RNA-SEQ ASSISTED GENE MODELING AND ANNOTATION, TRANSCRIPTOME STUDY AND FUNCTIONAL ANALYSIS OF STRESS RESPONSIVE PEPTIDES OF *MANDUCA SEXTA*

By

XIAOLONG CAO

Bachelor of Biological Science
University of Science and Technology of China
Hefei, Anhui
2011

Master of Science in Biochemistry & Molecular Biology
Oklahoma State University
Stillwater, Oklahoma
2015

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
December, 2016

RNA-SEQ ASSISTED GENE MODELING AND ANNOTATION, TRANSCRIPTOME STUDY AND FUNCTIONAL ANALYSIS OF STRESS RESPONSIVE PEPTIDES OF *MANDUCA SEXTA*

Dissertation Approved:

Dr. Haobo Jiang

Dissertation Adviser

Dr. Junpeng Deng

Dr. Ulrich Melcher

Dr. José Luis Soulages

Dr. Guolong Zhang

ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Haobo Jiang for providing me a chance to study and do research in this laboratory. Also, during the five and half years of my PhD study, he has kindly guided my research and study, remained patience when sometimes I became emotional, and provided much help. He is not only my advisor, but also a good friend, willing to share experience and happiness in life with us.

I would like to thank my committee members, Dr. Junpeng Deng, Dr. Ulrich Melcher, Dr. José Luis Soulages, Dr. Guolong Zhang, and previous member, Dr. Ramamurthy Mahalingam. Their guidance and assistance have helped a lot in my study and research.

Many thanks to members of the Biochemistry and Molecular Biology department, including our department head, Dr. John E. Gustafson, Graduate Program Coordinator, Dr. Robert Matts, and the office staff. I appreciate knowledge I learned from teachers, suggestions and technical supports from Dr. Steve Hartson and Janet Rogers, and friendship and great time with Junho Cho and other graduate students.

I would like to thank current and previous lab members, especially Yang Wang, who is our lab manager and helped us with all kinds of research works. We work together more like a big family, with everyone caring and helping each other. These great people include Yingxia Hu, Xiufeng Zhang, Shuguang Zhang, Yan He, Xuesong He, Ramesh Gunaratna, Krishna Bhattarai and Mansi Gulati. Special Thanks to Yan He as we've been good friends for nearly ten years. I will cherish the valuable friendship forever.

I would like to thank my friends in Stillwater, especially many previous members of Chinese Friendship Association. I enjoyed the time we worked together to help Chinese students and promote Chinese culture. I have great memories of playing badminton and joining parties in Stillwater, this great peaceful town.

Finally, I would like to thank my parents. They raised me and taught me to be a good man. They respect and support the decisions I make. With their love, I will always try to be a strong man. Science is great. I hope I will never lose my curiosity and passion for science, and do something good for society.

Name: XIAOLONG CAO

Date of Degree: DECEMBER, 2016

Title of Study: RNA-SEQ ASSISTED GENE MODELING AND ANNOTATION, TRANSCRIPTOME STUDY AND FUNCTIONAL ANALYSIS OF STRESS RESPONSIVE PEPTIDES OF *MANDUCA SEXTA*

Major Field: BIOCHEMISTY AND MOLECULAR BIOLOGY

Abstract:

Manduca sexta is a widely used model insect. The genome sequence was determined using 454 sequencing technology, and Official Gene Set (OGS) 2.0 was generated with help from RNA-seq data and manually annotation by researchers all over the world. To improve gene models, we developed methods to compare and select gene models by **M**AKER2, **C**ufflinks, **O**ases and **T**rinity, and generate a new gene set, called MCOT 1.0. Compared with OGS 2.0, MCOT 1.0 has higher quality score as evaluated by BUSCO, and with nearly 50% more unique proteins being predicted.

The immune signaling pathways are critical for proper defense against pathogens for insects. To facilitate systematic studies of *M. sexta* immune system, we have identified and verified participant genes in the genome of *M. sexta*. We annotated 186 genes which encode 199 proteins in Toll, Imd, MAPK-JNK-p38, JAK-STAT, autophagy, apoptosis and RNA interference pathways, analyzed their evolution and mRNA levels in different tissues and different developmental stages.

To date, 67 cDNA libraries have been sequenced from different tissues and different developing stages of *M. sexta*. However, there is no systematic analyzation of these RNA-seq data. We examined each library, found possible contaminant reads in each of these libraries, compared library similarity based on associated genes, and analyzed gene expression in different libraries. We found that most genes were expressed in library-specific manner, and their expression patterns would help functional study in the future.

Stress Responsive Peptides (SRPs) are cytokines activated under biotic and abiotic stresses, which may act as key signaling molecules for humoral and neural regulation of immune or other responses. Eight SRPs were identified in the genome of *M. sexta*. With similar amino acid sequence, their functions are very different. SRP6 can inhibit the growth of larvae, while SRP1 and SRP2 can induce the expression of different anti-microbial peptides. We verified activation site of SRP1 and SRP2 with MALDI-MS, and identified PAP3 as the upstream enzyme which can activate them. This study will help understand the roles of insect cytokines.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Chapter IV

CHAPTER I

INTEGRATED MODELING OF PROTEIN-CODING GENES IN THE MANDUCA SEXTA GENOME USING RNA-SEQ DATA FROM THE BIOCHEMICAL MODEL INSECT

Xiaolong Cao[a], Haobo Jiang[b]

[a] Department of Biochemistry and Molecular Biology, Oklahoma State University, Stillwater, OK 74078, USA
[b] Department of Entomology and Plant Pathology, Oklahoma State University, Stillwater, OK 74078, USA

Key words: gene annotation; de novo assembly; tobacco hornworm; automated gene modeling; arthropod genomics.

***Abbreviations*:** OGS, official gene set; ORF, open reading frame; L, length; ML, match length; QL, query length; SL, subject length; M, MAKER; C, Cufflinks; T, Trinity; O, Oases; U, UniProt Arthropoda; Y, C/T/O; S, similarity ratio of lengths; MLI, match length index; S1/S2, Selection 1 or 2; "P", perfect; "N", near perfect; "O", okay; "B", bad; "W", worst.

Abstract

The genome sequence of *Manduca sexta* was recently determined using 454 technology. Cufflinks and MAKER2 were used to establish gene models in the genome assembly based on the RNA-Seq data and other species' sequences. Aided by the extensive RNA-Seq data from 50 tissue samples at various life stages, annotators over the world (including the present authors) have manually confirmed and improved a small percentage of the models after spending months of effort. While such collaborative efforts are highly commendable, many of the predicted genes still have problems which may hamper future research on this insect species. As a biochemical model representing lepidopteran pests, *M. sexta* has been used extensively to study insect physiological processes for over five decades. In this work, we assembled *Manduca* datasets Cufflinks 3.0, Trinity 4.0, and Oases 4.0 to assist the manual annotation efforts and development of Official Gene Set (OGS) 2.0. To further improve annotation quality, we developed methods to evaluate gene models in the **M**AKER2, **C**ufflinks, **O**ases and **T**rinity assemblies and selected the best ones to constitute MCOT 1.0 after thorough crosschecking. MCOT 1.0 has 18,089 genes encoding 31,666 proteins: 32.8% match OGS 2.0 models perfectly or near perfectly, 11,747 differ considerably, and 29.5% are absent in OGS 2.0. Future automation of this process is anticipated to greatly reduce human efforts in generating comprehensive, reliable models of structural genes in other genome projects where extensive RNA-Seq data are available.

1. Introduction

With five larval instars, a large body size and hemolymph volume, and a simple larval body structure, the tobacco hornworm *Manduca sexta* has been widely employed as a model organism to study basic physiological processes in insects, such as cuticle formation, neural transmission, hormonal regulation, nutrient transport, intermediary metabolism, and immune responses (Hopkins et al., 2000; Shield and Hildebrand, 2001; Riddiford et al., 2003; Kanost et al., 1990; Arrese and Soulages, 2010; Jiang et al., 2010). Acquired knowledge of the molecular mechanisms underlying these processes would lead to new means of pest control, because *M. sexta* may be a good representative of some serious agricultural pests in the order of Lepidoptera. Several transcriptome analyses have yielded sequences and expression patterns of genes related to immunity, digestion, and olfaction (Zou et al., 2008; Pauchet et al., 2010; Zhang et al., 2011; Grosse-Wilde et al., 2011; Gunaratna and Jiang, 2013), but the potential of this model species is far from fulfillment partly due to the lack of its genome sequence. The shortage of complete protein sequences based on correctly modeled genes substantially hampers proteomic studies, for instance, of the immune complex formed around entomopathogens.

Recently, the genomic DNA isolated from a single male pupa of *M. sexta* was pyrosequenced at >20-fold coverage and assembled into *Manduca* Genome Assembly 1.0 (Msex 1.0) using Newbler with Atlas-GapFill (Kanost et al., 2016). Sixty cDNA libraries, representing mRNA samples of whole larvae, organs and tissues at various developmental stages, were sequenced using Illumina technology, yielding >350 gigabyte data. Some of these RNA-Seq datasets and other known *M. sexta* cDNA sequences were aligned to the reference genome to generate *Manduca* Cufflinks Assembly 1.0 and 1.0b using Bowtie, TopHat, and Cufflinks. Aided by the available sequence data from *M. sexta* and other arthropod species, approximately 18,000

genes in Msex 1.0 were predicted by MAKER2 generating the *Manduca* Official Gene Set 1.0 (OGS 1.0). Some of the OGS 1.0 models were examined by annotators to detect errors using *Manduca* Cufflinks 1.0/1.0b, Trinity 3.0, and Oases 3.0 sequences. The latter two sets of gene transcripts, assembled solely based on the RNA-Seq datasets, were extensively used along with Cufflinks 1.0/1.0b to improve annotation quality. Over a period of more than one year, 2,498 structural genes were successfully curated by approximately 70 researchers (Kanost et al., 2016). PASA2 (http://pasa.sourceforge.net/) was then used to select the best models from the MAKER2, Cufflinks, Trinity, Oases, and manual assemblies to generate *Manduca* OGS 2.0 (Kanost et al., 2016).

During the course of gene cross-examination, we came to realize that some of the lessons learned can be valuable to future genome projects. For example, as extensive RNA-Seq data are becoming a norm, genome-dependent and independent assemblies are critically important in the validation and perfection of MAKER2 gene models. Due to limitations of the programs used to produce OGS 2.0 (Table 1), an integration of their outputs using computer programs may greatly reduce human efforts in sequence cross-examination and considerably increase the percentage of crosschecked gene models. To achieve these goals, we have developed methods to evaluate models in the **M**AKER, **C**ufflinks, **O**ases and **T**rinity assemblies. As proof of principle, a reliable, nearly complete set of protein sequences (MCOT 1.0) is generated to facilitate proteomic research in this model insect. In the following, we report the generation of Cufflinks 3.0, Oases 4.0 and Trinity 4.0 gene models, discuss their advantages, shortcomings and integration, and describe how MCOT 1.0 was developed and compared with OGS 2.0.

2. Materials and Methods

2.1. Data and program acquisition

Manduca Genome Assembly 1.0 (Msex 1.0) and gene models in *Manduca* Official Gene Sets 1.0 (OGS 1.0, Table S1) and 2.0 (OGS 2.0) and Cufflinks Assembly 1.0 (Cufflinks 1.0) (Kanost et al., 2016) were downloaded from *Manduca* Base (ftp://ftp.bioinformatics.ksu.edu/pub/Manduca/). Universal protein sequences in UniProtKB Arthropoda (Table S1) were downloaded from ftp://ftp.ebi.ac.uk/pub/databases/fastafiles/uniprot/. The RNA-Seq datasets (Kanost et al., 2016) were acquired from Dr. Gary Blissard at Cornell University. SAMtools (0.1.19) (Li et al., 2009), Bowtie2 (2.2.1) (Langmead and Salzberg, 2012), TopHat (2.0.11) (Trapnell et al., 2009), Cufflinks (2.1.1) (Trapnell et al., 2012; Roberts et al., 2011), Trinity (20131110) (Grabherr et al., 2011), Oases (0.2.08) (Schulz et al., 2012), and BLAST+ (2.2.29) (Camacho et al., 2009) were downloaded from http://samtools.sourceforge.net/, http://bowtie-bio.sourceforge.net/bowtie2/index.shtml, http://ccb.jhu.edu/software/tophat/index.shtml, http://cufflinks.cbcb.umd.edu/, http://trinityrnaseq.sourceforge.net/, https://www.ebi.ac.uk/~zerbino/oases/, ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ and installed on a local supercomputer according to their manuals.

2.2. Generation of Cufflinks 3.0

The 60 RNA-Seq datasets were aligned to Msex 1.0 using TopHat at settings for three different read types: single end, paired end, and strand specific. "--read-realign-edit-dist 0" was selected to increase accuracy of read alignments. Cufflinks was used to translate the accepted hits

generated by TopHat to separate GTF files, with the "-u" command enabled to allow more accurate handling of multiple reads mapped to the same region. Cuffmerge was employed to combine GTF files of all the libraries to make the final GFF file (see scripts in the Supplemental Materials), from which transcript sequences were extracted using gffread to form Cufflinks 3.0 dataset (Table S1).

2.3. Reads treatment, normalization, and de novo assembling

Paired end reads were trimmed to 80 bp using FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html), with the forward reads combined in one file and the reverse ones in another. To handle the RNA-Seq data with 256 GB RAM of the supercomputer, the number of the reads was reduced according to Haas et al (2013). The Perl scripts provided in Trinity were used to perform *in silico* read normalization with maximum coverage set to 500. The single end and strand-specific reads were combined in one file for normalization at the same maximum coverage. After all normalized reads were pooled, Trinity was used to assemble the reads as paired end reads, generating Trinity 4.0 (Table S1). For Oases, four hash lengths (*k*: 25, 27, 29, 31) were chosen to assemble the reads as single end reads in four separate runs. Scaffolding was not allowed, preventing the stretches of Ns in assembled transcripts. The transcript files were then merged according to the Oases manual, generating Oases 4.0 (Table S1). In addition, reads that cannot be aligned to Msex 1.0 by TopHat were combined, trimmed to 80 bp, and assembled as paired end reads using Trinity. This new assembly (Trinity 4.0b, Table S1) was later used to identify unmapped genes, some of which may reside on the unsequenced W chromosome.

2.4. Gene translation and sequence comparison

Gene transcripts in Trinity 4.0 and Oases 4.0 were translated to polypeptide sequences using TransDecoder in Trinity (http://transdecoder.sourceforge.net/) (Haas et al., 2013), with minimum protein length set at 60. For removing redundant sequences in the *de novo* assemblies, identical proteins were identified in one batch using Python scripts and only one of each group was kept in the final set of unique protein sequences. To identify the best sequences in the comparisons between assemblies, the BLOSUM62 scoring matrix in the BLAST source code was changed to -100 for all 190 non-identical residue pairs. As such, only identical or near identical sequences would be detected by BLASTP with a positive score of alignment. The gap opening penalty was set to the maximum (32,767) to avoid gapped matches. A batchwise BLASTP comparison of the two sets of translated sequences was performed, with the tabular outputs (*e.g*. match length, query length, subject length) exported to Excel for further analysis. Cufflinks 3.0 translations were used as queries to search Trinity 4.0 or Oases 4.0 translations.

2.5. Cross-examination and selection of protein sequences from different assemblies

As illustrated in the flowchart (Fig. 1), the BLASTP results from comparisons of the unique protein sequences in **C**ufflinks 3.0, **T**rinity 4.0, and **O**ases 4.0 were examined by two methods to establish Selections 1 and 2. The results from one method were then cross-examined by the other to yield a dataset COT, later becoming a major part of MCOT 1.0.

In the length-based method, the Cufflink-Trinity comparison resulted in pairs with **m**atch **l**engths (TMLs, T for **T**rinity) and **C**ufflinks **l**engths (CLs), and their ratios were used to determine whether or not the Trinity hits would be kept (Fig. 1A). If TML/CL was ≤ 0.7, the Trinity hits were ignored and corresponding Cufflinks sequences were further processed: for

ones without ambiguous residues (Xs), their lengths (CLs) were directly used as CL*s; for the others with Xs, 70% of the CL values were used as CL*s. On the other hand, if TML/CL was > 0.7, the Trinity sequences were considered in the next step. The same procedure was carried out to compare Cufflinks and Oases translations and select the Oases ones (OML/CL > 0.7) for further consideration, together with the selected Cufflinks and Trinity sequences. The ones with the largest values (CL*, TL, or OL) were kept in Selection 1. If the values were equal, retention priority was given to the concerning sequences in the following order: Cufflinks, Trinity, and then Oases.

In the ratio-based method, Cufflinks 3.0 translations were used as queries to search arthropod **u**niversal/**U**niProt (U) sequences using BLASTP with the original BLOSUM62 matrix (Fig. 1B). Results were kept if identity >35% and ML/QL >0.7 or ML >200. When several regions were matched, ML equals the sum of match lengths between the same query and subject sequences. Up to five top hits were used to calculate UL (for **U**niProt **l**ength: mean ± SD) and ID of the best match was kept. Lengths (CL, TL, OL and UL) of the Cufflinks 3.0, Trinity, Oases, and UniProt proteins, correlated by the BLASTP searches, were used to calculate similarity ratios CUS, TUS, and OUS. For example, TUS (*i.e*. **S**imilarity ratio of lengths in a **T**-**U** comparison) was defined as TL/UL or UL/TL, whichever is between 0 and 1, so that a TUS close to 1 indicates high similarity between this Trinity-UniProt pair. Depending on the absence or presence of Xs in the Cufflinks translations, CUS was directly used or adjusted to 70% as CUS*. The proteins with the highest ratios (CUS*, TUS or OUS) will be kept in Selection 2 and, if the values were equal, the priority order of C > T > O was used to determine which ones to retain.

To cross-examine the two selections, the **l**ength (L) and **m**atch **l**ength (ML) of sequence Y (C or T or O) in Selection 2 (S2), UL of its correlated UniProt sequence, L and ML of its correlated sequence in Selection 1 (S1) were used to calculate $YUS_{S2} - YUS_{S1}$ (Fig. 1C). YUS = L/UL or UL/L, whichever is 0 to 1. Sequences in S1 were kept if their $YUS_{S2} - YUS_{S1} < 0.3$, $ML_{S1}/CL > 0.95$, or $ML_{S1}/CL > 0.8$ when Cufflinks sequence contains Xs (route 1). Sequences in S2 were retained, if their $YUS_{S2} - YUS_{S1} > 0.5$ and $L_{S2}/CL > 0.7$ (route 2). The remaining sequence pairs in the two selections were manually scrutinized to determine which ones to keep (route 3). In most cases, S1 and S2 were identical ($YUS_{S2} = YUS_{S1}$).

2.6. Classification of sequence comparison results

If the lengths of a query sequence (QL), subject sequence (SL), and match length (ML) were identical (QL = SL = ML), the match was considered as "P" (for perfect). If (ML/QL) ×(ML/SL) > 0.95 (*e.g.* when ML = QL, ML/SL > 0.95), the match was "N" (for near perfect). The 3rd and 4th categories "O" (for okay) and "B" (for bad) were separated based on **m**atch **l**ength **i**ndex (MLI), defined as (ML/QL)/0.7 + ML/200. If MLI was ≥ 1, the match was "O". In other words, even if QL is much greater than ML, >200 residues match is significant. Or, when ML/200 is small, >70% of QL falls into the matched region is considerable. If MLI was < 1, the match was "B". In the last category of "W" (for worst), the query sequences had no match. When OGS 1.0 and Cufflinks 3.0 datasets were compared, OGS 1.0 IDs with "B" and "W" matches were recorded.

2.7. Identification of proteins present only in OGS 1.0

Although accuracy of the gene models in OGS 1.0 is relatively low, some are unique (Table 1). Since Cufflinks is more sensitive than Trinity and Oases (Yandell and Ence, 2012),

MAKER2 proteins were used as queries to search the Cufflinks 3.0 translations using BLASTP with the modified scoring matrix, according to *Section 2.4*. Based on the results, those sequences in the categories of "B" or "W" were stored as "M" (for MAKER2 unique proteins), later incorporated into MCOT 1.0.

2.8. Identification of unmapped genes in Trinity 4.0b

Since a male pupa was used for genome sequencing, genes located on the W chromosome are not present in Msex 1.0. In addition, the genome assembly probably lacks genes or gene pieces on other chromosomes, as gaps between scaffolds or NNN regions. Trinity 4.0b was used to uncover transcripts of such unmapped genes. Based on results of the MCOT-Trinity 4.0b comparison, Trinity 4.0b protein sequences in the categories of "B" or "W" were kept for BLASTP search of arthropod UniProt sequences using the original BLOSUM62 scoring matrix. Hits with ML > 100, identity > 35%, and minimum/maximum of ML, QL and SL > 0.7 were combined with the proteins in "M" (*Section 2.7*) and "COT" (*Section 2.5*) to generate MCOT 1.0 (Table S1) after redundant sequences were removed. The redundant ones were identical sequences or shorter sequences (with zero or three residues trimmed off from both ends) identical to a part of longer ones.

3. Results and discussion

3.1. Manduca Genome Assembly 1.0

Shotgun sequencing of *M. sexta* genomic DNA fragments by the 454 technology resulted in a dataset at >20-fold of the genome size (422 ± 12 Mb), which was then assembled into Msex

1.0 (Kanost et al., 2016). The genome assembly consists of 20,891 scaffolds (Table 2) with N_{50} at 664 kb, much longer than the size of a typical lepidopteran insect gene. While this sequence set is good enough for gene modeling, other features may complicate the process: 1) 50.5% and 41.0% of the scaffolds are <1 kb and 1 kb to 10 kb, accounting for 1.70% and 4.05% of the 419 Mb assembly size, respectively (Fig. 2A); 2) over 17,000 undetermined nucleotide (NNN) regions (average: 1,118 bp; range: 1–124,308 bp) (Fig. 2B) may contain genes or gene elements, even though they only account for 4.71% of the entire assembly; 3) conserved and novel repetitive elements, accounting for 25% of Msex 1.0 (Kanost et al., 2016), and other highly similar sequences may cause errors in this assembly (Cao et al., 2015). Consequently, gene modeling can be a challenge in some cases.

3.2. Manduca Cufflinks Assembly 3.0

Cufflinks uses RNA-Seq data to model genes in a genome assembly (Table 1) (Trapnell et al., 2012; Roberts et al., 2011). We took advantage of Msex 1.0 and all 60 RNA-Seq datasets (Kanost et al., 2016) to generate a new assembly, namely Cufflinks 3.0. As an update of *Manduca* Cufflinks 1.0, assembled using 33 of the 60 libraries, Cufflinks 3.0 contains 36,027 genes and 62,497 transcripts (Table 3). Cufflinks 1.0 has 37,281 genes and 64,301 transcripts. Perhaps, lacking RNA-Seq data support from scarcely expressed genes has split some genes and their transcripts into two or more pieces in Cufflinks 1.0. Analysis of Cufflinks 3.0 dataset indicates that 75% of the genes have one transcript form and 16% have 2 or 3 splicing alternates (Fig. 3). Thus, alternative splicing appears to be a minor concern for the genes predicted autonomously. In comparison, 96% of the MAKER2 gene models in OGS 1.0 have no splice variant, indicating this program is not good at predicting such variations.

