I0J5	5.28	7.28
Setdb1	TBIU053165	Histone-lysine N- methyltransferase SETDB1	SWISS;ACC :O88974	0	5.59
JMJD4	TBIU053571	JmjC domain- containing protein 4	SWISS;ACC :Q5ZHV5	15.12	41.52
NDUFS5	TBIU053671	NADH dehydrogenase [ubiquinone] iron- sulfur protein 5	SWISS;ACC :Q0MQH3	342.69	967.27
RTase	TBIU054001	Probable RNA-directed DNA polymerase from transposon BS	SWISS;ACC :Q95SX7	0	2.17
Down-regulated					
genes					
let-268	TBIU052928	Procollagen-lysine,2- oxoglutarate 5- dioxygenase	SWISS;ACC :Q20679	2.74	0
pol	TBIU052994	RNA-directed DNA polymerase from mobile element jockey	SWISS;ACC :P21328	4.64	0

X-	TBIU053569	Probable RNA-directed	SWISS;ACC	5.93	1.21
element\ORF2		DNA polymerase from	:Q9NBX4		
		transposon X-element			
Hgsnat	TBIU053727	Heparan-alpha-	SWISS;ACC	4.82	2.37
		glucosaminide N-	:Q3UDW8		
		acetyltransferase			
exba	TBIU053780	Protein extra bases	SWISS;ACC	17.28	0
			:Q9VNE2		

Table S5 Selected up-regulated and down regulated genes between samples (pupa vs. moth).

Gene name	Unigene ID	Gene description	Source	Expression – Pupa (FPKM)	Expression – Moth (FPKM)
Up-regulated					
genes					
RpS25	TBIU052930	40S ribosomal protein S25	SWISS;ACC :Q962Q5	0	2.77
Slc2a13	TBIU053815	Proton myo-inositol cotransporter	SWISS;ACC :Q3UHK1	1.91	22
alg9	TBIU054025	Alpha-1,2- mannosyltransferase alg9	SWISS;ACC :Q9P7Q9	0	2.03
ABCF2	TBIU054126	ATP-binding cassette sub-family F member 2	SWISS;ACC :Q2KJA2	0	13.25
zc3h15	TBIU054262	Zinc finger CCCH domain-containing protein 15	SWISS;ACC :Q803J8	0	2.21
Down-regulated					
genes					
SDCBP	TBIU045865	Syntenin-1	SWISS;ACC :O00560	1.35	0
Rpl12	TBIU055268	60S ribosomal protein L12	SWISS;ACC :P35979	4.95	0
4CLL4	TBIU056797	4-coumarate--CoA ligase-like 4	SWISS;ACC :P0C5B6	6.99	0

Bap60	TBIU056833	Brahma-associated protein of 60 kDa	SWISS;ACC :Q9VYG2	4.87	0
Prm	TBIU057029	Paramyosin, long form	SWISS;ACC :P35415	4.13	0