3.3. Trinity and Oases assemblies

Based on the same reference genome, Cufflinks and MAKER2 may incorrectly predict genes if there are flaws in their corresponding genomic regions (Table 1, *Section 3.1*). To discover and repair this problem, we *de novo* assembled transcripts using the 60 RNA-Seq datasets. Totally, 317,062 transcripts corresponding to 193,161 genes were established using Trinity and 552,733 from 88,397 genes by Oases (Table 3). Due to characteristics of the Trinity and Oases programs (Table 1), the transcript numbers were 5.1 to 27.2-fold higher than those in Cufflinks 3.0 and OGS 1.0. The percentages of short transcripts (< 512 bp) were 48% in Trinity and 30% in Oases, much higher than 14% in Cufflinks 3.0 or OGS 1.0 (Fig. 4A, Table S2). Many of the short contigs in the genome-independent assemblies were probably caused by how these different programs handle problems such as single nucleotide polymorphisms, low quality reads, and posttranscriptional modifications. While Oases allows multiple hash levels, merging them does not necessarily produce a better assembly than Trinity did. The gene number was 88,397 or 45.8% of the Trinity models, but the protein number (total: 304,367, unique: 130,474) was 1.95- and 2.27-fold of the Trinity proteins (total: 155,825, unique: 57,593) (Table 3). Nonetheless, the numbers of transcripts and unique proteins in different size ranges (Fig. 4, A and B) did indicate that the RNA-Seq datasets were large and diverse enough for modeling a majority of the active genes and, in some cases, their splicing variants, all based on experimental evidence.

3.4. Translation of the gene model sets

We focus our efforts on structural genes to make *M. sexta* amenable to proteomic studies in the future. By translating their transcripts and setting the size limit to > 60 residues, we expect

to detect antimicrobial peptides (*e.g*. cecropins) but not some neuropeptides that are too small to tell apart from the noise of short open reading frames (ORFs). Some of the transcripts contain two or more ORFs, in most cases due to the merging of adjacent genes. As an extreme example, MAKER2 merged eleven adjacent genes into one coding for a gigantic "polyprotease". While the transcript numbers in Trinity and Oases are 5.1 and 8.8 times of that in Cufflinks, the numbers of unique proteins are just 1.5 to 3.5 times respectively (Table 3, Fig. 4B), suggesting that differences in the non-coding regions may also be responsible for the high transcript counts. Based on the protein size distribution (Table S3), Cufflinks outperforms the other three programs in modeling proteins longer than 2,049 residues, owing to its high sensitivity and reliance on Msex 1.0 (Table 1). The unique proteins shorter than 2,048 residues in Oases 4.0 are significantly higher in number than those in Trinity 4.0, then Cufflinks 3.0, and OGS 1.0 at last (Fig. 4B). Although part of this could be an artifact caused by Oases and Trinity to a lesser extent, the *de novo* assemblies well complement the other two assemblies by closing the gaps in Msex 1.0 (Table 1). MAKER2, primarily designed to model structural genes, has generated OGS 1.0. Albeit the smallest, this assembly contains unique genes. These genes are either scarcely expressed in the 52 tissue samples or expressed in unsampled tissues or stages so that they are not detected even by Cufflinks. In summary, an integration of the assemblies is necessary to generate a reliable, concise, and complete set of structural genes.

3.5. Comparison of proteins in OGS 1.0 and Cufflinks 3.0

To facilitate comparison among the four *M. sexta* assemblies, we modified the scoring matrix of BLASTP so that all non-identical residue pairs (*e.g*. Leu and Ile) score -100 (*Section 2.4*). Consequently, unless there is a long stretch of identical or near identical amino acid sequence in a query and a subject, the comparison always yields a negative score, allowing us to ignore

the less-than-perfect matches that cause complications. After the proteins in OGS 1.0 and Cufflinks 3.0 were compared, 17,907 of the pairs were 100% identical, 226 were 98.0 to 99.9% identical, and these two groups together accounted for 99.95% of the total matches (Table 4). In this way, match length (ML) in the query (Q) and subject (S) were directly used to calculate (ML/QL) ×(ML/SL) and (ML/QL)/0.7 + ML/200 (*i.e*. MLI or **m**atch **l**ength **i**ndex), without any concern about the exact percentage identity. The ML, QL, SL, (ML/QL) ×(ML/SL) and MLI values were then used to categorize the matches into "P", "N", "O", "B", and "W" (*Section 2.6*). Among the 22,310 unique proteins from the MAKER2 models, 6,481 perfectly and 2,245 near perfectly matched those from Cufflinks 3.0 (Table 5). Together, they account for 39.1% of the total. Another 39.1% fall into the "O" category. Proteins in the categories "B" (678) and "W" (4,177) are considered to be unique, as they are not modeled by Cufflinks, Trinity, or Oases. The latter two are less sensitive than Cufflinks (Table 1).

3.6. Comparison of proteins in Trinity 4.0, Oases 4.0, and Cufflinks 3.0

Using the same method, we separately compared proteins in Cufflinks 3.0 with Trinity 4.0 and Oases 4.0 translations. Because translations of the MAKER2 models (*Section 3.5*), *de novo* assemblies, and arthropod UniProt sequences (*Section 3.7*) were all compared with translations of Cufflinks 3.0, identifications of the Cufflinks hits from these BLASTP searches serve as a liaison for all these datasets. The correlated protein sequences can then be evaluated to find the best model (Fig. 1).

In the comparison of Cufflinks 3.0 with Oases 4.0 translations, for example, 67.8% of the total matched sequences had ML/QL > 0.95 (Fig. 5A). The rest of hits fell into the realms of 0.95-0.7 (19.7%) and 0.7-0 (12.5%). We arbitrarily set the ML/QL threshold at 0.7 to identify Q

and S sequences representing the same gene and kept the longer ones in Selection 1 (Fig. 1A). Likewise we found that 39.9% of the total O-C matches had (ML/QL)×(ML/SL) > 0.95 (Fig. 5B); 1.7% of the total had (ML/QL)/0.7 + ML/200 (*i.e*. match length indices or MLIs) less than one (Fig. 5C). Using cutoff values of 1.0 for ML/QL, 0.95 for (ML/QL)×(ML/SL), and 1 for MLI, we categorized the matches into "P", "N", "O", "B" or "W". By correlating the results from T-C (Trinity 4.0 vs. Cufflinks 3.0) and O-C comparisons (Fig. 1A), we found 5,516 and 968 of the proteins in Cufflinks 3.0 perfectly and near perfectly matched both Trinity and Oases models (Table 6), respectively. Among the 37,316 total hits, 26,702 (71.6%) fell into the same categories (P, N, O, B or W) from the comparisons, indicating that Trinity and Oases models are consistent in the protein-coding region at least. While 7,094 or 19.4% of the proteins were highly reliable (PP, NP, PN, and NN), 1,944 or 5.2% (BB, BW, WB, and WW) were probably modeled by Cufflinks only due to its high sensitivity (Table 1). The P/N/O proteins distributed normally over a broad size range; 68.7% of the B/W were short (<128 residues) (Table S4 and Fig. 6). Possibly the short proteins came from untranslated regions of some genes, noncoding RNAs, or small protein genes expressed but undetected. In contrast to these extreme categories, 18,509 or 49.6% of the 37,316 total hits belong to the OO comparison and further efforts were made to select useful information from these sequences.

3.7. Comparison of proteins in UniProtKB Arthropoda and Cufflinks 3.0

Reliable proteins from other arthropods are useful for validating gene models. Therefore, we used BLASTP algorithm and the original BLOSUM62 matrix to compare query (Q) proteins in Cufflinks 3.0 translations with UniProtKB Arthropoda (*i.e*. **U**niProt or U) as described in *Section 2.5*. Of the 37,316 unique proteins in the Cufflinks 3.0, 30,313 or 81.2% had one to five matches; 7,003 had no match and may be unique in *M. sexta*. Their length distributions

were normal distributions for the ones with 1 to 5 matches, but not so for those with 0 match (Table S5, Fig. 7) – 3,149 or 45.0% of them were shorter than 128 residues. Some of the small proteins may not exist and it is also possible that BLASTP at the default settings has bias against short proteins. Nonetheless, assuming the sequence lengths of orthologous proteins are similar, we can exploit the links among UniProt, Cufflinks, Trinity, and Oases datasets to choose models by the ratio-based method to generate Selection 2 (Fig. 1B).

3.8. Model selection among Cufflinks 3.0, Trinity 4.0, and Oases 4.0

For all hits with ML/CL > 0.7, we chose the longest models for Selection 1 (S1, Fig. 1A, *Section 2.5*). When Xs (caused by NNNs) were present in the Cufflinks translations, the use of CL* (*i.e.* 0.7CL), instead of CL, allowed the *de novo* proteins to survive and replace the ambiguous Cufflinks models. To complement S1, lengths of the Trinity, Oases, Cufflinks, and UniProt (U) proteins, correlated through Cufflinks IDs from the T-C, T-O, and T-U comparisons, were used to calculate the similarity ratios TUS, CUS* and OUS (*Section 2.5*, Fig. 1B). The models with ratios closest to 1.0 were kept in Selection 2 (S2). Cross-examination of the correlated proteins in S1 and S2 by ratio comparison (YUS_{S2}-YUS_{S1}) resulted in the retention of 36,205 proteins without Xs (Fig. 1C, route 1). Crosschecking S2 contributed 35 proteins (route 2); manual checking improved the other 77 in S1 or S2 (route 3). Of the 999 sequences with Xs, 996 were selected via route 1 and three via route 2. Of 36,317 proteins without Xs, 29,612 have the same S1 and S2 result, and the rest, 6,593 keep S1 (route 1), 35 keep S2 (route 2) and only 77 needed manual checking (route 3).

3.9. Generation of MCOT 1.0

During the comparison of OGS 1.0 and Cufflinks 3.0 translations (*Section 3.5*), we found that 4,855 B/W proteins in OGS 1.0 were not properly modeled by Cufflinks, possibly due to the limitation of detection sensitivity or scope. However, after these sequences were used as queries to search the *de novo* datasets with the length-based method, only 2,230 had B/W matches in both Trinity 4.0 and Oases 4.0 translations; the other ones were P/N/O. Because some of the P/N/O proteins were detected in the Cufflinks transcripts by TBLASTN, we realized that, due to its settings, TransDecoder filtered out 2,625 proteins, accounting for 4.94% of the 53,102 Cufflinks 3.0 proteins. These 4,855 B/W proteins in OGS 1.0 were compared with translations of Trinity 4.0 and Oases 4.0 and model selection was performed as per the Cufflinks 3.0 translations.

After comparing Trinity 4.0 and Oases 4.0 translations with Cufflinks 3.0 translations (*Section 3.8*), we selected the best model for each of the 37,316 Cufflinks proteins (COT). Pooling the 4,855 MAKER2 models (M) with B/W matches to Cufflinks resulted in 42,171 IDs, some of which were selected more than once. After removing them, we found 35,567 IDs, removed 2,036 redundant sequences, eliminated 2,763 and 764 (100% identical to a part of another after removal of 0 and 3 residues from each end, respectively), and obtained 30,004 protein sequences.

The intermediate BAM files generated by TopHat indicated that 20 to 30% of the RNA-Seq reads were not mapped to Msex 1.0 and may represent: 1) exons in the gaps, NNNs and W chromosome, 2) mitochondrial RNAs, or 3) others (*e.g*. polyA, mRNA of symbionts). To identify unmapped nuclear genes of *M. sexta*, we generated Trinity 4.0b using the unmapped reads (*Section 2.3*) and adopted relatively strict standards to scrutinize the Trinity sequences (*Section 2.8*). Of the 39,809 unique proteins (> 60 residues) translated from Trinity 4.0b,

10,534 had no match (W) with the 30,004 proteins; 212 had bad matches. In these 10,746 B/W proteins, only 1,183 (1,162 unique) had good UniProt matches. Some of the other 9,563 came from bacteria. The 1,162 proteins were combined with the 30,004 to generate MCOT 1.0. Of the 31,166 protein sequences in MCOT 1.0, 1,162 are from Trinity 4.0b, 7,118 are from Trinity 4.0, 2,559 from Oases 4.0, 3,715 from OGS 1.0 and 16,612 from Cufflinks 3.0. 31% of those proteins from Trinity or Oases were updated from their original versions in Cufflinks 3.0 translations. 3.7% were newly added genes in unsequenced genome regions including the W chromosome.

3.10. Comparison of MCOT 1.0 with OGS 2.0

There are 31,166 protein sequences in MCOT 1.0. Since they are originally from MAKER2, Cufflinks or Trinity 4.0b models, we traced back to their gene IDs, and found 18,089 protein-coding genes gave rise to 28,449 transcripts after model selection and 31,166 proteins (Table 7). In comparison, there are 14,165 genes, 18,979 transcripts, and 20,888 proteins in OGS 2.0 (after being filtered by the same method for MCOT 1.0). There are 21.7% fewer genes, 29.7% fewer transcripts and 33.0% fewer proteins in OGS 2.0 compared to MCOT 1.0 after counting genes, transcripts and proteins with exactly the same standard as MCOT 1.0. We then used the protein sequences in MCOT 1.0 as queries to search OGS 2.0 using BLASTP and the modified scoring matrix. The results showed 8,034 P, 2,178 N, 11,747 O, 996 B, and 8,211 W, indicating that 32.8% were P/N, 37.7% were O, and 29.5% were B/W (Table 8). The differences are substantial between the two assemblies. MCOT 1.0 is more inclusive than OGS 2.0 in terms of covering proteins. To facilitate the usage of MCOT 1.0 for proteomic studies, we have developed a naming system, which provides information of their sources, identification, matching qualities, and reference to OGS 2.0 (Fig. 8).

To estimate gene modeling quality, we use BUSCO to check the quality of all gene models used discussed above (Simão et al., 2015). With a single-copy orthologous gene database, BUSCO can assess completeness of genome assembly and modeled gene set. The result shows that MCOT 1.0 outperforms all other programs, with 93% complete genes modeled whereas OGS 2.0 only have 84% (Table 9).

3.11. Additional information from Cufflinks 3.0

A major part of MCOT 1.0 is refined from Cufflinks 3.0 models which includes 36,027 genes and 62,497 transcripts (Table 3). Using Transdecoder, we found 20,289 of the Cufflinks genes were not translated to proteins (based on the definition in *Section 2.4*), suggesting that most of them are noncoding. While 22,615 of the Cufflinks genes are absent in MCOT 1.0, the difference of 2,326 indicated that some of them may have been correctly merged during MCOT 1.0 generation. Of the 20,289 noncoding genes, the most complex gene (4,000 bp in length, 71.5% of A/T) have 33 alternative splicing forms and could be a long, noncoding RNA. Length distributions of the coding and noncoding transcripts in Cufflinks 3.0 (Fig. 9) were strikingly different: the coding ones are a lot longer. Surprisingly, 4,144 noncoding genes are 2,049–8,192 bp and 183 are > 8,193 bp. While MCOT 1.0 focuses on structural genes, the non-coding genes are another world to explore in the future.

3.12. Summary

We developed an integrated approach to select the best models based on BLASTP comparison of the Cufflinks dataset with sequences in OGS 1.0 and the *de novo* assemblies. The modified scoring matrix greatly simplified the sequence comparison by keeping pairs with >98% identity. Correlated by Cufflinks IDs, the models in different assembles (Trinity 4.0, Oases

4.0, OGS 1.0, and UniProt) were compared and chosen based on length-derived parameters. By incorporating unique sequences in OGS 1.0 and unmapped genes in the Trinity 4.0b, we generated MCOT 1.0, which has 60% more proteins than OGS 2.0. As extensive RNA-Seq data are available for most genome projects nowadays, automation of our procedures will produce comprehensive models of protein-coding genes in the future.

Acknowledgements

This study was supported by NIH grant GM58634. We thank Drs. Ulrich Melcher and Jamie Walters for their critical comments of the manuscript. The *Manduca* Genome Project, which provided Msex 1.0, OGS 1.0, OGS 2.0, Cufflinks 1.0, and RNA-Seq datasets, was funded by DARPA (Gary Blissard, Boyce Thompson Institute) and NIH grant GM41247 (Michael Kanost, Kansas State University). This work was approved for publication by the Director of Oklahoma Agricultural Experimental Station, and supported in part under project OKLO2450 (to H. Jiang). Computation for this project was performed at OSU High Performance Computing Center supported in part through NSF grant OCI-1126330.

Tables

Table 1. Comparison of the four gene prediction programs

Program	Algorithm	Advantages	Disadvantages
Cufflinks	map reads to the reference genome with TopHat and Bowtie to identify splice sites, and then use outputs of TopHat to create gene models	most sensitive; accurate splicing sites; GTF file for gene annotation; fast, less computation; more tolerant to low quality reads	carry errors in the genome scaffolds (gaps, NNNs, misassembling, etc.); many isoforms from closely located and related genes do not exist
Maker2	align EST and protein sequences to genome to produce *ab initio* gene predictions and can use RNA-Seq data to improve the prediction.	less redundant; model genes poorly represented in the RNA-Seq datasets; GTF file for gene annotation	low quality of predictions, such as extra or skipped exons, inaccurate splicing junctions, and merging of adjacent genes; biased on proteins
Trinity	De novo assemble transcripts using RNA-Seq data	not influenced by problems in the genome assembly	single hash level (*k*: 25); less sensitive than Cufflinks; redundant transcripts; no GTF file; SNPs etc.
Oases	De novo assemble transcripts using RNA-Seq data, and use Velvet for contig assembling	accurate, not influenced by problems in the genome assembly, multiple hash levels to improve quality of transcript assembly	less sensitive than Cufflinks, redundant transcripts; intense computation for large datasets; no GTF file; SNPs and other variations

Table 2. Summary statistics of *M. sexta* scaffolds in Msex 1.0 (data from Kanost et al., 2016)

size range	number	% of total number	length	% of total length	NNN number	NNN length	% of NNN length
<1 x10^3	10,543	50.5	7,516,906	1.8	13	13	0.00
10^3–10^4	8,572	41.0	16,986,901	4.1	340	551,049	3.24
10^4–10^5	1,083	5.2	40,475,711	9.6	3,568	4,970,857	12.28
10^5–10^6	604	2.9	209,932,343	50.0	9,576	10,185,979	4.85
$>10^6$	89	0.4	144,530,018	34.5	4,188	4,061,178	2.80
total	20,891	100	419,441,879	100	17,685	19,769,076	4.71

Table 3. Numbers of genes, transcripts, and proteins predicted by different programs

program	assembly	genes	transcripts	proteins	unique proteins
MAKER2	OGS 1.0	18,750	20,317	22,310	22,310
Cufflinks	Cufflinks 3.0	36,027	62,497	53,102	37,316
Trinity	Trinity 4.0	193,161	317,062	155,825	57,593
Oases	Oases 4.0	88,397	552,733	304,367	130,474

Table 4. Distribution of numbers of matched proteins over sequence identity in the BLASTP comparison of the protein sequences in OGS 1.0 and Cufflinks 3.0

Identity (%)	Count	% of total counts
96-97	1	0.01
97-98	8	0.04
98-99	58	0.32
99-100	168	0.94
100	17,672	98.69

Table 5. BLASTP comparison of OGS 1.0 and Cufflinks 3.0 models

category	count	% of total counts
P (perfect)	6,481	29.05
N (near perfect)	2,245	10.06
O (okay)	8,729	39.13
B (bad)	678	3.04
W (worst)	4,177	18.72
total	22,310	100

Table 6. BLASTP comparison of Cufflinks 3.0, Trinity 4.0, and Oases 4.0 models

Cufflinks		**Oases**				
		P	N	O	B	W
Trinity	P	5,516	407	3,490	178	228
	N	203	968	1,511	39	22
	O	1,592	824	18,509	796	361
	B	22	6	213	325	89
	W	151	21	315	146	1,384

Table 7. Summary statistics of MCOT 1.0 and OGS 2.0

	MCOT 1.0	OGS 2.0
gene #	18,089	14,165
transcript #	28,449	18,979
final protein #	31,166	20,888

Table 8. Comparison of MCOT 1.0 and OGS 2.0

Query to Subject	P	N	O	B	W
MCOT1.0 to OGS2.0	8,034	2,178	11,747	996	8,211

Table 9. BUSCO estimation of different gene models

Gene set	**Size**	**BUSCO assessment results**
OGS 1.0	20,137	C:82%[D:10%],F:9.0%,M:8.4%
OGS 2.0	27,404	C:84%[D:37%],F:8.5%,M:7.1%
CufflinksV3	53,102	C:76%[D:47%],F:8.7%,M:14%
TrinityV4	155,825	C:90%[D:51%],F:4.6%,M:4.3%
OasesV4	304,367	C:78%[D:72%],F:13%,M:7.9%
TrinityW	43,871	C:54%[D:22%],F:21%,M:24%
MCOT 1.0	31,166	C:93%[D:32%],F:3.4%,M:2.5%

*C, complete. D, duplicated. F, fragmented. M, missing.

Table S1. Datasets generated or used in this study and their descriptions

Name of Dataset	Description
Msex 1.0	Manduca Genome Assembly 1.0 generated by Newbler with Atlas-GapFill
Cufflinks 3.0	RNA-Seq reads aligned to the genome by TopHat; genes modeled by Cufflinks
Trinity 4.0	RNA-Seq reads normalized and assembled by Trinity
Trinity 4.0b	Trinity assembly of RNA-Seq reads not aligned to the genome
Oases 4.0	RNA-Seq reads normalized and then assembled by Oases with four different hash lengths
OGS 1.0	Genes modeled based on the genome sequence by MAKER2
OGS 2.0	OGS 1.0 improved by manual annotation and PASA2 using Cufflinks and de novo assemblies
Uniprot Arthropoda	Downloaded from Uniprot database on April 15, 2014
MCOT 1.0	Assembled using Msex 1.0, Cufflinks 3.0, Trinity 4.0 & 4.0b, Oases 4.0, and Uniprot Arthropoda

Table S2. Transcript length distribution of different modeling programs

size range (bp)	frequency				percentage			
	MAKER2	Cufflinks	Trinity	Oases	MAKER2	Cufflinks	Trinity	Oases
<128	64	1,391	0	151	0.32	2.23	0.00	0.00
128–256	674	2,365	64,982	45,369	3.35	3.78	20.50	8.21
257–512	2,154	5,162	88,035	121,868	10.70	8.26	27.77	22.05
513–1,024	3,897	10,476	54,326	131,931	19.35	16.76	17.13	23.87
1,025–2,048	4,593	13,572	39,500	118,958	22.81	21.72	12.46	21.52
2,049–4,096	4,566	13,990	35,656	88,667	22.67	22.39	11.25	16.04
4,097–8,192	3,263	11,530	27,919	38,658	16.20	18.45	8.81	6.99
8,193–16,384	850	3,547	6,450	6,481	4.22	5.68	2.03	1.17
>16,385	76	464	193	650	0.38	0.74	0.06	0.12
total	20,317	62,497	317,062	552,733	100	100	100	100

Table S3. Unique protein length distribution of different modeling programs

size range (aa)	frequency				percentage			
	MAKER2	Cufflinks	Trinity	Oases	MAKER2	Cufflinks	Trinity	Oases
<64	580	869	1,740	4,737	2.60	2.33	3.02	3.63
64–128	4,825	6,175	13,293	39,111	21.63	16.55	23.08	29.98
129–256	5,201	7,026	11,437	35,350	23.31	18.83	19.86	27.09
257–512	6,460	11,859	16,816	32,388	28.96	31.78	29.20	24.82
513–1,024	3,771	7,820	10,255	14,580	16.90	20.96	17.81	11.17
1,025–2,048	1,212	2,711	3,362	3,729	5.43	7.27	5.84	2.86
2,049–4,096	232	626	592	515	1.04	1.68	1.03	0.39
4,097–8,192	25	159	83	52	0.11	0.43	0.14	0.04
8,193–16,384	3	61	15	12	0.01	0.14	0.03	0.00
16,385–32,768	1	10	0	0	0.00	0.03	0.00	0.00
total	22,310	37,316	57,593	130,474	100	100	100	100

Table S4. Length distribution of the Cufflinks 3.0 proteins with P/N/O or B/W matches

length	count		percentage	
	B/W	P/N/O	B/W	P/N/O
<64	293	94	15.07	1.33
64–128	1,042	802	53.60	11.31
129–256	258	1,266	13.27	17.85
257–512	296	2,796	15.23	39.41
513–1024	40	1,708	2.06	24.08
1,025–2,048	15	382	0.77	5.38
2,049–4,096	0	40	0	0.56
4,097–8,192	0	6	0	0.08

Table S5. Length distribution of Cufflinks 3.0 proteins with different numbers of hits in UniProt

length	match number					
	0	1	2	3	4	5
<64	524	53	44	28	21	199
64–128	2,625	448	473	309	326	1,994
129–256	1,174	636	838	523	661	3,194
257–512	1,590	1,092	1,259	806	1,234	5,878
513–1,024	827	945	1,061	561	874	3,552
1,025–2,048	223	334	506	238	270	1,140
2,049–4,096	36	106	89	63	97	235
>4,096	4	101	33	37	26	29
total	7,003	3,715	4,303	2,565	3,509	16,221

Figures

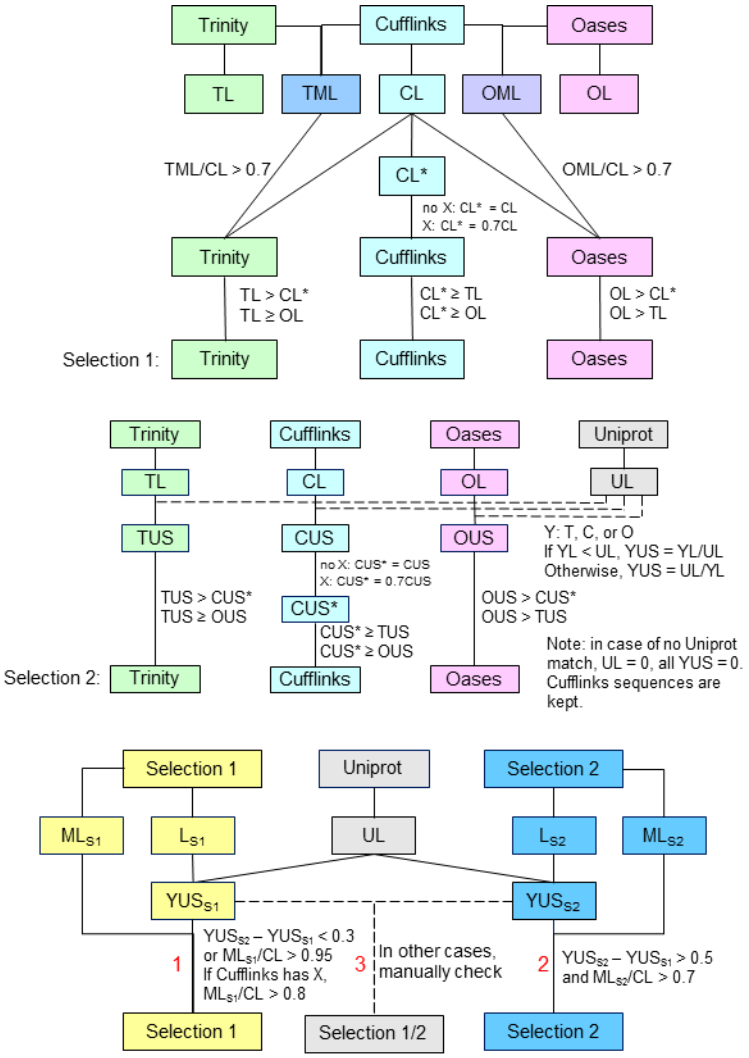


Fig. 1. Scheme of sequence comparison and selection. **A**) The length-based comparison of Cufflinks (C), Trinity (T), and Oases (O) protein sequences to generate Selection 1 (S1). **B**) The ratio-based comparisons (C-U, T-U and O-U) to generate Selection 2 (S2). **C**) Cross-examination of S1 and S2 to generate COT, a major component of MCOT 1.0.

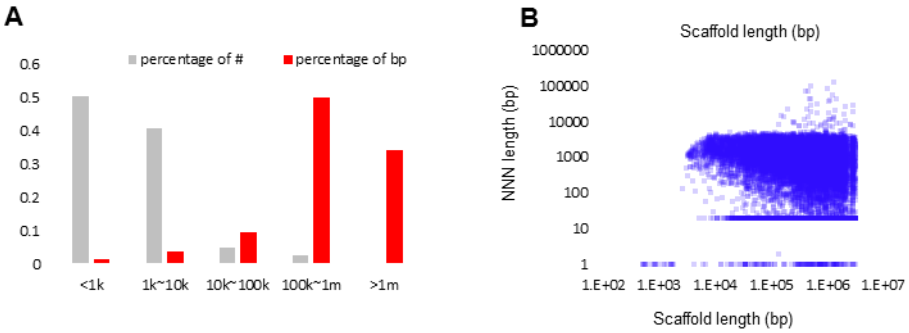


Fig. 2. Length distributions of Scaffolds and NNN regions. **A**) Percentage of scaffold numbers and sizes; **B**) lengths of NNN regions and corresponding scaffolds.

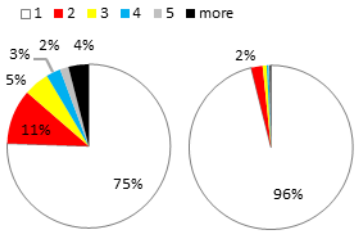


Fig. 3. Percentages of genes with 1, 2, 3, 4, 5, or ≥ 6 splicing forms based on Cufflinks 3.0 (*left*) and MAKER2-generated OGS 1.0 (*right*)

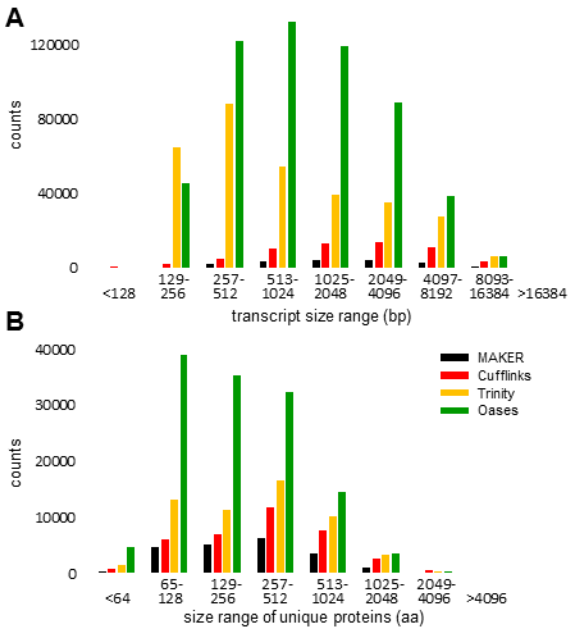


Fig. 4. Size distributions of transcripts (**A**) and unique proteins (**B**) predicted by the four programs.

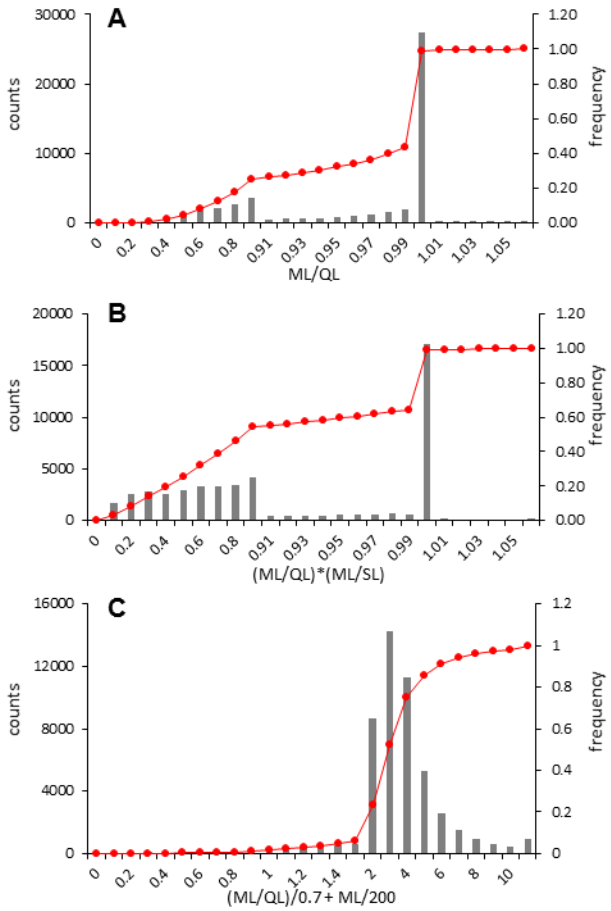


Fig. 5. Distributions of ML/QL (**A**), (ML/QL)×(ML/SL) (**B**), and (ML/QL)/0.7 + ML/200 (**C**) values from Cufflinks-Oases comparison.

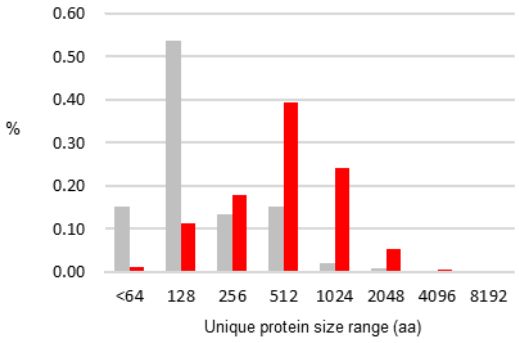


Fig. 6. Size distributions of unique Cufflinks proteins in the P/N/O (red) and B/W (gray) categories after comparison with the *de novo* assemblies.

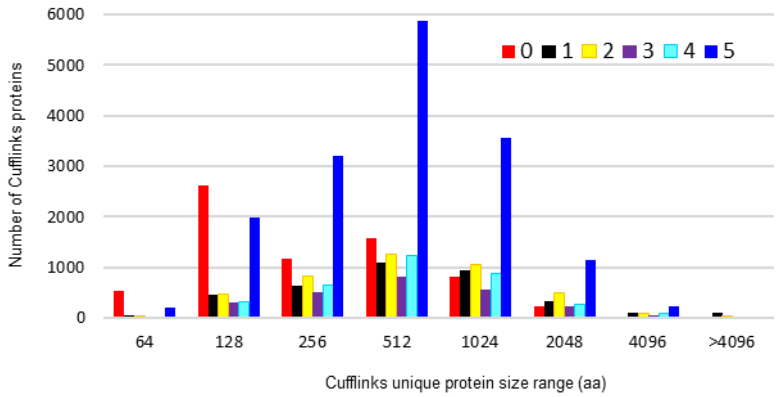


Fig. 7. Size distributions of unique Cufflinks proteins with 0, 1, 2, 3, 4, and ≥ 5 UniProt hits.

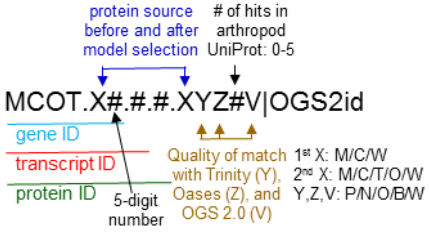


Fig. 8. Naming of MCOT 1.0 sequences. In gene "MCOT.X#", X stands for "M" (MAKER2), "C" (Cufflinks 3.0) or "W" (Trinity 4.0b) to indicate its original source (before BLAST search and model selection), and # is the 5-digit ID (*e.g.* 02367). Transcripts are named "MCOT.X#.#", where the second # (1, 2 …) stands for the 1st/2nd/… transcript from the same gene. Likewise proteins are named "MCOT.X#.#.#.XYZ#V", where the third # represents the 1st/2nd/… protein from the same transcript. If one gene generates one transcript and then one protein, the second and third #'s are marked as "0". Multicistronic genes are rare, but do exist in insects. The 2nd X indicates the final sequence source of "M", "C", "T (Trinity 4.0)", "O" (Oases 4.0), or "W" (unmapped, including those on the W chromosome). Y and Z are the quality of matching with Trinity 4.0 and Oases 4.0, respectively: "P" (perfect), "N" (near perfect), "O" (okay), "B" (bad), "W" (worst), or "X" (data unavailable). The fourth # is the number of kept UniProt hits (0 to 5 or X for data unavailable). V marks the quality of matching with OGS 2.0: if "P", "N" or "O", the corresponding OGS 2.0 ID is added next to "|"; otherwise "X" is added to indicate no good match.

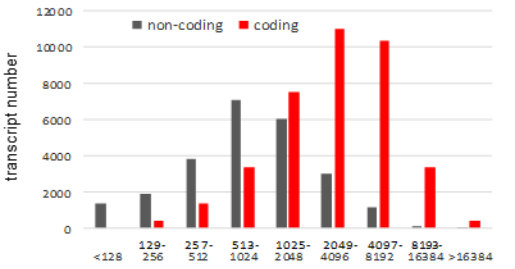


Fig. 9. Size distributions of the coding and noncoding transcripts in Cufflinks 3.0.

References

Arrese, E.L., Soulages, J.L., 2010. Insect fat body: energy, metabolism, and regulation. Ann Rev Entomol. 55, 207–225.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., 2009. BLAST+: architecture and applications. BMC Bioinformatics. 10, 421.

Cao, X., He, Y., Hu, Y., Zhang, X., Wang, Y., Zou, Z., Chen, Y., Blissard, G.W., Kanost, M.R., Jiang, H., 2015. Sequence conservation, phylogenetic relationships, and expression profiles of nondigestive serine proteases and serine protease homologs in Manduca sexta. in press.

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

Grosse-Wilde, E., Kuebler, L.S., Bucks, S., Vogel, H., Wicher, D., Hansson, B.S., 2011. Antennal transcriptome of Manduca sexta. Proc Natl Acad Sci USA. 108, 7449–7454.

Gunaratna, R. Jiang, H., 2013. A comprehensive analysis of Manduca sexta immunotranscriptome. Dev Com Immunol. 39, 388–398.

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., Macmanes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., Leduc, R.D., Friedman, N., Regev, A., 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 8, 1494–1512.

Hopkins, T., Krchma, L., Ahmad, S., Kramer, K., 2000. Pupal cuticle proteins of Manduca sexta: characterization and profiles during sclerotization. Insect Biochem Mol Biol. 30, 19–27.

Jiang, H., Vilcinskas, A., Kanost, M.R., 2010. Immunity in lepidopteran insects. In "Invertebrate Immunity" (K. Söderhäll ed.), Adv Exp Med Biol. 708, 181–204.

Kanost, M.R., Kawooya, J.K., Law, J.H., Ryan, R.O., Van Heusden, M.C., Ziegler, R., 1990. Insect hemolymph proteins. Adv Insect Physiol. 22, 299–396.

Langmead, B, Salzberg, SL, 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods. 9, 357–359.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25, 2078–2079.

Pauchet, Y., Wilkinson, P., Vogel, H., Nelson, D.R., Reynolds, S.E., Heckel, D.G., ffrench-Constant, R.H., 2010. Pyrosequencing the Manduca sexta larval midgut transcriptome: messages for digestion, detoxification and defence. Insect Mol Biol. 19, 61–75.

Riddiford, L., Hiruma, K., Zhou, X., Nelson, C.A., 2003. Insights into the molecular basis of the hormonal control of molting and metamorphosis from Manduca sexta and Drosophila melanogaster. Insect Biochem Mol Biol. 33, 1327–1338.

Roberts, A., Trapnell, C., Donaghey, J., Rinn, J.L., Pachter, L., 2011. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol. 12, R22.

Schulz, M.H., Zerbino, D.R., Vingron, M., Birney, E., 2012. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 28, 1086–1092.

Shields, V., Hildebrand, J.G., 2001. Recent advances in insect olfaction, specifically regarding the morphology and sensory physiology of antennal sensilla of the female sphinx moth Manduca sexta. Microsc Res Tech. 55, 307–329.

Trapnell, C, Pachter, L, Salzberg, SL, 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 25, 1105–1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., Pachter, L., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 7, 562–578.

Kanost, M.R., Arrese, E.L., Cao, X., Chen, Y.-R.R., Chellapilla, S., Goldsmith, M.R., Grosse-Wilde, E., Heckel, D.G., Herndon, N., Jiang, H., Papanicolaou, A., Qu, J., Soulages, J.L., Vogel, H., Walters, J., Waterhouse, R.M., Ahn, S.-J.J., Almeida, F.C., An, C., Aqrawi, P., Bretschneider, A., Bryant, W.B., Bucks, S., Chao, H., Chevignon, G., Christen, J.M., Clarke, D.F., Dittmer, N.T., Ferguson, L.C., Garavelou, S., Gordon, K.H., Gunaratna, R.T., Han, Y., Hauser, F., He, Y., Heidel-Fischer, H., Hirsh, A., Hu, Y., Jiang, H., Kalra, D., Klinner, C., König, C., Kovar, C., Kroll, A.R., Kuwar, S.S., Lee, S.L., Lehman, R., Li, K., Li, Z., Liang, H., Lovelace, S., Lu, Z., Mansfield, J.H., McCulloch, K.J., Mathew, T., Morton, B., Muzny, D.M., Neunemann, D., Ongeri, F., Pauchet, Y., Pu, L.-L.L., Pyrousis, I., Rao, X.-J.J., Redding, A., Roesel, C., Sanchez-Gracia, A., Schaack, S., Shukla, A., Tetreau, G., Wang, Y., Xiong, G.-H.H., Traut, W., Walsh, T.K., Worley, K.C., Wu, D., Wu, W., Wu, Y.-Q.Q., Zhang, X., Zou, Z., Zucker, H., Briscoe, A.D., Burmester, T., Clem, R.J., Feyereisen, R., Grimmelikhuijzen, C.J., Hamodrakas, S.J., Hansson, B.S., Huguet, E., Jermiin, L.S., Lan, Q., Lehman, H.K., Lorenzen, M., Merzendorfer, H., Michalopoulos, I., Morton, D.B., Muthukrishnan, S., Oakeshott, J.G., Palmer, W., Park, Y., Passarelli, A.L., 2016. Multifaceted biological insights from a draft genome sequence of the tobacco hornworm moth, Manduca sexta. Insect biochemistry and molecular biology.

Yandell, M., Ence, D., 2012. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 13, 329–342.

Zhang, S., Gunaratna, R.T., Zhang, X., Najar, F., Wang, Y., Roe, B., Jiang, H., 2011. Pyrosequencing-based expression profiling and identification of differentially regulated genes from Manduca sexta, a lepidopteran model insect. Insect Biochem Mol Biol. 41, 733–746.

Zou, Z., Najar, F., Wang, Y., Roe, B., Jiang, H., 2008. Pyrosequence analysis of expressed sequence tags for Manduca sexta hemolymph proteins involved in immune responses. Insect Biochem Mol Biol. 38, 677–682.

CHAPTER II

THE IMMUNE SIGNALING PATHWAYS OF *MANDUCA SEXTA*

Xiaolong Cao[1], Yan He[2], Yingxia Hu[1], Yang Wang[2], Yun-Ru Chen[3],

Bart Bryant[4], Rollie J. Clem[4], Lawrence M. Schwartz[5], Gary Blissard[3], Haobo Jiang[2]

[1] Department of Biochemistry and Molecular Biology, Oklahoma State University, Stillwater, OK 74078, USA

[2] Department of Entomology and Plant Pathology, Oklahoma State University, Stillwater, OK 74078, USA

[3] Boyce Thompson Institute, Cornell University, Ithaca, NY 14853, USA

[4] Division of Biology, Kansas State University, Manhattan, KS 66506

[5] Department of Biology, University of Massachusetts, Amherst, MA 01003

Key words: insect immunity, RNA-Seq, transcriptome, gene annotation, expression profiling.

Abbreviations: Atg, autophagy-related protein; aPKC, atypical protein kinase C; CF, control fat body; CH, control hemocytes; IF, induced fat body; IH, induced hemocytes; Deaf, deformed epidermal autoregulatory factor; Dnr1, defense repressor-1; Dsp1; Dorsal switch protein-1; ECSIT, evolutionarily conserved intermediate in Toll pathways; FPKM, fragments per kilobase of exon per million fragments mapped; GPRK, G protein-coupled receptor kinase; IAP, inhibitor of apoptosis; Imd, immunodeficiency; IKK, IκB kinase; JNK, Jun N-terminal kinase; Jra, Jun-related antigen; MAPK, mitogen-activated protein kinase; MASK, multiple ankyrin repeats single KH domain; ML, MD2-like; MLK, mixed-linage kinase; NFκB and IκB, nuclear factor-κB and its inhibitor; NTF, nuclear transport factor; PIAS, protein inhibitor of activated STAT; PIRK, poor Imd response upon knock-in; PVF, platelet-derived and vascular endothelial growth factor; PVR, PDGF/VEGF receptor; STAT, signal transducer and activator of transcription; PG and PGRP, peptidoglycan and peptidoglycan recognition protein; SUMO, small ubiquitin-like modifier; TAK, transforming growth factor β activated kinase; TAMP, Toll activation mediating protein; TIR, Toll/interleukin-1 receptor; TRAF, tumor necrosis factor receptor-associated factor.

Abstract

Signal transduction pathways and their coordination are critically important for proper functioning of animal immune systems. Our knowledge of the constituents of the intracellular signaling network in insects mainly comes from genetic analyses in *Drosophila melanogaster*. To facilitate future studies of similar systems in the tobacco

hornworm and other lepidopteran insects, we have identified and examined the homologous genes in the genome of *Manduca sexta*. Based on 1:1 orthologous relationship in most cases, we hypothesize that the Toll, Imd, MAPK-JNK-p38 and JAK-STAT pathways are intact and operative in this species, as are most of the regulatory mechanisms. Similarly, cellular processes such as autophagy, apoptosis and RNA interference probably function in similar ways, because their mediators and modulators are mostly conserved in this lepidopteran species. We have annotated a total of 186 genes encoding 199 proteins, studied their domain structures and evolution, and examined their mRNA levels in tissues at different life stages. Such information provides a genomic perspective of the intricate signaling system in a non-drosophiline insect.

1. **Introduction**

Insects fight against invading pathogens and parasites via their innate immune system (Gillespie et al., 1997; Lemaitre and Hoffmann, 2007). Like other physiological processes, insect immune responses involve sensors, effectors, and signal transducers, linking pathogen recognition with cellular and humoral responses. Some of the responses occur in minutes while others involve transcriptional activation of genes that are not highly expressed under normal conditions, and thus may provide responses in hours to days. In the latter case, a relay system must exist to transduce the extracellular signals of wounding or invasion into the nuclei of cells, where transcriptional regulation occurs. If pathogens are sensed by receptors (*e.g*. PGRP-LC) on the cell surface, responses are more direct than if recognition occurs in hemolymph. In the latter scenario, receptors (*e.g*. PGRP-SA) in hemolymph bind to the pathogens and initiate extracellular signal transduction to generate

active cytokines. The cytokines then interact with their receptors on the cell surface to induce cellular responses including phagocytosis, encapsulation, apoptosis, autophagy, and synthesis of immune effectors (Strand, 2008; Jiang et al., 2010). Consequently, the intracellular signal transduction network is essential for mediating immune responses in insects.

Receptor-mediated Toll, Imd, MAPK-JNK-p38, JAK-STAT and other pathways are widely conserved in metazoans, functioning as regulators and mediators of humoral and cellular immune responses (Buchon et al., 2014). Extensive studies in *Drosophila melanogaster* have revealed many details of the signal transduction network. The Toll pathway was discovered in the screens that identified mutations in genes affecting the establishment of dorsoventral axis and later found to regulate the expression of immunity-related genes through Dorsal and Dif, transcription activators of the Rel family (Valanne et al., 2011). Gram-positive bacteria and fungi trigger this pathway via an extracellular serine protease cascade that activates the cytokine Spätzle through limited proteolysis. This activated cytokine binds to the Toll receptor, leading to antimicrobial peptide synthesis and differentiation of certain hemocytes into lamellocytes. These lamellocytes are capable of encapsulating and killing parasites such as parasitoid wasps (Sorrentino et al., 2004). In the case of Gram-negative bacteria, DAP-type peptidoglycans (PGs) elicit the Imd pathway via transmembrane PGRP-LC and intracellular signal mediators (Kaneko et al., 2006; Rämet et al., 2002). Activated Relish, another Rel factor, then migrates into the nucleus to turn on a set of immunity-related genes overlapping with that induced by Dorsal and Dif (Imler and Hoffmann, 2001; Mellroth et al., 2005). Cytokines, growth factors, or stress signals stimulate the MAPK-JNK-p38 pathway to regulate apoptosis, Imd pathway, and

cell differentiation (Ragab et al., 2011; Chen et al., 2010; Dong et al., 2002). The JAK-STAT pathway, RNA interference, autophagy and other defense mechanisms are involved in antiviral responses (Kisseleva et al., 2002; Baeg et al., 2005; Kingsolver et al., 2013). Based on the available information, these pathways are mostly conserved among insects but differences do exist. For instance, the honeybee *Apis mellifera* has considerably fewer immunity-related genes (Evans et al., 2006). *A. mellifera* has five Toll genes compared with the nine found in *D. melanogaster*. The pea aphid *Acyrthosiphon pisum* lacks the entire Imd pathway (Gerardo et al., 2010). With such plasticity observed among the few genomes available in the Insecta, it is therefore critically important to examine and characterize the immune signaling components in different major orders of insects.

Lepidoptera comprises about 160,000 described species of moths and butterflies in 126 families and 46 superfamilies (Kristensen et al., 2007). Larvae of many lepidopterans are serious agricultural pests but they are susceptible and can be controlled by biological agents such as entomopathogens (*e.g*. viruses, bacteria, fungi) and parasitoid wasps. Studies of lepidopteran immune systems and the associated signaling pathways are extremely important for developing effective biological control methods. *Manduca sexta* and *Bombyx mori* have been used as powerful biochemical models to explore various aspects of innate immunity (Jiang et al., 2010). Immunity-related genes in the silkworm were previously compared with those in *D. melanogaster*, *Anopheles gambiae* and *A. mellifera* (Tanaka et al., 2008) and analyses of the *M. sexta* hemocyte and fat body transcriptomes revealed a set of 232 genes encoding proteins for pathogen recognition, signal transduction, microbe killing (Gunaratna and Jiang, 2013), and modulation of mRNA levels in response to an immune challenge (Zhang et al., 2011). Recently, the *M. sexta* genome assembly became

available along with 52 RNA-Seq datasets of tissues at various life stages (X et al., 2015). To better understand immune signal transduction in this undomesticated pest species, we have annotated genes for the putative pathway members, studied their expression patterns, and proposed a signal transduction network based on 1:1 orthology. The results represent working models for future studies on *M. sexta* and other lepidopteran pests.

2. Materials and methods

2.1. Gene identification, sequence improvement, and feature prediction

Manduca Genome Assembly 1.0 and gene models in *Manduca* Official Gene Sets (OGS) 1.0 and 2.0 and Cufflinks Assembly 1.0 (X et al., 2015) were downloaded from *Manduca* Base (ftp://ftp.bioinformatics.ksu.edu/pub/Manduca/). Protein sequences of the putative signal transducers from *M. sexta* (Gunaratna and Jiang, 2013) and other insects were used as queries to search Cufflinks 1.0, OGS 1.0 and OGS 2.0 using the TBLASTN algorithm (http://darwin.biochem.okstate.edu/blast/blast_links.html). Hits with aligned regions longer than 30 residues and identity over 40% were retained for retrieving corresponding cDNA sequences. Errors resulting from problematic regions (*e.g*. NNN…) in the genome assembly were manually corrected after BLASTN search of *Manduca* Oases and Trinity Assemblies 3.0 of the RNA-Seq data (Cao and Jiang, 2015). The two genome-independent RNA-Seq assemblies were developed to cross gaps between genome scaffolds or contigs and detect errors in the gene models. In some complex cases, all exons of a gene were examined based on the GT-AG rule and sequence alignment to identify the splicing junctions. Correct open reading frames in the improved sequences were identified using ORF Finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html) and validated by BLASTP

search against GenBank (http://www.ncbi.nlm.nih.gov/) or Uniprot (http://www.uniprot.org/). Signal peptides were predicted using SignalP4.1 (Petersen et al., 2011). Conserved domain structures were identified using SMART (http://smart.embl-heidelberg.de/smart/set_mode.cgi) and InterProScan (http://www.ebi.ac.uk/Tools/pfa/iprscan/).

2.2. Sequence alignment and phylogenetic analysis

Multiple sequence alignments of immune signal transducers from *M. sexta* and other insects were performed using MUSCLE, a module of MEGA 6.0 (Tamura et al., 2013) at the following settings: refining alignment, gap opening penalty (–2.9), gap extension penalty (0), hydrophobicity multiplier (1.2), maximal iterations (100), UPGMB clustering (for iterations 1 and 2) and maximum diagonal length (24). The aligned sequences were used to construct neighbor-joining trees by MEGA 6.0 with bootstrap method for the phylogeny test (1000 replications, Poisson model, uniform rates, and complete deletion of gaps or missing data).

2.3. Gene expression profiling

Coding DNA sequences from the improved gene models were retrieved and employed as templates for mapping reads in the 52 *M. sexta* RNA-Seq datasets, representing mRNA samples from whole insects, organs or tissues at various developmental stages. Illumina reads (*M. sexta* genome and transcriptome project; http://www.ncbi.nlm.nih.gov/bioproject/PRJNA81039) were trimmed to 50 bp and mapped to the coding regions using Bowtie (0.12.8) (Langmead et al, 2009). Numbers of the mapped reads were used to calculate FPKM (**f**ragments **p**er **k**ilobase of exon per **m**illion fragments mapped) values using RSEM (Li and Dewey, 2011). Hierarchical clustering of the $\log_2$(FPKM+1) values was performed

using MultiExperiment Viewer (v4.9) (http://www.tm4.org/mev.html) with the Pearson correlation-based metric and average linkage clustering method. To study transcript level changes after immune challenge, the entire CDS set was used to search for corresponding contigs in the CIFH09 database (http://darwin.biochem.okstate.edu/blast/blast_links.html) (Zhang et al., 2011) by TBLASTN. The numbers of CF, CH, IF, and IH reads (C for control, I for induced after injection of bacteria, F for fat body, H for hemocytes) assembled into these contigs were retrieved for normalization and calculation of IF/CF and IH/CH ratios. When a polypeptide sequence corresponded to two or more contigs, sums of the normalized read numbers were used to calculate its relative mRNA abundances in fat body and hemocytes (Gunaratna and Jiang, 2013).

3. Results and discussion

3.1. Spätzle-1−7, cytokines with distinct structures, functions, and expression patterns

There are seven genes encoding Spätzle-like proteins in *M. sexta* (Table S1), which differs from the number present in *Tribolium castaneum* (9), *D. melanogaster* (6), and *A. gambiae* (6), *B. mori* (3) and *A. mellifera* (2) (Tanaka et al., 2008). The *M. sexta* proteins contain a signal peptide, a 50 to 360-residue segment with 0−4 low complexity regions, and a cystine-knot cytokine domain (Fig. 1). For Spätzle-1, cleavage between QR and LG results in a dimer of the C-terminal fragment that induces antimicrobial peptide synthesis (An et al., 2010), presumably via a Toll receptor. While Spätzle-2−7 may be activated by trypsin-like serine proteases, Spätzle-3 and 5 can also be processed by furin-like enzymes next to their recognition sequences, RHAR and RPRR, respectively. The C-terminal fragments of Spätzle-3−6 contain an even number of Cys residues, which might allow them

to possibly dimerize via additional disulfide bonds. Molecular modeling suggests that Spätzle-1−5 and 7 adopt a similar fold with three pairs of antiparallel β-strands stabilized by 3 or 4 intrachain disulfide bonds (data not shown). Phylogenetic analysis of the entire proteins indicates that Spätzle-3−6 each forms a tight group with their orthologs from the other insects (Fig. 1A), suggestive of conserved functions. From parallel studies (Cao et al., 2015; Rao et al., 2015; He et al., 2015), we have noticed that the mRNA levels of many immunity-related genes in fat body and midgut greatly increase at the onset of wandering stage and peak on day 1 of the pupal stage. This infection-independent gene up-regulation during metamorphosis also occurs in other lepidopterans such as *Galleria mellonella* (Altincicek and Vilcinskas, 2008). Spätzle-1, 2 and 7 transcripts display this pattern with the highest FPKM values of 224, 760 and 564, respectively (Fig. 2A). These three genes are induced upon immune challenge, whereas Spätzle-3−6 mRNAs were detected only at very low levels (Gunaratna and Jiang, 2013; Zhang et al., 2011; Table 1). Spätzle-1B mRNA level is low in ovary, higher in eggs and down-regulated after hatching. In contrast, Spätzle-2 mRNA levels are high in ovary, lower in eggs, and become higher in 1st instar larvae. The expression patterns of Spätzle-3 and 5 are similar to each other. Spätzle-4 and 6 are almost exclusively produced in the midgut of 2nd and 3rd instar larvae. The detection of Spätzle-2, 3, 5 and 7 mRNAs in head is interesting, since *Drosophila* Toll6, Toll7 and Toll8 act as receptors of neurotrophins (*Drosophila* Spätzle-2, 3 and 5) (McIlroy et al., 2013; Ballard et al., 2014).

3.2. Structure, expression, and evolution of Toll receptors

Toll receptors are a group of transmembrane proteins with extracellular Leu-rich repeats (LRRs) and a cytoplasmic Toll/interleukin-2 receptor (TIR) homology domain

(Fig. 3A). We have identified sixteen such genes and named them Toll1−6, 7_1−3, 8, 9_1, 9_2, 10_1−3 and 12 (Table S1). These names are based on and mostly consistent with their orthologs in *B. mori* (Tanaka et al., 2008). Toll1 is reported as an immune-inducible gene that is predominantly expressed in hemocytes (Ao et al., 2008b; Gunaratna and Jiang, 2013) (Table 1). Along with Toll2−5 and *B. mori* Toll3_1−3, *M. sexta* Toll1 is grouped with *D. melanogaster* Toll1, 3−5, *A. gambiae* Toll1A, 1B, 5A and 5B, and *T. castaneum* Toll1−4 (Fig. 3B). Nonetheless, *M. sexta* Toll1, 3 and 4 have only 4 to 5 LRRs (Fig. 3B), instead of the 12 LRRs and 2 Cys-rich C-terminal domains that are present in *Drosophila* Toll1. The mRNA levels of Toll1, 3 and 4 are very low in the 52 libraries (Fig. 2A). Hence, the putative roles as Spätzle-1 or 2 receptors need validation. In contrast, Toll2 and 5 transcripts are highly abundant in fat body and their profiles of expression are closely similar to those of Dorsal, Serpent and Spätzle-1B. Interestingly, *Manduca* Dorsal and Serpent may interact with each other to activate moricin gene transcription (Rao et al., 2011). Toll2 and 5, containing a Cys-rich C-terminal domain, are more similar in domain structure to *Drosophila* Toll1. Based on this and other evidence, we suggest *M. sexta* Toll2 and 5 are better candidates than Toll1 as receptors of Spätzle-1, 2 and 7. In *D. melanogaster*, Toll6, 7 and 8 (*i.e.* Tollo) are involved in neurotrophism (McIlroy et al., 2013; Ballard et al., 2014) and recent studies suggest that Toll7 may also be a pattern recognition receptor for vesicular stomatitis virus, activating cellular autophagy of the virus (Nakamoto et al., 2012). Their orthologous genes (Toll6, 7_1−3 and 8) are expressed in heads at levels higher than other tissues (Fig. 2A) and may play similar roles in *M. sexta*. The *M. sexta* Toll9_1 mRNA levels are high in Malpighian tubules of pre-wandering larvae and adults, as well as in midgut of feeding larvae. Human myeloid differentiation factor-2

(MD2) forms a complex with Toll-like receptor-4 to recognize lipopolysaccharide and lead to inflammation and cytokine production. A MD2-like protein (ML1) from *A. gambiae* specifically regulates the resistance to *Plasmodium falciparum* (Dong et al., 2006). We have identified five MD2-like proteins (MLs) in *M. sexta*, which contain a signal peptide and may increase binding specificity of the Toll receptors (Ao et al., 2008a).

Despite the fact that the coding regions being 2.2−4.0 kb in length, half of the 16 genes (Toll6, 7_1−3, 8, 10_1−3) only contain a single exon (Fig. 3C). They correspond 1:1 with their orthologous genes on chromosome 23 in *B. mori*. *M. sexta* Toll7_1, 10_3, 10_2, 10_1 and 6 on Scaffold (S) 00066 have the same orientations as those in the silkworm, flanked by Toll7_3 (S00185), 7_2 (S00183), and 8 (S00166) (Fig. 3C). When we compared the orthologous genes in *A. gambiae*, *D. melanogaster* and *T. castaneum*, similar gene orders were found. These orthologous genes include: Toll7_3 to 1, 10_3 to 1, 6 and 8 in the lepidopterans; Toll11&10, 7, 8 and 6 in the mosquito; Toll2&7, 8 and 6 in the fruit fly; Toll6, 8, 10, and 7 in the beetle. The underlined genes result from lineage-specific gene duplications. Except for AgToll8, TcToll8 and TcToll10 (with 5, 2 and 2 exons, respectively), the remaining genes are intronless. In comparison, MsToll1−5 have 7 or 8 exons, BmToll3_1−3 have 7, 5 and 8 exons, DmToll1, 3−5 have 2 or 4 exons, AgToll1A, 1B, 5A and 5B have 3 exons, and TcToll1−5 have 3 or 4 exons. Together, these observations reveal a dramatic evolutionary history of this ancient family of genes along the lineages of holometabolous insects.

3.3. Intracellular members of the Toll pathway and their regulation

We have identified 1:1 orthologs for most of the intracellular pathway members and modulators known so far. These include MyD88, Tube, Pelle, Pellino, Cactus, G protein-

coupled receptor kinase-2 (GPRK2), Tollip-1&2, Cactin, Aos, Uba2, Smt3, Lesswright, and deformed epidermal autoregulatory factor-1 (Deaf1) (Fig. 4A, Table S1) in *Drosophila*. In the current model, activated Toll receptor associates with its adaptor MyD88 via their TIR domains. MyD88, Tube and Pelle (a kinase-like protein) form a complex via their death domains to phosphorylate Cactus. Pellino, with a RING E3 ubiquitin ligase domain, may ubiquitinate Pelle to enhance the Toll signaling. Unlike its ortholog in the fruit fly, the C-terminal Ser/Thr protein kinase domain of *M. sexta* Tube is predicted to be active catalytically and thus, actively involved in the pathway activation. The phosphorylation of Cactus by Pelle and perhaps Tube, causes it to dissociate from Dorsal or Dif become polyubiquitinated and degraded by the proteasome. Dorsal and Dif appear to be the products of a lineage-specific gene duplication (data not shown). GPRK2 may interact with Cactus to enhance signaling. Atypical protein kinase C (aPKC), together with its partners Ref2P and TRAF2 (TNF-receptor-associated factor-2), may interact with Pelle and directly activate Dorsal/Dif (Avila et al., 2002). Free, active Dorsal/Dif translocates into the nucleus to activate target gene transcription along with Deaf1 and other transcription factors (*e.g*. U-shaped and Toll activation mediating protein, TAMP). This pathway is likely regulated at other steps. For instance, Tollips may associate with the Toll receptor and suppress the kinase activity of Pelle (Zhang and Ghosh, 2002). In *D. melanogaster*, Cactin may bind Cactus to block its function and cause embryonic ventralization (Lin et al., 2000). Conjugation of Dorsal/Dif by Smt3, a small ubiquitin-like modifier (SUMO), may potentiate function of Dorsal/Dif (Bhaskar et al., 2002). Aos1 and Uba2 may form a dimer which acts as an E1 SUMO-activating enzyme (Paddibhatla et al., 2010). The Lesswright homolog of Ubc9, an E2 SUMO-conjugating enzyme, negatively

impacts the pathway (Chiu et al., 2005). The E3 SUMO ligase, Ulp1 peptidase and its helper Kurtz, reduces SUMO conjugation and response level of Dorsal/Dif-induced genes (Anjum et al., 2013).

Three Dorsal and two Dif variants are generated via alternative splicing (Table S1). The major Dorsal A is widely produced in tissues whereas B- and C-forms are preferentially expressed in fat body and head, respectively (Fig. 2A). Dif mRNA levels are lower compared to Dorsal. Like MyD88, Aos1 and Smt3, *Manduca* Tube, Pelle, Pellino, Lesswright, Uba2, Ref2Ps, aPKC-A, TRAF2, Cactus, Dorsal-A, ML2, cactin, Tollip-1 and 2 are widely expressed in all the tissues examined. However, mRNA levels of the latter genes (Tube through Tollip-2) increase considerably in fat body during the wandering stage and reach peaks in pupae at day 1. As well, most of these genes are induced by 24 h following an immune challenge (Table 1).

3.4. The Imd pathway, JNK branch, and their regulation

The Imd pathway, considered specific for Gram-negative bacteria, regulates the transcription of a set of immunity-related genes that overlaps with that controlled by the Toll pathway (Kleino and Silverman, 2014). This pathway is also branched to JNK and apoptosis (Fig. 4B). We have identified 1:1 orthologs for nearly all of the pathway components (Table S1) and, therefore, propose that the *M. sexta* Imd pathway is triggered by DAP-PG, a component of the cell wall in most Gram-negative bacteria as well as Gram-positive *Bacillus* and *Listeria* species. Since there is no PGRP-LE ortholog in the moth (Zhang et al., 2015), membrane-bound PGRP-LCa and LCb may work together to detect them. The longer splicing variant LCa contains two transmembrane domains, raising the possibility that it detects intracellular bacteria. Upon DAP-PG binding, a cytosolic portion

of these variants may interact with the adaptor Imd and then FADD through their death domains. FADD recruits Dredd, the mammalian caspase-8 homolog, which cleaves Imd and Relish (Ertürk-Hasdemir et al., 2009) or a pro-caspase that leads to apoptosis. Cleaved Imd is susceptible to ubiquitination by IAP2 (inhibitor of apoptosis-2, an E3 ubiquitin ligase), Uev1A, Ubc13/Bendless and Ubc5/Effete (E2 ubiquitin-conjugating enzymes) (Paquette et al., 2010). Following ubiquitination, Imd likely recruits transforming growth factor β-activated kinase-1 (TAK1) and its binding protein TAB2 (Aggarwal, 2003). The dimer of TAB2 and TAK1 may then phosphorylate both Kenny/IKKβ and IRD5/IKKγ/NEMO in a complex, and JNK and Basket through MKK4 or MKK7/hemipterous (Hep) (Geuking et al., 2009). JNK may activate Aop and the AP-1 complex of Jra/Jun and Fos/Kayak to regulate downstream genes (*e.g*. PIRK). The IKK complex may phosphorylate the cleaved Relish to cause chain separation. While the C-terminal ankrin repeats and death-like domain are destined to be degraded, the N-terminal fragment (Relish-N), assisted by nuclear transport factor 2 (NTF2), could translocate into the nucleus and activate expression of immunity-related genes (*e.g*. antimicrobial peptides) via its Rel homology domain.

Additional regulatory mechanisms are known for the Imd pathway in *Drosophila* (Kleino and Silverman, 2014). PIRK interferes with the association of Imd, FADD, and Dredd (Kleino et al., 2008). Dnr1 (defense repressor-1) inhibits the caspase Dredd while Sickie and Caspar have opposite effects on Dredd-induced activation of Relish (Foley and O'Farrell, 2004). USP36 deubiquitinates Imd for its degradation and, thus, represses Imd signaling (Thevenon et al., 2009). Another deubiquitinase, CYLD (for Cylindromatosis), modulates the IKK complex to control Relish phosphorylation (Tsichritzis et al., 2007).

POSH controls the complex of TAK1 and TAB2; an SCF complex of **S**kp1, **C**ullin and **F**-box protein regulates the phosphorylated Relish-N; Akirin and Relish-N co-regulate some target genes of the Imd pathway (Tsuda et al., 2005; Cardozo and Pagano, 2004; Bonnay et al., 2014).

Most genes in the putative Imd pathway are widely expressed in different tissues at various life stages (Fig. 2B). The mRNA levels of Imd, FADD, Dredd, Relish, and many other genes are considerably higher in midgut than in fat body. This is consistent with the finding that local immune response of epithelial cells is Imd pathway-dependent, as the Imd pathway is fast and can be activated within minutes following a challenge (Kleino and Silverman, 2014; Paquette et al., 2010). While mRNA levels of a few genes are higher at 24 h after the immune challenge, others are similar to or even lower than the control levels (Table 1). This contrasts drastically with most of the Toll pathway genes, whose induced expression in fat body and hemocytes lasts longer than 24 h. Consistent with their immune inducibility, most Imd pathway members are highly expressed in fat body from the pre-wandering larval stage to the early pupal stage (Fig. 2B). Their up-regulation in midgut is less pronounced and varies among the Imd pathway members during the same period, perhaps due to gut purging.

3.5. MAPK-JNK-p38 pathways

MAPK pathways are responsive to growth factors, cytokines and stress signals, and thereby regulate cell proliferation, differentiation, inflammation, and death. In *Drosophila*, components of these pathways activate MAPKs (Rolled, JNK and p38), down-regulate the Imd pathway, and stimulate hemocyte proliferation and lamellocyte formation (Fig. 4C) (Ragab et al., 2011; Dong et al., 2002; Lee and Ferrandon, 2011). We have identified

homologs of two platelet-derived and vascular endothelial growth factors (PVFs), a PDGF/VEGF receptor (PVR), two small GTPases (Ras85D and Rac1), three kinases (Polehole, Dsor1 and Rolled), and a transcription factor (Pointed) that induces PIRK (poor Imd response upon knock-in) production. By interfering with Imd-FADD-Dredd association, PIRK, a small protein with no known domain structure, may inhibit Imd signaling. JNK may be activated through an Imd branch (Fig. 4B) and perhaps also by MLK1, MKK4, PVR or Alk (PVR and Alk are receptors with a Ser/Thr kinase domain). We have also found putative members of the cytokine-triggered MAPK pathway, namely Eiger, Wengen, TRAF1, and Misshapen that may recruit and sequentially activate TAK1-TAB2 dimer, MKK7/hemipterous, and JNK (Liu et al., 1999; Geuking et al., 2009). A protein called ECSIT (evolutionarily conserved intermediate in Toll pathways) is linked to the Toll receptor through TRAF2, and may activate a kinase cascade of MEKK1, MKK3, and p38 to induce the formation of the AP-1 complex (Kopp et al., 1999). In addition, Spitz and Vein may induce MAPK signaling in the presence of reactive oxygen species but their receptors are unknown in *Manduca*.

Certain members of the putative MAPK-JNK-p38 pathways (*i.e*. Eiger, Rac1, MASK, Rolled, JNK, p38, Aop, Jra, Fos) in *M. sexta* are transcriptionally activated in larval fat body or hemocytes after an immune challenge (Gunaratna and Jiang, 2013) (Table 1). In addition to these, PVF2, PVR, Wengen, Ras85D, Cdc42, Dsor1, Misshapen, MLK and Pointed show mRNA level increases in fat body from pre-wandering to early pupal stage (Fig. 2C). Transcript levels for most of these genes in midgut are similar to or higher than those in fat body. Levels of PVR, Rac1, Misshapen-B&C, p38B, Ras85D, Jra and Fos mRNAs reach peak levels during pupation. Expression of the pathway members in head,

muscles, Malpighian tubules, testis, and ovary clearly indicates that roles of the MAPK-JNK-p38 pathways are beyond immunity.

3.6. JAK-STAT pathway and other antiviral mechanisms

3.6.1. JAK-STAT pathway and its regulation

The JAK-STAT pathway is involved in antiviral immune responses in insects (Dostert et al., 2005; Kingsolver et al., 2013). In *Drosophila*, an extracellular protein, Unpaired3, binds to Domeless, causes receptor dimerization, and recruits STAM and Hopscotch/JAK, which in turn phosphorylates itself and then STAT (Fig. 4D). We did not find an Unpaired3 ortholog in *M. sexta* or *T. castaneum* (Zou et al., 2007). However, the *M. sexta* ortholog of Vago may bind to an unknown receptor to activate JAK and STAT in a way similar to the unknown ligand of Domeless. After phosphorylation, the STAT dimer translocates into the nucleus to induce antiviral gene expression. SOCS (a JAK inhibitor) and PIAS (protein inhibitor of activated STAT) may down-regulate the pathway. Except for the ligand, orthologs of all the pathway components are present in *M. sexta* (Table S1). Domeless and SOCS mRNA levels increased 2.6-fold in larval fat body at 24 h after the injection of a mixture of bacteria (Gunaratna and Jiang, 2013) (Table 1). We also found that their mRNA levels became more abundant in fat body and midgut between wandering larval and early pupal stages (Fig. 2D). Similar increases were observed for other members of the predicted pathway, including JAK, STAT and STAM.

3.6.2. RNA interference (RNAi) pathways

RNA interference plays important roles in limiting viral infection in insects (Kingsolver et al., 2013; Fablet, 2014). There are three RNAi pathways (Fig. 4E): 1) small interfering RNAs (siRNAs) are generated from double-stranded RNA (dsRNA) of viruses and siRNAs

degrade or inhibit viral RNA and thereby disrupt the viral infection cycle; 2) microRNAs (miRNAs) are produced from cellular gene transcripts and typically function to control the translation or half-life of their target transcripts, including those regulating immune responses; 3) Piwi-interacting RNAs (piRNAs) provide epigenetic control of transposable elements and viral transcripts in germ-line cells in order to prevent genome disruption. The siRNA pathway is mostly responsible for antiviral activity in insects. Viral RNAs may form double stranded RNAs due to innate secondary structures or via replication intermediate, and these dsRNAs are recognized and cleaved by Dicer-2 to generate siRNAs, which are then loaded into RNA-induced silencing complexes consisting of Argonaute-2 and other proteins. Unwinding of the duplex occurs along with guide strand selection. After target RNA recognition by the guide RNA, the targeted viral RNA is degraded by Argonaute-2. We have identified 28 putative pathway members suggesting that these pathways are functional in *M. sexta* (Fig. 4E, Table S1). Since R2D2 is not found in *M. sexta*, we suggest that R3D1 (an ortholog of *Drosophila* Loquacious) acts as a Dicer-1 partner in the miRNA pathway, as well as a Dicer-2 partner in the siRNA pathway. Unlike *Drosophila*, which has distinct Piwi and Aubergine genes, lepidopteran insects have a single PIWI-clade protein that we refer to as Aub/Piwi. Transcript levels for members of the siRNA pathway are relatively higher than those for either piRNA or miRNA pathways (Fig. 2E), consistent with its greater role in antiviral immunity (Kingsolver et al., 2013). Expression profiles of these pathways do not exhibit fat body- and midgut-specific up-regulation from wandering to early pupal stage, except for Dicer-2 and Argonaute-2. The Argonaute-2 mRNA levels increased moderately in induced fat body and hemocytes (Table 1). Although transcript abundances for piRNA pathway components vary, they are almost

always higher in testis and ovary than the other tissues, consistent with their roles in the germline cells.

3.6.3. Autophagy

Autophagy is a cellular process in which dysfunctional or unnecessary cellular materials or components are selectively targeted, then separated from the cytoplasm in double membrane vesicles (autophagosomes), and ultimately degraded by lysosomes (Mulakkal et al., 2014). Some pathogens may also be targeted to autophagosomes. Autophagy recycles the cellular materials and maintains cellular homeostasis under a variety of conditions. It is implicated in cellular responses to stress by nutrient-restriction, developmental changes involving tissue reorganization during metamorphosis, and certain pathological processes. The signaling of autophagy is mediated through the phosphoinositide 3-kinase (PI3K)-Akt pathway (Fig. 4F), which phosphorylates TOR to suppress autophagy. Autophagy itself involves about 20 components conserved throughout eukaryotes from yeast to mammals. In *Drosophila*, autophagy is induced upon infection by some viruses, intracellular bacteria (*e.g. Listeria monocytogenes*), and other pathogens (Yano et al., 2008; Kingsolver et al., 2013), suggesting that in addition to other cellular functions, it may also serve as an ancient cellular immune response. We have identified orthologs of all known autophagy pathway members (Fig. 4F, Table S1) and examined their expression profiles (Fig. 2F). As components of a ubiquitination complex, Atg3, 4, 5, 7, 8, 10, 12 and 16 are highly expressed in all the tissue samples used for RNA-Seq analyses. The mRNA levels of these autophagy pathway genes are generally higher in midgut than in fat body, testis and ovary. Since there is no major increase in mRNA levels in the pupal stage, autophagy may be partly supported by pre-existing proteins. Transcript

levels of Atg2 through 6, 8, 9, and 16 are up-regulated in fat body and midgut from wandering larvae and young pupae, and decrease in the later stages. These changes may correlate with cellular reorganization in cells undergoing metamorphosis. In contrast, the PI3K, Akt, TOR, Vps34, Atg1, 7, 10, 12, 13, 17, 18, and 101 mRNA levels remain high from pupal to adult stage. Based on our current data (Fig. 2F), expression of autophagy-related genes appears to be a development-regulated process. There is no strong correlation with their immune inducibility, perhaps due to the fact we did not use viruses or intracellular bacteria to challenge the larvae.

3.6.4. Apoptosis

Apoptosis, the best characterized mechanism of programmed cell death, is a part of normal developmental processes such as tissue modeling and homeostasis, but apoptosis can also participate in pathological processes including cancer and defense against pathogens (Opferman and Korsmeyer, 2003). In *Drosophila*, the initiator caspase Dronc and an adaptor protein (Ark) form a large protein complex (apoptosome) in response to intrinsic signals (Hay and Guo, 2006). It is not clear how the other *Drosophila* initiator caspases, Dredd and Strica, are activated. Once Dronc is activated, it cleaves and activates effector caspases such as Drice and Dcp1 to cleave other protein substrates that lead to the downstream events of programmed cell death. Negative regulators of caspases (*e.g*. IAPs, Dnr1) control the pathway by inhibiting the activation of initiator caspases through either direct binding or by ubiquitination-induced degradation (Orme and Meier, 2009). Likewise, IAP antagonists (*e.g*. Reaper, Hid, Grim and Sickle) inactivate IAPs and, thus induce apoptosis. We have identified 12 members of the core apoptosis pathway in *M. sexta*, including Reaper, IAP1, IAP2, Deterin/IAP3, Dnr1, Ark, Dredd/caspase-6,

Dronc/caspase-5, caspase-1, -3, and -4 (Fig. 4G) (Courtiade et al., 2011). While Dnr1, Dredd, and IAP2 are likely involved in the balance between the Imd and apoptosis pathways, the other proteins may be devoted to programmed cell death. Reaper, an indirect pathway activator, is produced in the embryo, pupal fat body and midgut, as well as adult head, Malpighian tubules, testis and ovary (Fig. 2G), suggesting a possible role of apoptosis in tissue remodeling. The IAP3 mRNA, which is related to Survivin, a mitotic spindle-associated protein, is strikingly high and may perhaps regulate embryonic development. With a similar expression profile, IAP1 may block caspase-3 and -4 in cells of midgut, fat body, and other tissues. The high transcript abundances in midgut of feeding and wandering larvae, pupae and adults could indicate that the tissue is poised to undergo or carefully regulate active programmed cell death and regeneration. In addition, the caspase-1 and IAP1 mRNA peaks in fat body and midgut from wandering to early pupal stage correlate with their immune inducibility (Table 1).

3.7. Concluding remarks

Our search of the *M. sexta* genome has yielded 187 genes encoding 198 putative members of the immunity-related signal transduction pathways, namely Toll, Imd, MAPK-JNK-p38, JAK-STAT, piRNA, siRNA, miRNA, autophagy and apoptosis. Analysis of the expression profiles reveals differences among the proposed pathways (*e.g.* Toll, Imd, and MAPK-JNK-p38) and among some of the components (*e.g.* Spätzles, Tolls). These results suggest that the intracellular signaling system is functional in this undomesticated insect, and thus pave the way for understanding and potentially modulating similar pathways in pest lepidopteran species. The proposed signaling network needs experimental validation using biochemical, molecular and cellular biological methods.

Acknowledgments

This work was supported by NIH grant GM58634 (to H. Jiang) and a DARPA/NSF (IOS-1354421) grant (to G. Blissard). Computation for this project was performed at OSU High Performance Computing Center at Oklahoma State University supported in part through NSF grant OCI-1126330. This work was approved for publication by the Director of Oklahoma Agricultural Experimental Station, and supported in part under project OKLO2450.

Tables

Table 1. Relative mRNA abundances of the intracellular signaling pathway members in induced (I) and control (C) fat body (F) and hemocytes from the larvae of *M. sexta*.

Name	IF/CF	IH/CH	Name	IF/CF	IH/CH	Name	IF/CF	IH/CH	Name	IF/CF	IH/CH
Spätzle1*	1.4	3.9	Dredd*	1.8	1.2	MLK1*	1	2.1	R3D1/Loqs	0.8	1.3
Spätzle2	1.5	2.2	Relish*	5.2	1.5	MKK4*	0.8	0.5	Dicer1	0	0.7
Spätzle7	4.1	3.6	NTF2*	1.2	1.7	JNK*	1.8	1.6	Ago1	1	2.6
Toll1*	2.5	6.2	TAK1*	0.5	3.6	Basket	1.3	2.7	Drosha	1.3	1.1
Toll2	0.7	0.8	Tab2*	2.5	1	ECSIT*	1	1.7	Pasha	-	1.5
Toll3	2.5	6.2	IKKβ*	0.5	0.2	MEKK1*	0	0.6	Expotin5	4.1	1.7
Toll4	2.5	6.4	IKKγ*	2.5	1.2	MKK3*	1.4	1.1	Nibbler	0.7	1.4
Toll5	2.9	0.9	NEMO	1	1.2	p38*	2.2	1	Gawky	0.9	0.9
ML1	2.2	0.6	Dnr1	2.2	0.5	Aop*	3.1	1.2	Me31B	1	0.8
ML2	1.8	0.8	Sickie*	1	33.2	FOS*	2	1.9	Ge-1	0.5	1
MyD88*	1.5	1.5	Caspar*	3.1	1.6	Jra*	1.6	0.9	Atg1	2.8	0.9
Tube*	8.4	0.8	IAP2*	0.3	0.9	Ebi	1	1.4	Atg13	0.4	1.3
Pelle*	5.1	2.2	Bendless*	1.2	2.1	Smrter	1.8	1.2	Atg101	1.3	0.6
Pellino*	2.3	1.3	Uev1A*	1.4	1.1	Rpd3/HDAC1	0.2	1.2	Atg17	0.6	0.7
Cactus*	9.2	1.8	Effete	0.8	1.1	Domeless*	2.4	0.8	Vps34	0.5	1.2
Dorsal*	1.3	1.2	USP36	0.5	1.4	Stam*	2	1.1	Vps15	0.5	1.6
Tollip-1*	1	-	POSH1*	1.7	0.9	JAK/Hopscotch*	1	0.5	Atg6	0.5	1.2
Tollip-2*	0.8	1.1	POSH2	0.8	1.3	STAT*	0.4	0.7	Atg18	3.6	0.8
Ref2P*	1.5	1	CYLD	3.6	0.8	PIAS*	1.6	1	Atg12	0.5	3
aPKC*	-	1	SkpA	0.5	1.7	SOCS*	2.5	0.8	Atg7	0.5	0.5
GPRK2	0.1	0.7	Cullin	0.9	1.3	ZHF1	0.7	1	Atg5	0.5	1
Cactin	0.5	2.5	SlimB	1.5	0.7	Piwi	0.4	0.8	Atg4	0.2	1.1
Aos1*	0.7	1.7	Akirin	9.2	1.7	Armitage	1.4	1	Atg8	3.1	0.6
Uba2*	2.5	1.5	Dsp	2	1.4	Yb	0.5	0.2	Atg3	0.3	1
Lesswright*	4.6	1.2	Eiger*	0.8	8	Shu	0.5	5.9	Atg2	1.2	0.2
Ulp1	5.6	0.8	PVR*	1	1.4	Qin	0.7	0.9	Atg9	1.5	0.9
Kurtz	1	0.7	Ras85D*	0.7	1.7	Dicer2	1.3	1.5	Akt	1	1
Smt3*	1.6	1.6	Rac1*	2.5	1.3	Ago2	0.8	1	TOR	0.2	0.8
Deaf1	-	0.9	Cdc42*	1.3	1.4	Vig	0.9	1.1	PI3K	1.5	0.9
Serpent*	0.4	0.9	MASK*	1.2	1.4	TSN	0.6	1.2	IAP1	1.6	1.1
Pannier-1	0.5	2.4	Polehole	0.5	1	Ars2	3.1	1.7	Deterin	-	0.7
Pannier-2	1	2.4	Dsor1*	0.5	1.1	CBC	-	3.1	Dronc	0.5	1.2
GATAe	0.5	2.4	Rolled	1.9	1.1	Belle/Cap	1.2	1.5	Ark	2	0.6
U-shaped	0.3	1.1	Pointed	0.3	0.9	Blanks	2.5	1.1	Caspase-1	1.5	1.4
Imd*	2.7	1	Misshapen*	1.6	0.8	Translin	0.5	0.9			
FADD*	0.6	1.3	Hep/MKK7*	1.5	1.2	Tis11	1.3	0.9			

As described in *Section 2.3*, the transcriptome data of larval fat body and hemocytes before and after the immune challenge (Zhang et al., 2011) were processed again according to Gunaratna and Jiang (2013), based on the BLAST search using 196 complete coding sequences as queries. The ones with no hit in the CIFH library are omitted from the table. Note that, due to the increase in query sizes and contig hits, the reported relative abundances (*) (*i.e*. IF/IH and CF/CH) (Gunaratna and Jiang, 2013) may be different for certain genes. "-": C and I = 0.

Figures

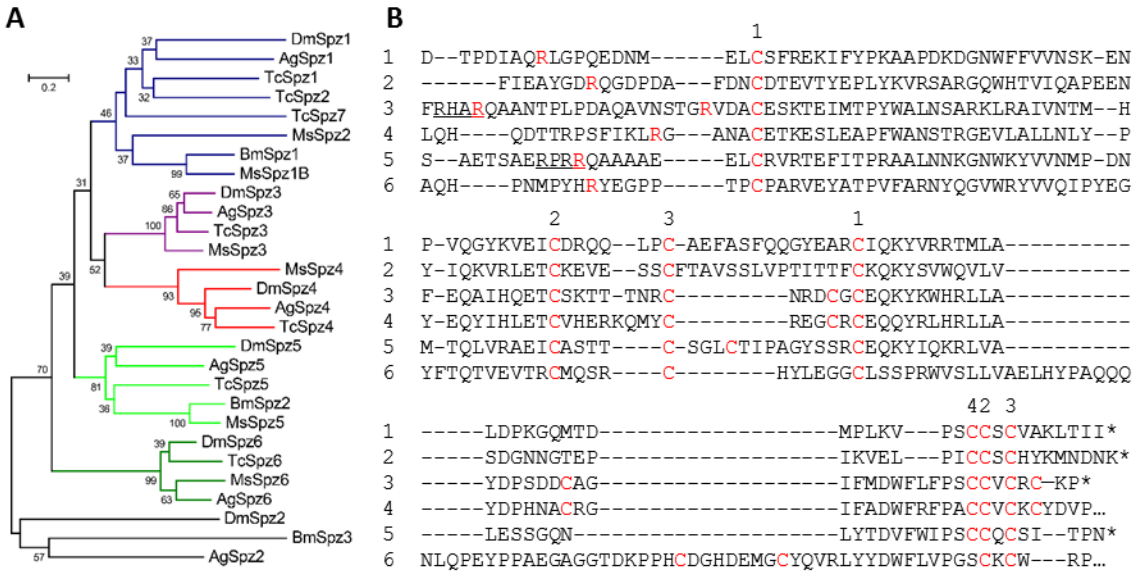
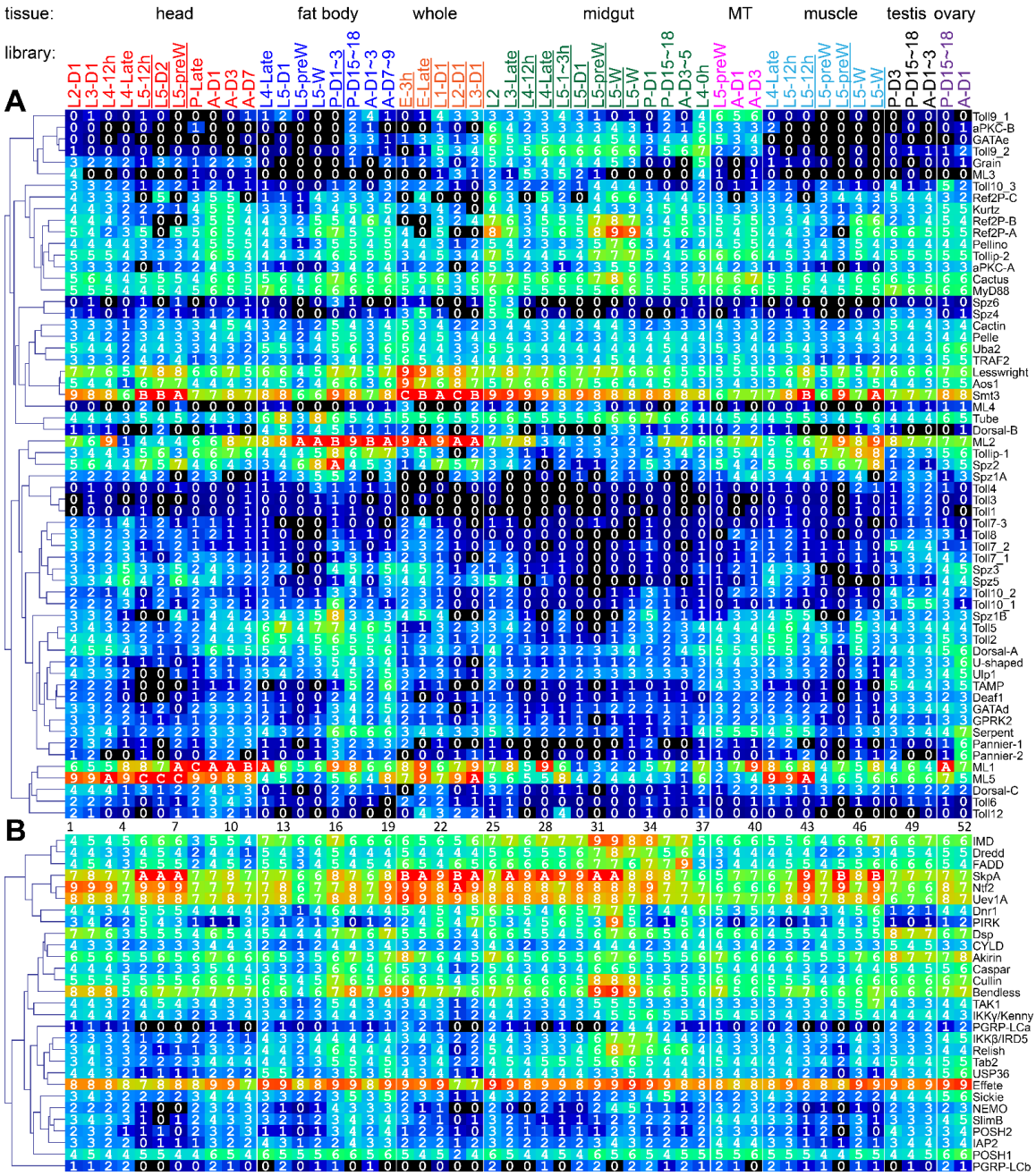


Fig. 1. Phylogenetic relationships of Spätzles in *M. sexta*, *B. mori*, *T. castaneum*, and *D. melanogaster*. (**A**) Tree. Based on the sequence alignment of 29 full-length Spätzles, a tree was generated with branches shown in colors representing closely related groups. (**B**) Aligned sequences of the cystine-knot cytokine domains in *M. sexta* Spätzles-1 through 7. Cys residues are indicated in a red font. Some Cys residues may form intra- (1−1, 2−2, 3−3) and inter- (4) chain disulfide bonds. Proteolytic activation sites, known for Spätzle-1, are predicted to be next to the Arg (red) in Spätzle-2 through 6. The putative processing site (RXXR) is underlined in Spätzle-3 and 5.



(continued on the next page)

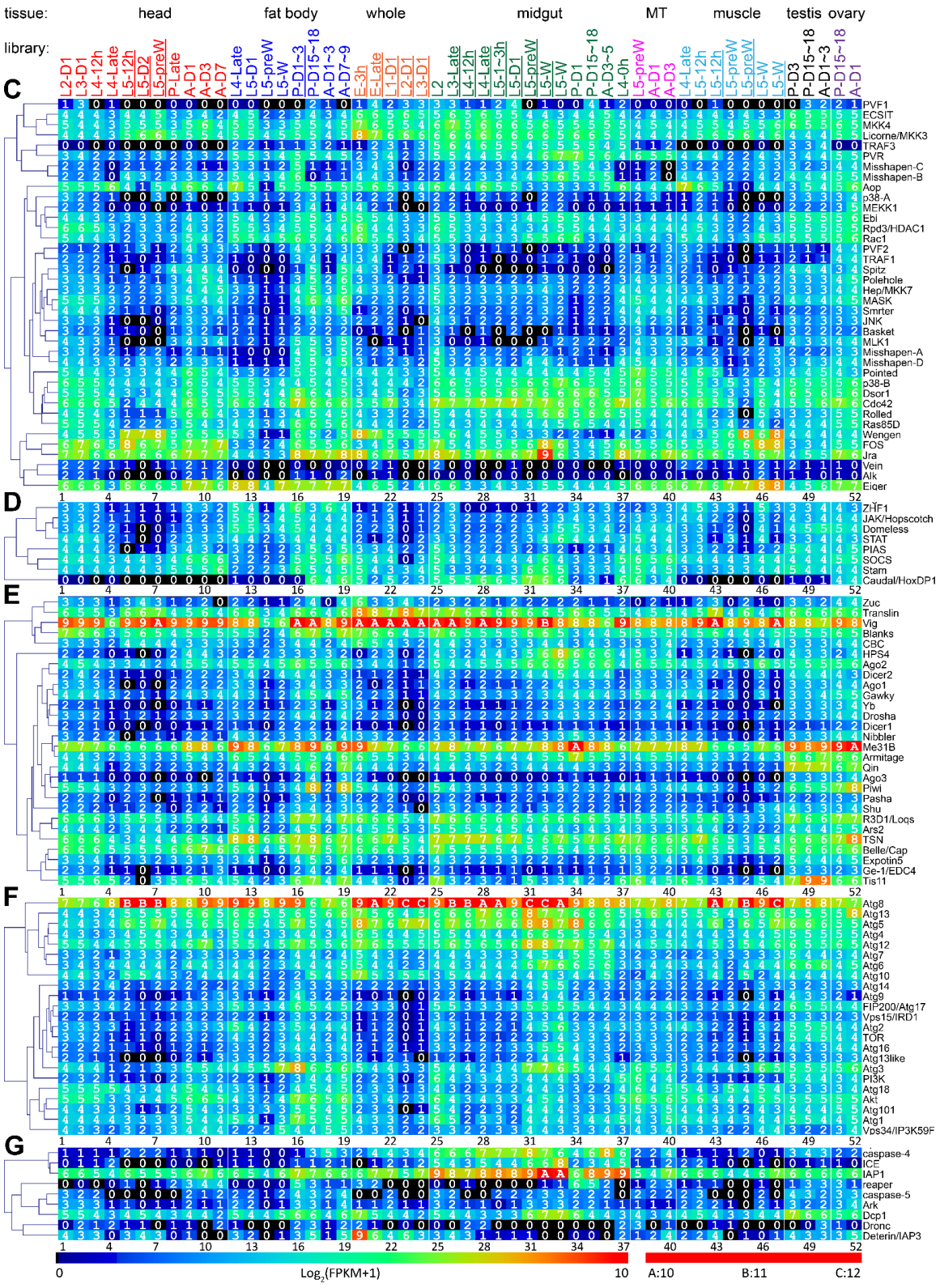


Fig. 2. Transcript profiles of the putative signaling protein genes in the 52 tissue samples. The mRNA levels, as represented by $\log_2$(FPKM+1) values, are shown in the gradient heat

map from blue (0) to red (⩾10). The values of 0−0.49, 0.50−1.49, 1.50−2.49 ⋯ 8.50−9.49, 9.50−10.49 10.50−11.49, and 11.50−12.49 are labeled 0, 1, 2 … 9, A, B and C, respectively. The cDNA libraries are constructed from the following tissues and stages: head [2nd (instar) L (larvae), d1 (day 1); 3rd L, d1; 4th L, d0.5; 4th L, late; 5th L, d0.5; 5th L, d2; 5th L, pre-W (pre-wandering); P (pupae), late; A (adults), d1; A, d3; A, d7], fat body (4th L, late; 5th L, d1; 5th L, pre-W; 5th L, W; P, d1-3; P, d15-18; A, d1-3; A, d7-9), whole animals [E (embryos), 3h; E, late; 1st L; 2nd L; 3rd L), midgut (2nd L; 3rd L; 4th L, 12h; 4th L, late; 5th L, 1-3h; 5th L, 24h; 5th L, pre-W; 5th L, W; P, d1; P, d15-18; A, d3-5; 4th L, 0h), Malpighian tubules (MT) (5th L, pre-W; A, d1; A, d3), muscle (4th L, late; 5th L, 12h; 5th L, pre-W; 5th L, W), testis (P, d3; P, d15-18; A, d1-3), and ovary (P, d15-18; A, d1). Some libraries (underlined) are from single-end sequencing; the others are from paired-end sequencing. Note that some synonymous libraries exhibit different FPKMs due to method differences. Panel **A**, Toll; **B**, Imd with JNK branch; **C**, MAPK-JNK-p38; **D**, JAK-STAT; **E**, pi- si- and mi-RNA pathways, **F**, autophagy; **G**, apoptosis.

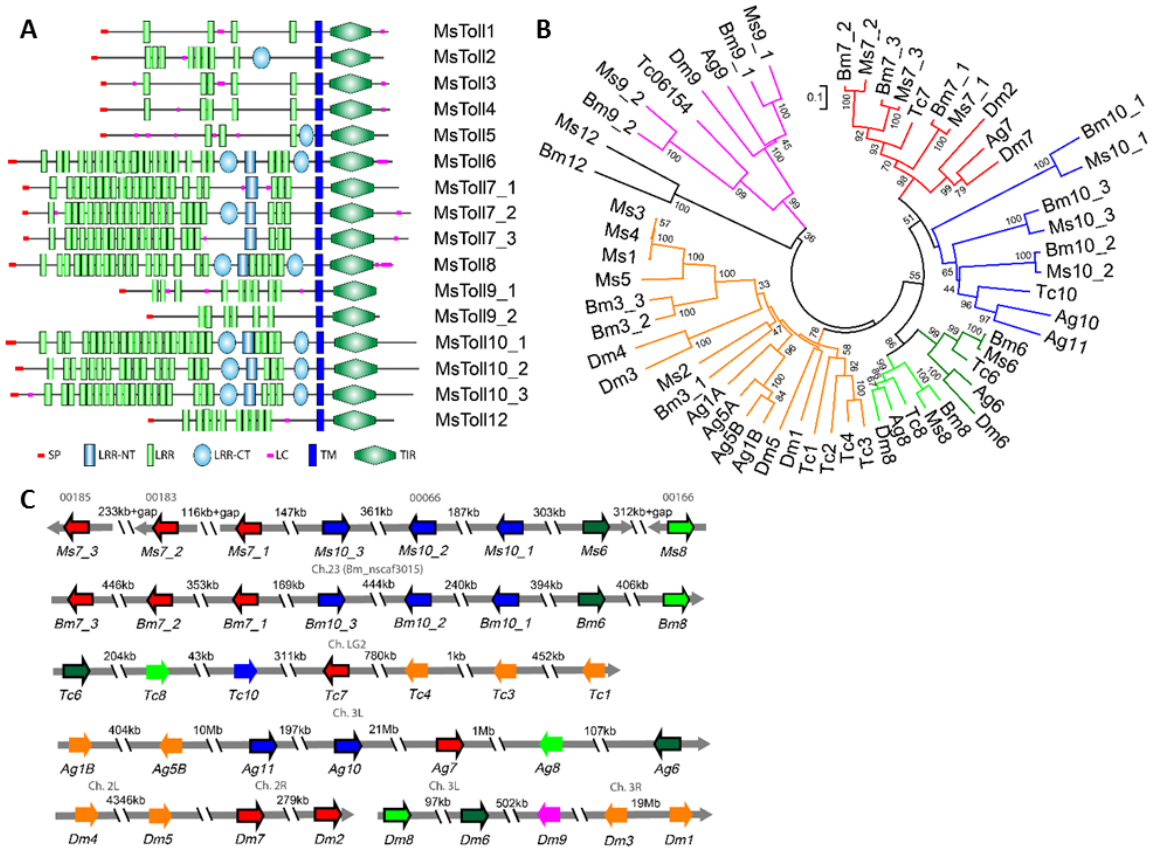


Fig. 3. Domain structures (**A**), phylogenetic relationships (**B**), and gene orders (**C**) of Tolls in *M. sexta*. (**A**) Signal peptide (SP), Leu-rich repeat (LRR), amino- and carboxyl-terminal (NT & CT) LRRs, low complexity (LC) region, transmembrane (TM) segment, and TIR (Toll/interleukin-1 receptor) domain are shown in different colors and shapes as indicated. (**B**) Amino acid sequences of the 58 full-length Toll proteins from *M. sexta*, *B. mori*, *T. castaneum*, *A. gambiae*, and *D. melanogaster* are aligned to generate the tree with its branches in different colors for closely related groups. (**C**) Orientations and orders of the Toll genes in the five insects are schematically shown as arrows in the same colors as in panel B. Arrows for the single exon genes are in black frame.

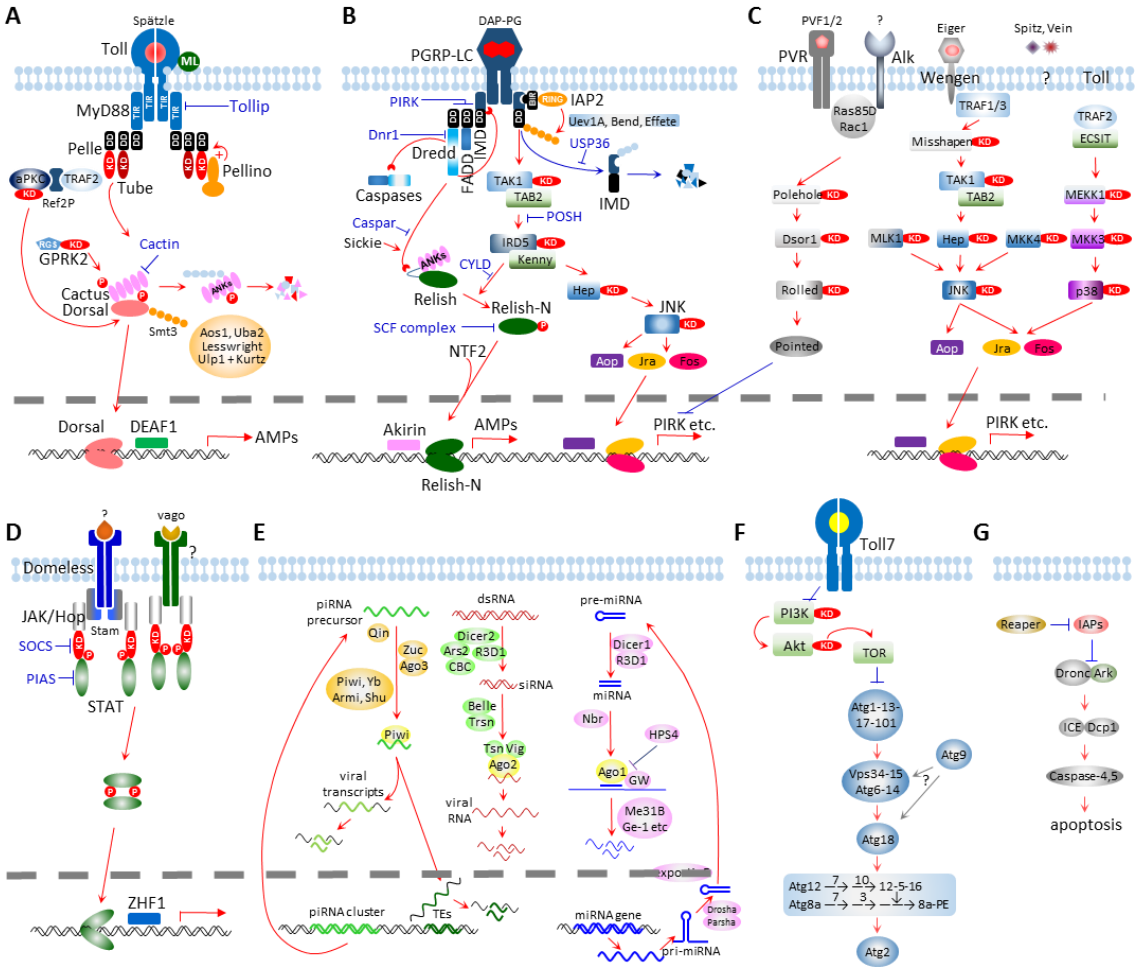


Fig. 4. Putative signaling pathways and regulators for antimicrobial immune responses in *M. sexta*. Panels **A**, Toll; **B**, Imd with JNK branch; **C**, MAPK-JNK-p38; **D**, JAK-STAT; **E**, pi- si- and mi-RNA pathways, **F**, autophagy; **G**, apoptosis. Panels A through G are described in the text.

References

Aggarwal, B.B., 2003. Signalling pathways of the TNF superfamily: a double-edged sword. Nat Rev Immunol. 3, 745–756.

Altincicek B, Vilcinskas A., 2008. Identification of a lepidopteran matrix metalloproteinase with dual roles in metamorphosis and innate immunity. Dev Comp Immunol. 32, 400–409.

An, C., Jiang, H., Kanost, M.R., 2010. Proteolytic activation and function of the cytokine Spätzle in the innate immune response of a lepidopteran insect, *Manduca sexta*. FEBS J. 277, 148–162.

Anjum, S.G., Xu, W., Nikkholgh, N., Basu, S., Nie, Y., 2013. Regulation of Toll signaling and inflammation by β-arrestin and the SUMO protease Ulp1. Genetics. 195, 1307–1317.

Ao, J.-Q., Ling, E., Rao, X.J., Yu, X.Q., 2008a. A novel ML protein from *Manduca sexta* may function as a key accessory protein for lipopolysaccharide signaling. Mol Immunol. 45, 2772–2781.

Ao, J.-Q., Ling, E., Yu, X.-Q., 2008b. A Toll receptor from *Manduca sexta* is in response to *Escherichia coli* infection. Mol Immunol. 45, 543–552.

Avila, A., Silverman, N., Diaz-Meco, M.T., Moscat, J., 2002. The *Drosophila* atypical protein kinase C-ref(2)p complex constitutes a conserved module for signaling in the toll pathway. Mol Cell Biol. 22, 8787–8795.

Baeg, G.H., Zhou, R., Perrimon, N., 2005. Genome-wide RNAi analysis of JAK/STAT signaling components in *Drosophila*. Genes Dev. 19, 1861–1870.

Ballard, S.L., Miller, D.L., Ganetzky, B., 2014. Retrograde neurotrophin signaling through Tollo regulates synaptic growth in *Drosophila*. J Cell Biol. 204, 1157–1172.

Bhaskar, V., Smith, M., Courey, A.J., 2002. Conjugation of Smt3 to Dorsal may potentiate the *Drosophila* immune response. Mol Cell Biol. 22, 492–504.

Bonnay, F., Nguyen, X.H., Cohen-Berros, E., Troxler, L., Batsche, E., Camonis, J., Takeuchi, O., Reichhart, J.M., Matt, N., 2014. Akirin specifies NF-κB selectivity of *Drosophila* innate immune response via chromatin remodeling. EMBO J. 33, 2349–2362.

Buchon, N., Silverman, N., Cherry, S., 2014. Immunity in *Drosophila melanogaster* - from microbial recognition to whole-organism physiology. Nat Rev Immunol. 14, 796–810.

Cao, X., He, Y., Hu, Y., Zhang, X., Wang, Y., Zou, Z., Chen, Y., Blissard, G., Kanost, M.R., Jiang, H., 2015. Sequence conservation, phylogenetic relationships, and expression

profiles of nondigestive serine proteases and serine protease homologs in *Manduca sexta*. Insect Biochem Mol Biol. in press.

Cao, X., Jiang, H., 2015. Integrated modeling of protein-coding genes in the *Manduca sexta* genome using RNA-Seq data from the biochemical model insect. Insect Biochem. Mol. Biol. in press.

Cardozo, T., Pagano, M., 2004. The SCF ubiquitin ligase: insights into a molecular machine. Nat Rev Mol Cell Biol. 5, 739–751.

Chen, J., Xie, C., Tian, L., Hong, L., Wu, X., Han, J., 2010. Participation of the p38 pathway in *Drosophila* host defense against pathogenic bacteria and fungi. Proc Natl Acad Sci USA. 107, 20774–20779.

Chiu, H., Ring, B.C., Sorrentino, R.P., Kalamarz, M., Garza, D., Govind, S., 2005. dUbc9 negatively regulates the Toll-NF-kappa B pathways in larval hematopoiesis and drosomycin activation in *Drosophila*. Dev Biol. 288, 60–72.

Courtiade, J., Pauchet, Y., Vogel, H., Heckel, D.G., 2011. A comprehensive characterization of the caspase gene family in insects from the order Lepidoptera. BMC Genomics. 12, 357.

Dong, C., Davis, R.J., Flavell, R.A., 2002. MAP kinases in the immune response. Ann Rev Immunol. 20, 55–72.

Dong, Y., Aguilar, R., Xi, Z., Warr, E., Mongin, E., Dimopoulos, G., 2006. *Anopheles gambiae* immune responses to human and rodent *Plasmodium* parasite species. PLoS Pathog. 2, e52.

Dostert, C., Jouanguy, E., Irving, P., Troxler, L., Galiana-Arnoux, D., Hetru, C., Hoffmann, J.A., Imler, J.-L., 2005. The JAK-STAT signaling pathway is required but not sufficient for the antiviral response of *Drosophila*. Nat Immunol. 6, 946–953.

Ertürk-Hasdemir, D., Broemer, M., Leulier, F., Lane, W.S., Paquette, N., Hwang, D., Kim, C.H., Stöven, S., Meier, P., Silverman, N., 2009. Two roles for the *Drosophila* IKK complex in the activation of Relish and the induction of antimicrobial peptide genes. Proc Natl Acad Sci USA 106, 9779–9784.

Evans, J.D., Aronstein, K., Chen, Y.P., Hetru, C., Imler, J.L., Jiang, H., Kanost, M., Thompson, G.J., Zou, Z., Hultmark, D., 2006. Immune pathways and defence mechanisms in honey bees *Apis mellifera*. Insect Mol Biol. 15, 645–656.

Fablet, M., 2014. Host control of insect endogenous retroviruses: small RNA silencing and immune response. Viruses. 6, 4447–4464.

Foley, E., O'Farrell, P.H., 2004. Functional dissection of an innate immune response by a genome-wide RNAi screen. PLoS Biol. 2, e203.

Gerardo, N.M., Altincicek, B., Anselme, C., Atamian, H., Barribeau, S.M., de Vos M., Duncan, E.J., Evans, J.D., Gabaldón, T., Ghanim, M., Heddi, A., Kaloshian, I., Latorre, A.,

Moya, A., Nakabachi, A., Parker, B.J., Pérez-Brocal, V., Pignatelli, M., Rahbé, Y., Ramsey, J.S., Spragg, C.J., Tamames, J., Tamarit, D., Tamborindeguy, C., Vincent-Monegat, C., Vilcinskas, A., 2010. Immunity and other defenses in pea aphids, *Acyrthosiphon pisum*. Genome Biol. 11, R21

Geuking, P., Narasimamurthy, R., Lemaitre, B., Basler, K., Leulier, F., 2009. A non-redundant role for *Drosophila* MKK4 and hemipterous/MKK7 in TAK1-mediated activation of JNK. PLoS One. 4, e7709.

Gillespie, J.P., Kanost, M.R., Trenczek, T., 1997. Biological mediators of insect immunity. Ann Rev Entomol 42, 611–643.

Gunaratna, R.T., Jiang, H., 2013. A comprehensive analysis of the *Manduca sexta* immunotranscriptome. Dev Comp Immunol. 39, 388–398.

Hay, B.A., Guo, M., 2006. Caspase-dependent cell death in *Drosophila*. Ann Rev Cell Dev Biol. 22, 623–650.

He, Y., Cao, X., Li, K., Hu, Y., Chen, Y., Blissard, G., Kanost, M.R., Jiang, H., 2015. A genome-wide analysis of antimicrobial effector genes and their transcription patterns in *Manduca sexta*. Insect Biochem Mol Biol. in press

Imler, J.L., Hoffmann, J.A., 2001. Toll receptors in innate immunity. Trends Cell Biol. 11, 304–311.

Jiang, H., Vilcinskas, A., Kanost, M.R., 2010. Immunity in lepidopteran insects. Adv Exp Med Biol. 708, 181–204.

Kaneko, T., Yano, T., Aggarwal, K., Lim, J.-H., Ueda, K., Oshima, Y., Peach, C., Erturk-Hasdemir, D., Goldman, W.E., Oh, B.-H., 2006. PGRP-LC and PGRP-LE have essential yet distinct functions in the *Drosophila* immune response to monomeric DAP-type peptidoglycan. Nat Immunol. 7, 715–723.

Kingsolver, M.B., Huang, Z., Hardy, R.W., 2013. Insect antiviral innate immunity: pathways, effectors, and connections. J Mol Biol. 425, 4921–4936.

Kisseleva, T., Bhattacharya, S., Braunstein, J., Schindler, C.W., 2002. Signaling through the JAK/STAT pathway, recent advances and future challenges. Gene 285, 1–24.

Kleino, A., Myllymäki, H., Kallio, J., Vanha-aho, L.M., Oksanen, K., Ulvila, J., Hultmark, D., Valanne, S., Rämet, M., 2008. Pirk is a negative regulator of the *Drosophila* Imd pathway. J Immunol. 180, 5413–5422.

Kleino, A., Silverman, N., 2014. The *Drosophila* Imd pathway in the activation of the humoral immune response. Dev Comp Immunol. 42, 25–35.

Kopp, E., Medzhitov, R., Carothers, J., Xiao, C., 1999. ECSIT is an evolutionarily conserved intermediate in the Toll/IL-1 signal transduction pathway. Genes Dev. 13, 2059–2071.

Kristensen, N.P., Scoble, M.J., Karsholt, O., 2007. Lepidoptera phylogeny and systematics: the state of inventorying moth and butterfly diversity. Zootaxa 1668, 699–747.

Langmead, B., Trapnell, C., Pop, M., Salzberg, S.L., 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25.

Lee, K.Z., Ferrandon, D., 2011. Negative regulation of immune responses on the fly. EMBO J. 30, 988–990

Lemaitre, B., Hoffmann, J., 2007. The host defense of *Drosophila melanogaster*. Ann Rev Immunol 25, 697–743.

Li, B., Dewey, C., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC bioinformatics 12, 323.

Lin, P., Huang, L.H., Steward, R., 2000. Cactin, a conserved protein that interacts with the *Drosophila* IκB protein cactus and modulates its function. Mech Dev. 94, 57–65.

Liu, H., Su, Y.C., Becker, E., Treisman, J., Skolnik, E.Y., 1999. A *Drosophila* TNF-receptor-associated factor (TRAF) binds the Ste20 kinase Misshapen and activates Jun kinase. Curr Biol. 9, 101–104.

McIlroy, G., Foldi, I., Aurikko, J., Wentzell, J.S., Lim, M.A., Fenton, J.C., Gay, N.J., Hidalgo, A., 2013. Toll-6 and Toll-7 function as neurotrophin receptors in the *Drosophila melanogaster* CNS. Nat Neurosci. 16, 1248–1256.

Mellroth, P., Karlsson, J., Håkansson, J., 2005. Ligand-induced dimerization of *Drosophila* peptidoglycan recognition proteins *in vitro*. Proc Natl Acad Sci USA. 102, 6455–6460.

Mulakkal, N.C., Nagy, P., Takats, S., Tusco, R., Juhász, G., Nezis, I.P., 2014. Autophagy in *Drosophila*: from historical studies to current knowledge. BioMed Res Int. 2014, 273473,

Nakamoto, M., Moy, R.H., Xu, J., Bambina, S., Yasunaga, A., Shelly, S.S., Gold, B., Cherry, S., 2012. Virus recognition by Toll-7 activates antiviral autophagy in *Drosophila*. Immunity 36, 658–667.

Opferman, J.T., Korsmeyer, S.J., 2003. Apoptosis in the development and maintenance of the immune system. Nat Immunol. 4, 410–415.

Orme, M., Meier, P., 2009. Inhibitor of apoptosis proteins in *Drosophila*: gatekeepers of death. Apoptosis 14, 950–960

Paddibhatla, I., Lee, M.J., Kalamarz, M.E., Ferrarese, R., Govind, S., 2010. Role for sumoylation in systemic inflammation and immune homeostasis in *Drosophila* larvae. PLoS Pathog. 6, e1001234.

Paquette, N., Broemer, M., Aggarwal, K., Chen, L., Husson, M., Ertürk-Hasdemir, D., Reichhart, J.M., Meier, P., Silverman, N., 2010. Caspase-mediated cleavage, IAP binding,

and ubiquitination: linking three mechanisms crucial for *Drosophila* NF-κB signaling. Mol. Cell 37, 172–182.

Petersen, T., Brunak, S., von Heijne, G., Nielsen, H., 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 8, 785–786.

Ragab, A., Buechling, T., Gesellchen, V., Spirohn, K., Boettcher, A.L., Boutros, M., 2011. *Drosophila* Ras/MAPK signalling regulates innate immune responses in immune and intestinal stem cells. EMBO J. 30, 1123–1136.

Rämet, M., Manfruelli, P., Pearson, A., Mathey-Prevot, B., Ezekowitz, R.A., 2002. Functional genomic analysis of phagocytosis and identification of a *Drosophila* receptor for *E. coli.* Nature. 416, 644–648.

Rao, X-J., Cao, X., He, Y., Hu, Y., Zhang, X., Chen, Y., Blissard, G., Kanost, M.R., Yu, X-Q., Jiang, H., 2015. Structural features, evolutionary relationships, and transcriptional regulation of C-type lectin-domain proteins in *Manduca sexta*. Insect Biochem Mol Biol. in press.

Rao, X-J., Xu, X-X., Yu, X-Q., 2011. *Manduca sexta* moricin promoter elements can increase promoter activities of *Drosophila melanogaster* antimicrobial peptide genes. Insect Biochem Mol Biol. 41, 982–992.

Sorrentino, R.P., Melk, J.P., Govind, S., 2004. Genetic analysis of contributions of dorsal group and JAK-Stat92E pathway genes to larval hemocyte concentration and the egg encapsulation response in *Drosophila*. Genetics 166, 1343–1356

Strand, M.R., 2008. The insect cellular immune response. Insect Sci. 15, 1–15.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol Biol Evo. 30, 2725–2729.

Tanaka, H., Ishibashi, J., Fujita, K., Nakajima, Y., Sagisaka, A., Tomimoto, K., Suzuki, N., Yoshiyama, M., Kaneko, Y., Iwasaki, T., Sunagawa, T., Yamaji, K., Asaoka, A., Mita, K., Yamakawa, M., 2008. A genome-wide analysis of genes and gene families involved in innate immunity of *Bombyx mori*. Insect Biochem Mol Biol. 38, 1087–1110.

Thevenon, D., Engel, E., Avet-Rochex, A., Gottar, M., Bergeret, E., Tricoire, H., Benaud, C., Baudier, J., Taillebourg, E., Fauvarque, M.O., 2009. The *Drosophila* ubiquitinspecific protease dUSP36/Scny targets Imd to prevent constitutive immune signaling. Cell Host Microbe 6, 309–320.

Tsichritzis, T., Gaentzsch, P.C., Kosmidis, S., Brown, A.E., Skoulakis, E.M., Ligoxygakis, P., Mosialos, G., 2007. A *Drosophila* ortholog of the human cylindromatosis tumor suppressor gene regulates triglyceride content and antibacterial defense. Development 134, 2605–2614.

Tsuda, M., Langmann, C., Harden, N., Aigaki, T., 2005. The RING-finger scaffold protein Plenty of SH3s targets TAK1 to control immunity signaling in *Drosophila*. EMBO Rep. 6, 1082–1087.

Valanne, S., Wang, J.H., Rämet, M., 2011. The *Drosophila* Toll signaling pathway. J Immunol. 186, 649–656.

X et al., 2015

Yano, T., Mita, S., Ohmori, H., Oshima, Y., Fujimoto, Y., Ueda, R., Takada, H., Goldman, W.E., Fukase, K., Silverman, N., Yoshimori, T., Kurata, S., 2008. Autophagic control of *Listeria* through intracellular innate immune recognition in *Drosophila*, Nat Immunol. 9, 908–916.

Zhang, G., Ghosh, S., 2002. Negative regulation of toll-like receptor-mediated signaling by Tollip. J Biol Chem. 277, 7059–7065.

Zhang, S., Gunaratna, R., Zhang, X., Najar, F., Wang, Y., Roe, B., Jiang, H., 2011. Pyrosequencing-based expression profiling and identification of differentially regulated genes from *Manduca sexta*, a lepidopteran model insect. Insect Biochem Mol Biol. 41, 733–746.

Zhang, X., He, Y., Cao, X., Gunaratna, R.T., Chen, Y., Blissard, G.W., Kanost, M.R., Jiang, H., 2015. Phylogenetic analysis and expression profiling of the pattern recognition receptors: insights into molecular recognition of invading pathogens in *Manduca sexta*. Insect Biochem. Mol. Biol. in press.

Zou, Z., Evans, J., Lu, Z., Zhao, P., Williams, M., Sumathipara, N., Hetru, C., Hultmark, D., Jiang H., 2007. Comparative genome analysis of the Tribolium immune system. Genome Biol. 8, R177.

CHAPTER III

A DEEP LOOK INTO THE RNA-SEQ DATA AND DEVELOPMENTAL TRANSCRIPTOME OF MANDUCA SEXTA

Xiaolong Cao[a], Haobo Jiang[b]

[a] Department of Biochemistry and Molecular Biology, Oklahoma State University, Stillwater, OK 74078, USA
[b] Department of Entomology and Plant Pathology, Oklahoma State University, Stillwater, OK 74078, USA

Key words: tobacco hornworm; RNA-seq; transcriptome; insect genomics

Abbreviations: OGS, official gene set; FPKM, fragments per kilobase per million mapped reads. BPKM, bases per kilobase per million mapped bases. An, antenna; G, midgut; F, fat body; M, muscle; T, testis; O, ovary; H, head; W, whole body, MT, malpighian tubule.

Abstract

The tobacco hornworm, *Manduca sexta* has been widely used as a model insect to study insect immunity, metabolism, nervous system, hormonal regulation and other physiological processes. 67 cDNA libraries from different tissues and different developing stages were sequenced along with the genome project or by other research groups. After analyzing the relationship between genome transcribed ratio with mapped bases, we found the transcribed ratio could be influenced by number of mapped bases, sequencing method and the library tissue resources and developmental stages. During the previouse Cufflinks gene modeling, more than 40% of the total reads cannot be mapped to the genome. We did a careful analysis, and found most unmapped reads are from ribosomal RNAs. Similarity among libraries was measured based on associated genes, and there is a clear difference between different tissues at different developmental stages. We calculated gene expression level and analyzed the most highly expressed genes in different libraries. Majority of the highly expressed genes are cuticle, muscle or odorant-binding proteins, and some are proteins with known function. We analyzed tissue-specific gene expression and identified over 20 groups of genes with distinct expression patterns, which facilitate function prediction for many unknown proteins. This work will help future research of *M. sexta*.

1 Introduction

Next generation sequencing (NGS) is a powerful tool for molecular studies of living organisms, including insects. To the date of June 1st, 2016, over 52,542 sequencing experiments from about 1,400 insect species were submitted to Sequence Read Archive (SRA) of National Center for Biotechnology Information (NCBI) using NGS technology. The two obvious outlier of insect species in terms of number of sequencing studies are *Drosophila melanogaster* and *Anopheles gambiae*, which accounts for 43% sequencing runs, 29% of sequencing bases and 16% of sequencing runs and 23% of sequencing bases, respectively. This is reasonable, as *D. melanogaster* is the mostly widely used model insects for genetics and other biological researches, and *A. gambiae* is the vector of malaria, one of the most dangerous diseases in world. For other insects, due to the lack of research resources, they were not extensively sequenced. For some species, transcriptome sequencing, or RNA-seq, becomes an excellent choice for specific research goals, providing not only gene models but also gene expression information, for instance, the immunotranscriptome of *Manduca sexta* (Gunaratna and Jiang, 2013; Zhang et al., 2011) and *Helicoverpa armigera* (Xiong et al., 2015).

As a typical holometabolous lepidopteran insect with five larval instars, a large and simple larval body, *M. sexta* has advantages over other model insects in studying physiological processes of insects, especially cuticle formation, metabolism, metamorphosis, hormonal and neural regulation, and immunity (Arrese and Soulages, 2010; Hopkins et al., 2000; Jiang et al., 2010; Riddiford et al., 2003; Shields and Hildebrand, 2001). Under laboratory condition, *M. sexta* is easy to raise with simple artificial food and has a well conserved life

cycle, with each developmental stage in a clear time range (Reinecke et al., 1980). These developing stages are egg stage, five instar feeding stages and molting sleep between them, cessation of feeding, body wetting, wandering to burrowing, dorsal pigmentation, fluid excretion, burrowing, reduced movement, stationary stage, metathoracic bars, pupation, pupa stage and adult stage. Being not an insect for genomics study, the genome of lab-raised *M. sexta* have less genome diversity and their gene expression and regulation should be much conserved as long as they were raised similarly.

Recently, the draft genome sequence of *M. sexta* is published with 52 cDNA libraries of different tissues of different life stages (Kanost et al., 2016) , together with a series of papers focusing on different genes, including microRNAs (Zhang et al., 2015b), antimicrobial effector genes (He et al., 2015), pattern recognition receptors (Zhang et al., 2015a), nondigestive serine proteases (Cao et al., 2015b), immune signaling pathway (Cao et al., 2015a), C-type lectin-domain proteins (Rao et al., 2015), cuticle proteins (Dittmer et al., 2015), chitin metabolism enzymes (Tetreau et al., 2015a) and chitin binding proteins (Tetreau et al., 2015b). The transcriptome data has greatly helped the gene modeling, and improved gene annotation and function prediction for these individual studies. Additionally, 8 cDNA libraries studying sex-biased gene expression (Smith et al., 2014) and 7 libraries studying chemosensory receptor genes (Koenig et al., 2015) were public available in SRA. The expression of genes in Official Gene Set 2.0 (OGS 2.0) were provided along with the genome paper and gene expression of individual genes were analyzed in these individual papers. However, as the genome paper was focused on the immune system and cuticle/chitin metabolism, the RNA-seq data was not thoroughly described and analyzed, and overview gene expression with the transcriptome data is not

provided. Additionally, the MCOT 1.0 models contains some protein coding genes not well modeled by OGS 2.0 (Cao and Jiang, 2015), and both OGS 2.0 and MCOT 1.0 have bias for protein coding genes but not non-coding genes, so a large portion of genes are not analyzed in the genome paper. To make full use of the public available RNA-seq data to help researchers studying *M. sexta*, we did a thorough transcriptome study with these 67 datasets.

2 Materials and Methods

2.1 Data and program acquisition

Final version of *M. sexta* Genome Assembly 1.0 (Msex 1.0) and gene models in Manduca Official Gene Sets 2.0 (OGS2.0) were downloaded from *Manduca sexta* workspace at of National Agricultural Library (NAL) (https://i5k.nal.usda.gov/Manduca_sexta) (Kanost et al., 2016). The RNA-seq datasets were downloaded from NCBI SRA database with accession number listed in Table S1, or previously acquired from Dr. Gary Blissard at Cornell University. Trimmomatic (0.32) (Bolger et al., 2014), Samtools (1.3.1) (Li et al., 2009), Bowtie2 (2.2.6) (Langmead and Salzberg, 2012), TopHat (2.0.12) (Kim et al., 2013), Cufflinks (2.2.1) (Trapnell et al., 2012), STAR (2.5.2a) (Dobin et al., 2013), TransDecoder (3.0.0) (https://github.com/TransDecoder), BLAST+ (2.2.30) (Camacho et al., 2009), RSEM (1.2.29) (Li and Dewey, 2011), tRNAscan-SE (1.3.1) (Lowe and Eddy, 1997) were downloaded from their official sites and installed on a local supercomputer. MeV (Multiple Experiment Viewer 4.9.0, http://mev.tm4.org/) and Cluster 3.0 (by Michael B. Eisen) were installed in a local computer. The MCOT 1.0 gene models were generated in our previous study (Cao and Jiang, 2015).

2.2 Reads alignment and Generation of Cufflinks 4.0

Reads from 67 libaries were first trimmed with Trimmomatic to remove adaptors and low quality bases with the setting "SLIDINGWINDOW:4:20 LEADING:10 TRAILING:10 MINLEN:50". Trimmed paired and non-paired reads in each library were aligned to the genome with TopHat. Cufflinks and Cuffmerge was used to generate and combine GTF files to make the final gene models, Cufflinks 4.0 in same method as previously described

(Cao and Jiang, 2015). Cufflinks 4.0 GTF file was used to build the genome for STAR alignment. Trimmed reads were also aligned to the genome in the 2-pass mapping mode to insure the maximum alignment. Unmapped reads were stored in individual libraries for further analysis.

2.3 Reads aligned to mitochondria genome, mRNA, non-coding genes and rRNA genes

Gene models in Cufflinks 4.0 were analyzed and separated to 4 groups, mitochondria, mRNA, non-coding and rRNA genes. Basically, gene models located in the mitochondrial sequence of the genome were from mitochondria. Scaffolds AIXA01032915.1, AIXA01021581.1, AIXA01037114.1, AIXA01021582.1 were fragments of ribosomal RNAs genes, and genes models from them were rRNA. A gene would be considered mRNA if any transcript of it can be translated to protein by TransDecoder under the default setting (minimum protein length of 100), otherwise it was considered a non-coding gene. The reads count and FPKM value of each gene were calculated with RSEM. Reads counts of different gene groups were summed and plotted.

2.4 Genome coverage by mapped reads

The number of reads mapped to each scaffold of the genome were extracted using samtools idxstats function. The sequencing depth for each base of the genome in each library were extracted with “depth” function of samtools. Genome transcribed region was obtained by counting the non-zero numbers in each library. Transcripts in Cufflinks 2.0 were transcribed with TransDecoder in the genome-guided mode, which outputs a GTF file with coding sequence (CDS), mRNA, gene and UTR location. The length of a gene was defined as the maximum distance between its exons’ edges. For transcribed or CDS ratio in

Cufflinks 4.0, a base would be considered transcribed if it was within any exon or CDS region of Cufflinks 4.0 (regardless of positive or negative strand of the exon). For scaffolds longer than 200 kilobases (mitochondrial and rRNA scaffolds are shorter than 200 kilobases), mapping depth for each bases were normalized and represented by BPKM (bases per kilobase per million mapped bases) value used in transcriptome paper of *D. melanogaster* (Brown et al., 2014), which is equal to number of bases mapped to one base out of one billion mapped bases. The bases in the genome were sorted based on BPKM value and were divided into 20 groups in each library, which are top 400, 400×2^{n} to $400\times2^{n+1}$ where n equals 0 to 17, and below 104,857,600 which is equal to 400×2^{18}. Since all values in group 20 are 0, only the top 19 groups were used. Average BPKM for each group were calculated in each library. For each group, z-score were calculated across the 67 libraries and used to make the figure. Ratio of bases in each group were calculated and plotted in Fig. 3D.

2.5 Unmapped reads analysis

Unmapped reads from running STAR were blasted against the non-redundant nucleotide database (nt, 2016-07-18 version) of NCBI in a local supercomputer, with E-value threshold setting to 10^{-6} and one hit kept in the hit-table format. The number of reads with BLASTN hits were counted. The matched subject sequences were retrieved from NCBI with the accession number. The reads number mapped to each subject sequences were counted with simple python script in each library. The subject sequences were grouped to 7 categories, which were rRNA (“ribosomal RNA” or “rRNA”), mitochondrion (“mitochondrial” or “mitochondrion”), phage (“phage”), *M. sexta* (“M. sexta”, “manduca” or “sexta”), E. coli ("Escherichia coli”, “e.coli” or “e. coli"), Oryza (“oryza”), and Other.

Keywords were listed behind the group name and subject sequences were grouped by matching these keywords in case-ignored manner in order of the group names listed, which meant that a sequence of rRNA of *M. sexta* would be grouped to rRNA, instead of *M. sexta*. Percentage of reads mapped to each groups were calculated.

2.6 Gene expression calculation

Trimmed reads were aligned to OGS 2.0, MCOT 1.0 and Cufflinks 4.0 in different runs, and gene/transcript expression in different libraries was calculated with RSEM according to its manual. The FPKM value, expected reads count for each gene/transcript were summarized together for analysis.

2.7 Library associated genes and comparison between libraries

The same definition and method were used to identify library-associated genes and to compare different libraries (Li et al., 2014). Basically, z-scores were calculated from the FPKM values of each gene, with the formula $z_i=(x_i-\mu)/s$, where x_i is the FPKM value, μ is average FPKM and s is standard deviation. Genes with z-score over 1.5 and FPKM value over 1 will be considered as associated gene for that library.

The comparisons between different libraries were done by testing the dependence of associated genes. Library X and Y are two samples from a population, and the null hypothesis is that they are independent. Suppose total gene number is n, associated genes in X and Y are x and y, and they share c common associated genes. If X and Y are independent, c equals $(x \times y)/(n \times n)$, and the possibility of observing higher c will decrease as the value of c increases. The chance we observe over c common genes were calculated as

$$P = \sum_{i=c}^{\min(x,y)} \frac{\binom{n}{i}\binom{n-i}{x-i}\binom{n-x}{y-i}}{\binom{n}{x}\binom{n}{y}} = \sum_{i=c}^{\min(x,y)} \frac{x!y!(n-x)!(n-y)!}{n!i!(x-i)!(y-i)!(n+i-x-y)!}.$$

Bonferroni corrected P-value = P-value × number of pairwise comparison

Mapping score = $-\log_{10}$(Bonferroni corrected P-value)

Here, the pairwise comparison is 67×67=4,489. For mapping score over 10, the corrected p-value will be very small, and we can reject the null hypothesis and consider two libraries are dependent. The $\log_2$(mapping score) were calculated and plotted in the figure.

2.8 Library-specific gene expression

FPKM values OGS 2.0, MCOT 1.0, and Cufflinks 4.0 were calculated with RSEM. Z-score were calculated based on the FPKM value. MCOT 1.0 genes with bad or no match with OGS2.0 were considered as MCOT-specific genes, and non-coding genes in Cufflinks 4.0 were defined as those genes which cannot be translated to proteins. OGS2.0, MCOT-specific, and non-coding genes were combined, and Genes with at least one FPKM value over 100 of 67 libraries were selected for hierarchical clustering of the z-score with MeV (4.9.0). The clustered genes were split into three groups based on their source and plotted in individual figures.

2.9 tRNA gene modeling and codon usage

tRNAscan-SE was used to scan the genome to identify tRNA genes under the default setting for seach eukaryotic sequences. Genome-based Codon usage was calculated by adding up codon used of different genes in OGS 2.0 directly. Transcriptome-based Codon usage was calculated by first getting numbers of codons used by different transcripts with

the CDS sequencing from TransDecoder, multiplying these numbers by FPKM value of that transcript, summing up the product according to each codon, calculating the percentage of usage in each library, and averaging percentage across 67 libraries.

3 Results

3.1 Life cycle and origins of 67 RNA-seq libraries

In SRA database, there are 67 RNA-seq runs studying *M. sexta* with public RNA-seq reads available, which are 52 cDNA libraries of different tissues and developmental stages which are described in Table S8 of the genome paper of *M. sexta* (Kanost et al., 2016), 8 libraries of adult head studying sex-biased gene expression (Smith et al., 2014), and 7 libraries of male and female antennae studying chemosensory receptor genes (Koenig et al., 2015). Details about construction of these libraries were described in these papers. We label these libraries with number 1 to 67, of which 33 are paired-end reads and 34 single-end (Fig. 1). For libraries 1 to 52, 11 of them are from head, 8 from fat body, 5 from whole body, 13 from midgut, 3 from Malpighian tubule, 7 from muscle, 3 from testis and 2 from ovary. Typically, there are no biological replicates for 52 libraries, but some samples such as G-L5-W, M-L5-12h, M-L5-preW, M-L5-W were sequenced with both single-end and paired-end sequencing technology. Libraries 53 to 60 are 4 biological replicates for male or female head of day 1 adult. Libraries 61 to 67 are from antenna of larva or adult. From three different studies, these libraries can be divided into four groups, which are Group P and S for 33 Paired-end and 19 Single-end libraries in the genome paper, respectively, and Group H and A for 8 single-end libraries from Head and 7 from Antenna. Reads length in Group

P, S, H, and A are 100, 51, 51 and 94 (table S1). Samples for these libraries were collected from different kind of tissues at different developmental stages (Fig. 1, table S1).

3.2 Overview of 67 RNA-seq datasets

As shown in Fig. 2A, the number of reads in each library varies a lot, ranging from 4.2 million in G-L5-preW-S (Lib. #32) to 73 million in F-L5-preW (Lib. #14). The box-plot Fig. 2B shows the average, median and range of reads number in library group P, S, H and A. Generally, there are many more reads in group P than in S, with average number of reads 37 million versus 7.7 million. The variation of reads number is also big in group P, with Lib. #14 as an outlier, and small in group H and A, as these libraries are biological replicates from same sample types.

We then wanted to have an overview of the origins of reads in each library, including how much reads can be aligned to the genome, what are these mapped and unmapped reads. To reach that goal, reads need to be aligned to the genome first. We re-aligned all these reads and re-assembled the Cufflinks 4.0 model instead of our previous result in MCOT (Cao and Jiang, 2015) for several reasons. First, the newer version of official genome, Msex1.0, is slightly different from the previous version we used, with new IDs for scaffolds and with 3 sequences from mitochondrial. Additionally, we had extra 7 extra RNA-seq libraries to analyze. Finally, we decided to do quality control for reads before the alignment to make full use of reads and to improve gene models.

We trimmed reads with Trimmomatic, and only kept reads longer than 50 after quality control to reduce the chance of non-specific matching when we analyze unmapped reads later on. The survival rates were shown in Fig. 2B and 2C. Library group S has higher

survival rate compared to group P, and survival rates in group H are almost the same. The overall survival rates are higher than 85%, with Lib. #66 and #67 as outliers, with a survival rate around 61%.

We then mapped the trimmed high quality reads to the genome with TopHat, and get the Cufflinks 4.0 according to the manual (Trapnell et al., 2012). Because we were also interested in unmapped reads, we used STAR to map the reads to the genome again. With the help of GTP file generated from Cufflinks and running in 2-pass mapping mode, STAR mapped more reads to the genome, with an increase of nearly 10% for trim-survived reads (Fig. 2C and 2E), with library group P from 82% to 91%, and group A 83% to 96%. Lib. #11 is an outlier, with a mapping rate of 60% by TopHat and 69% by STAR.

For aligned reads, we first noticed that a large portion of reads were aligned to the 3 mitochondrial sequences in the genome after extracting count of reads mapped to each scaffold of the genome with samtools. In the beginning, we defined non-coding genes as Cufflinks 4.0 genes that cannot be translated by TransDecoder, and we found 4 genes non-coding genes with extremely high FPKM values. After blast search, we found these genes were rRNA genes, and they were actually 4 individual scaffolds in the genome. Since the 3 mitochondrial scaffolds also codes 3 individual genes, we finally separate genes in Cufflinks 4.0 to four groups, and got total reads ratios mapping to each group (Fig. 2F). The 33,378 genes in Cufflinks 4.0 include 3 from mitochondria, 4 from rRNA, 14,532 coding and 18,839 non-coding genes. Even through the number of non-coding genes is higher than coding genes in Cufflinks 4.0, as described previously, the non-coding genes were generally shorter than coding genes (Cao and Jiang, 2015), and in RNA-seq data, their contribution to total reads is only about 10% that of coding genes (Fig. 2F).

Surprisingly, the percentage of reads mapped to mitochondria and rRNA genes can be very high. Library group S has high percentage of reads mapped to mitochondria, some with more than 20%, and have almost no rRNA reads, and group P has average 20% rRNA reads, and around 5% mitochondrial reads with big variation and outliers, while groups H and A have both low rRNA and mitochondrial reads with Lib. #63 as an outlier which has 19% mitochondrial reads. Group P and S sequencing samples were prepared separately by different groups using techniques in different places, and were sequenced much earlier comparing to group H and A. The differences between these groups may be explained by different sample preparing methods and the improved mRNA purification techniques over time. We did not see biological explanations for these observations.

3.3 Genome transcription

In different tissue or developmental stages, different parts of the genome were actively transcribed to RNA which can be sequenced with RNA-seq technology. Based on Cufflinks 4.0 models, 51.73% of the genome consists of gene regions, 17.12% can be transcribed to mRNA and 5.31% is protein coding region. Interestingly, based on the mapped reads, the transcribed genome ratio goes up to 63.89%, and in different libraries, the ratios are very different, ranging from 1.54% in Lib. #45 (M-L5-preW-S) to 23.22% in Lib. #49 (T-P-D15~18) (Table S1).

It was reasonable to consider that the mapped ratio of the genome will increase as more RNA-seq bases were aligned to the genome, and the observations supports this idea. The relationship between genome mapped ratio and aligned bases were shown in Fig. 3A. We artificially added two linear regression lines for single-end libraries and paired-end

libraries. Libraries below the lines mean that compared to other libraries, less ratio of the genome is transcribed, which also means that there might be some very highly transcribed bases in the genome, while libraries above the line may have less highly transcribed bases. Testis libraries, especially Lib #49, are very far above the line. This is consistent with the observation that these libraries have less highly expressed genes as reported in the genome paper (Kanost et al., 2016). Overall, the mapped ratio in library group P is much higher than group S, likely because of the much higher number of aligned bases (Fig. 3B).

Genome mapped ratios are very different across different libraries, and despite the total number of aligned bases, the distribution of aligned bases across the genome may also be a big contributor to the variations. To test this hypothesis, we first obtained the sequencing depth for each bases in the genome in each libraries. Because rRNA and mitochondrial reads were over-represented in some libraries as discussed before, we remove bases from scaffolds less than 200kd before the analysis. We then normalized the sequence depth using the BPKM value, and sort the bases according to their sequencing depth, and divided them to 20 groups from high to low. To compare across libraries, z-scores were calculated for each of the 20 groups (Fig. 3C). The ratio of aligned bases in each group are shown in Fig. 3D. As expected, libraries above the regression line in Fig. 3A usually have higher BPKMs in highly transcribed base groups, such as library #14 and #15, they have higher BPKMs in group 1 to 6. For Lib. #5 and #16, which have similar numbers of aligned bases and very different genome aligned ratios, average BPKM of #5 is higher than #16 in the top-transcribed groups, and lower in the less-transcribed groups. Comparing libraries from midgut, we can see the transition from larva to adult. Testis and ovary have higher BPKM in low-transcribed groups, which is consistent with their higher position in Fig. 3A. Large

variations were observed from the ratios of each top-transcribed groups. Groups 1 to 4 are top 3,200 transcribed bases, which have the length of 1 to 2 genes on average, as the average length of genes in OGS 2.0 is about 2,000. This means that, in some libraries, two genes may contribute over 20% even close to 40% of total mRNA bases. Groups 1 to 12 are top 819,200 transcribed bases, which may represent 400 genes, 2.6% of genes in OGS 2.0. On average, they occupy over 63% of aligned bases. Groups 13 to 16 contributes to 32% of aligned bases, and there are about 6,000 genes in them. This result means that genes expression levels vary a lot, with few highly expressed genes contribute the major part of the sequenced RNAs, which is consistent with reports in the genome paper that very few highly expressed genes contribute a large part of total FPKM values (Kanost et al., 2016). Additionally, the variations between different libraries indicates that the highly expressed genes may be very different from library to library. It will be interesting to check those highly expressed and library-specific genes.

3.4 Unmapped reads

When we first tried to model genes with Cufflinks, we noticed that the ratio of reads which could map to the genome was very low, only around 60%. We were curious about these unmapped reads since that time. Here, with our improved method, we could almost totally explain this low mapping rate. First, about 10% of reads were discarded after quality control with Trimmamatic, and a certain percentage of reads were trimmed, which could not be aligned to the genome due to low quality of some bases. Secondly, TopHat might have a relatively high standard to consider a read as mapped, and the mapping rate might be lower without the help of the GTF file which stored the splicing site information. Without the GTF file, some reads near the splice junctions might not be mapped properly

due to the short anchoring sequences, while here STAR were provided not only with the GTF file generated by Cufflinks, but also allowed running in the 2-pass mode (Dobin et al., 2013). As a result, STAR mapped about 10% more reads comparing to TopHat. Finally, the mitochondrial sequences were not included in the previous version of the genome, and as described before, on average, nearly 7% of TopHat mapped reads were mapped to mitochondrial DNA. Still, on average, 7.4% of trimmed reads cannot be mapped to the genome, and in library #11, the unmapped rate was 31% even after these settings.

The unmapped reads from STAR were blasted to non-redundant nucleotide (nt) database, the subject genes were grouped to four and the number of reads matched to each gene groups were counted. One thing to note was that nt did not include sequences from OGS 2.0. If the unmapped rate by STAR was higher, the ratio of reads with blastn match would generally be higher, and paired-end libraries had higher unmapped rate and higher ratio of reads with blastn match, with libraries #48 and #50 from testis as outlier, and these un-matched reads might be originated from the W chromosome, which was not included in the official genome of M. sexta (Fig. 4A). Reads with no blastn matches might be AT rich sequences from poly-A tail of mRNA, or from un-sequenced part of the genome.

We divided the blastn target sequences to 7 groups after checking them carefully. The total number and composition of unmapped reads with blastn match are shown in Fig. 4B. It turned out that rRNAs were the major part (80%) of these reads (Table S2), and library group P had a higher ratio of rRNA of unmapped reads, just like they had higher rRNA ratio in mapped reads (Fig. 2G), and library #19 was an outlier which contained a lot of phage sequences. These rRNAs were mostly from other lepidopteran species. Phage sequences, majorly Enterobacteria phage phiX174, which were reported as a positive

control in DNA sequencing, accounted for 6.87%, and their ratios in different libraries were very different. For mitochondrion group, it turned out that majority reads were mapped to a more complete version of *M. sexta* mitochondrion in NCBI. Oryza group contained sequences from different plants, and they were only identified in midgut of larva, not adult. This may be explained by the fact that during wandering stage, all the content in the gut will be eliminated from the larva, and the adult of M. sexta is fed on different food. However, it was hard to explain that library G-L4-0h had few Oryza reads. G-L5-W-S and G-L5-W were supposed to be from the same sample. However, Oryza reads were only identified in single-end libraries. Group E. coli reads were also much higher in the gut, possibly related with the function of the gut. However, we cannot rule out the possibility of contamination, as they were also high in library #53 to #60 which were from the head. 2.79% of reads matched to sequences of *M. sexta*, including lysozyme, apolipophorin protein and other previously reported genes. Maybe these genes were not well sequenced in the genome, or these reads cannot map to the genome due to higher variation from single-nucleotide polymorphism. Group other included sequences various sources, including other lepidoptera species, different bacteria and even human beings. Result from our partner and this study show that no viral sequences were identified other than the phage genome, though it was believed that insects were commonly infected with various viruses and next generation sequencing can help identify insect virus (Liu et al., 2011).

3.5 Comparison between different libraries

Genes were selectively expressed in different tissues at different developmental stages, and result in comparing genome transcription depth implies that different libraries may have very different gene expression patterns. We followed the definition of library-associated genes as genes with FPKM over 1 and Z-score over 1.5 as in the previous study comparing *D. melanogaster* and *C. elegans* RNA-seq libraries (Li et al., 2014). Based on this standard, 15,289 out of 15,543 genes are associated with some libraries, and those un-associated genes are all very lowly expressed genes (FPKM <1). The number of associated genes in each library ranges from 200 to 3,000. Early egg, adult pupa and fat body generally have more associated genes and three libraries from testis have the highest number of associated genes. High expressed associated genes (FPKM >100) are mostly proportional to the number of associated genes (Fig. 5B). Table S3 stores the number of associated genes in each libraries and shared associated gene numbers between different libraries.

We used the same strategy to compare different libraries from *M. sexta* by calculating the mapping score (Li et al., 2014) summarized in Fig. 5A. We improved the heatmap by including mapping scores up to 100, compared to up to 10 in the reference paper, and we also included $\log_2$(mapping score) value as single letter in the figure. First, as expected, mapping scores close to the diagonal line were the highest, indicating libraries of closer developmental stage from same tissue types are more similar to each other. Secondly, there are square-shaped regions of different size along the diagonal line, with some squares share common elements in diagonal line and some not. For instance, library #1 to #3 and library #3 to 5 from head form two different squares, and they do not share a common element, while squares of library #32 to #36 and #34 to #37 from the midgut share several elements. This difference may be explained by the fact that starting from late 4th instar, the head of

larva already changes a lot to prepare for pupa stage. The gene expression profiles in 5th instar, wandering stage, pupa stage and adult stage are very different, maybe because all gut content was cleaned and liquid was secreted in wandering stage and a lot of cells wre disrupted and new tissues were grown in the pupa stage. For libraries from the same tissue/organ/body part, if they are too far away in terms of developing stage, they will share much less associated genes and behave like independent in this mapping score heatmap. Third, libraries share common tissues may have higher similarity, and they are mostly from the same development stages. Library #53 to #60 are from day 1 adult, and they show high similarity with library #9 and #10, which are from day 1 and day 2 adult. Antenna libraries, #61 to #67, both larva and adult have some level of similarity, but only libraries from adult antennae show similarity only to female head (library #53 to #56), and the only male antenna library has lower similarity with male head than female head. Maybe the male and female antennaes are very similar, while the heads are very different. The whole body libraries, #20 to #24, show more similarity with libraries from larva midgut, head and muscle. This similarity is reasonable as the major task for larva is thinking about how to move to eat as much as possible. Ovary libraries, O-P-D15~18 and O-A-D1 are very similar to fat body libraries F-P-D15~18 and F-A-D7~9, and surprisingly, F-A-D1~3 are more similar to MT-A-D1 and MT-A-D3. One possible explanation is that fat body is a very large organ or tissue, fat body from different parts of the body were used in these libraries. Finally, sequencing methods have obvious influence in determining library similarity. As mentioned before, library group P have more rRNA reads while group S have more mitochondrial reads, and the read numbers in group P are much higher than in group S. Library #42 and #43, #44 and #45, #46 and #47 had sequenced same sample with two

technologies, paired-end and single-end. Only the single-end libraries have high similarity with library #5 to #7, which were also single-end libraries, and there are other examples. This difference may be caused by the sequencing technique itself, or the way each sample was prepared before the sequencing, or simply due to the influence of sequencing depth.

3.6 Top expressed genes in different libraries

Different genes were very differently expressed in different libraries. The highly expressed genes, which contributes the major part of the RNA-seq reads and FPKM values, usually play vital functions. To limit the total number of genes for manually checking, we only included top 3 expressed genes in each library. After removing duplicate genes, 69 genes in OGS 2.0 were found to be the top 3 expressed genes in at least one of the libraries. Their expression level and descriptions were shown in Fig. 6. Of the 69 genes, some are housekeeping genes highly expressed in almost all libraries, including ribosomal proteins and energy metabolism related proteins. Another two big groups are odorant binding/chemosensory proteins and cuticle proteins, which can be specific or non-specific in larva and adult. Muscle libraries are from tissues with skin, and this can account for the high level of cuticle proteins in muscle libraries.

The other tissue specific proteins include digestion related proteins which are highly expressed in whole body and gut, serine protease 102 in O-P-D15~18, titin in MT-A-D1 and MT-A-D3, histone H2B and histone H4 in W-E-3h-S and W-E-Late-S, circadian clock-controlled genes in H-L5-preW-S, antimicrobial peptides including diapausin and lysozyme, and etc. The development of insects is controlled by diverse clocks, including the circadian clock (Numata et al., 2015), and the fact that circadian clock-controlled

proteins were extremely highly expressed in heads of pre-wandering stage larva indicates that protein-level control plays a vital role in wandering behavior and development of insects. A groups of diapausins of *M. sexta* were identified and reported to have antifungal activity (Al Souhail et al., 2016; He et al., 2015). They received the name diapausin for their diapause-specific expression when first identified in leaf beetle (Tanaka et al., 2003). These may explain the fact that two diapausins are extremely highly expressed not only in fat body, major resources for antimicrobial peptides, but also in the head, controlling center for diapause behavior.

We also noticed that library #63 has very high expression of actin and other muscle proteins, different from its biological replicate library #61 and #62. Gene Msex2.15420 is extremely highly expressed in many libraries, and though no homolog sequences were identified, we consider it as ribosomal RNA due to its high amount and FPKM values proportional to the ratio of rRNAs in different libraries. Msex2.13838 seems a short non-coding gene, and it is highly expressed in adults, indicating that non-coding genes are regulated similar to coding genes. There are six other uncharacterized proteins which show high specificity to some libraries. Functional studies of these genes may help understanding biological behavior of this model insect.

3.7 Library-specific expression of genes

Genes were differently expressed in different libraries. Theoretically, FPKM values are proportional to mRNA levels inside the cell, and the uncertainty for higher FPKM values is smaller. To have an overview of library-specific gene expression, we made a heat map with z-score of only those high-expressed genes with at least one FPKM value over 100

(Fig. 7). We manually grouped these genes to 22 cluster groups based on their expression patterns.

Clearly we can see many of these genes were very differently expressed in different libraries. 551 genes in cluster 1 are more highly expressed in all three testis libraries, while about 341 genes in cluster 2 are either highly expressed in T-P-D3 or T-P-D15~18 and T-A-D1~3. Cluster 4 are genes expressed in O-A-D1, adult ovary; cluster 9 in 3-hour egg; cluster 12 in larva midgut, which includes a lot of digestive serine proteases (data not shown); cluster 17 in pre-wandering head; clusters 20 and 21 in adult and larva antennae, respectively. Non-coding genes and MCOT specific genes show similar expression patterns (Fig. S1).

It's not easy to describe genes in each cluster group, and most of these genes were not studied, nor is there well-defined functions for each of them. Besides, generally different genes have different functions, so it might be hard to find a common term to describe diverse genes. We did a gene ontology enrichment assay for cluster 1 with Blast2GO (Götz et al., 2008), and found top 3 most significantly increased GO terms in molecular function were microtubule binding, ATP-binding and protein serine/threonine kinase activity (Table S3).

3.8 tRNA genes and codon usage

Different organisms have different codon preferences. Codon preference in the codon preference database were calculated simply based on the published sequences in NCBI. To get a more precise codon preference data table and to check the relationship between codon preference and tRNA gene numbers, we predicted tRNA genes and calculated codon

preference based on the OGS 2.0 sequence and based on the RNA-seq data (table 1). We can see that codon preference does not have high relationship with tRNA gene numbers. For some codons, the predicted tRNA gene number is 0, such as CTC, while the frequency of CTC in genome and transcriptome is 15.2 and 17.7 per thousand, respectively. The ratio of codon in genome and transcriptome are generally similar. We also calculated codon preference in different libraries based on the RNA-seq data, and did not observe clear global codon preference change. We also tried to check tRNA gene expression levels by calculating BPKM values of tRNA gene regions. However, these regions were almost not mapped by any reads, which means that tRNAs were cleaned from the sample and not sequenced in these libraries (data not shown).

4 Discussion

We did a thorough analysis of all currently public available RNA-seq data for *M. sexta*, described the quality, content of the reads. It is surprising that some libraries have high amount of rRNA and mitochondrial reads, which may significantly influence BPKM value of genome bases and FPKM value of genes. It reminds researchers to do the experiment carefully, to reduce these kinds of contamination as much as possible. Also, when calculating gene expression levels, we recommend to remove rRNA genes and mitochondrial genes, which might have too high FPKM value and will lower FPKM values for other genes. The dissection of different tissues aids in finding tissue-specific genes and in elucidating gene functions. It is very important to prepare the sample well, as this determines the RNA-seq quality.

Based on the RNA-seq data, up to 63.89% of the genome may be transcribed, which is consistent with the finding that 85% of human genome can be transcribed (Hangauer et al., 2013), and up to 51.73% of the genome is Cufflinks 4.0 gene regions, which is a relatively compact genome with not so many inter-gene regions. We calculated gene expression levels based on different gene models, including OGS 2.0, Cufflinks 4.0 and MCOT 1.0. Based on the clustering assay to check library-specific gene expression, we can clearly see that majority of highly expressed genes are very library-specific, some of which become extremely high only in one library. Part of the reason is that with a big body size, *M. sexta* is easy to dissect, and enough RNA can be extracted from a few insects which reduces the variation from differences of insects, and the relative long life cycle ensures that these insects are in very close developmental stage. Of course single cell sequencing will be good to reducing this problem, but this technique is still too expensive for most researchers. These genes are tightly regulated, and the cellular machinery are very specific and efficient in regulating their expression. Given the fact less inter-gene bases in the genome, *M. sexta* may be also a good model for studying transcription elements and factors regulating the development, which may also help study development of other insects and help control pests.

Since different genes are very differently expressed in these libraries, and gene expression in these diverse libraries can provide vital information with gene function, we suggest researchers look at information provided with RNA-seq data of genes they are interested in advance, such as gene expression, possible alternative splicing and co-regulated genes. Gene alternative splicing also plays important role in gene regulation, though we did not talk about that in this paper. It is almost impossible to look at the genes one by one. Thus,

it is very important to provide user friendly resources so anyone can check information they are interested in easily.

Egg, larva, pupa, and adult are three distinct stages for holometabolous insects. Our comparison across libraries clearly show that from gene level, they are very different. Different tissues have different genes expressed, and this is clearly supported by the library-specific gene expression heat map. This means genes can be separated to different small groups. Together with domain structure, and protein level information from mass spectrometry, gene functional study will be accelerated.

One of the highest expressed gene, Msex2.15420 is very likely a ribosomal protein, though it can be translated to a protein by TransDecoder. However, we did not found any homologous sequences in NCBI nt or nr database. It is hard to imagine that we do not have enough knowledge for rRNAs. Additionally, some of the highest expressed genes remain unknown, with no homolog sequences or homolog sequences un-studied. These highly expressed, and tissue-specific unknown genes are good targets for future research. What's more, although reads from non-coding genes account for around 10% of coding gene, we did see some highly-expressed and very tissue-specific non-coding genes. While function of most of them are unknown, they might be interesting research area in the future.

5 Summary

We comprehensively studied all current RNA-seq reads for *M. sexta*, checked the amount of rRNA, mitochondrial, mRNA and non-coding reads. We also explained the source for unmapped reads for all these libraries. We compared transcription activity from the genome view, and did similarity comparison across 67 libraries. We provided gene expression level

based on different gene modeling programs, and found that with so many tissue- and time-specific libraries, most genes are expressed in a library-specific manner. This information will greatly help experimental design of future RNA-seq work and basic research of *M. sexta* and other insects.

Acknowledgements

This study was supported by NIH grant GM58634. The *Manduca* Genome Project, which provided Msex 1.0, OGS 1.0, OGS 2.0, Cufflinks 1.0, and RNA-Seq datasets, was funded by DARPA (Gary Blissard, Boyce Thompson Institute) and NIH grant GM41247 (Michael Kanost, Kansas State University). This work was approved for publication by the Director of Oklahoma Agricultural Experimental Station, and supported in part under project OKLO2450 (to H. Jiang). Computation for this project was performed at OSU High Performance Computing Center supported in part through NSF grant OCI-1126330.

Tables

Table 1. Codon usage in *Manduca sexta*

	T					C					A					G					
	codon	AA	freq. G[*1]	freq. T[*2]	tRNA #[*3]	codon	AA	freq. G	freq. T	tRNA #	codon	AA	freq. G	freq. T	tRNA #	codon	AA	freq. G	freq. T	tRNA #	
T	TTT	F	13.7	11.0	1	TCT	S	12.6	12.3	19	TAT	Y	12.7	10.4	1	TGT	C	9.1	6.1	0	T
	TTC	F	20.8	24.8	26	TCC	S	11.3	13.4	0	TAC	Y	19.0	22.6	28	TGC	C	11.9	10.5	21	C
	TTA	L	14.5	10.7	14	TCA	S	12.8	9.8	9	TAA	-[*4]	0.9	2.0	0	TGA	-	0.6	1.2	1[*5]	A
	TTG	L	16.7	15.6	16	TCG	S	13.9	9.2	15	TAG	-	0.4	0.6	3[*6]	TGG	W	11.0	9.2	8	G
C	CTT	L	10.3	11.7	15	CCT	P	13.3	14.1	28	CAT	H	10.7	9.0	0	CGT	R	6.6	9.3	24	T
	CTC	L	15.2	17.7	0	CCC	P	11.9	16.1	0	CAC	H	15.0	14.1	23	CGC	R	13.9	13.1	0	C
	CTA	L	9.2	7.0	6	CCA	P	15.0	13.4	17	CAA	Q	19.1	17.1	23	CGA	R	7.1	4.9	14	A
	CTG	L	22.5	20.1	18	CCG	P	16.5	10.4	11	CAG	Q	19.6	20.2	20	CGG	R	7.3	4.7	0	G
A	ATT	I	15.5	15.4	27	ACT	T	14.9	13.2	23	AAT	N	22.4	19.0	1	AGT	S	12.2	9.6	1	T
	ATC	I	17.1	24.0	0	ACC	T	13.1	15.4	4	AAC	N	24.3	25.8	41	AGC	S	13.7	11.9	16	C
	ATA	I	19.3	12.0	11	ACA	T	16.3	14.5	28	AAA	K	34.1	33.6	22	AGA	R	13.0	10.9	9	A
	ATG	M	22.6	22.3	44	ACG	T	14.0	9.3	11	AAG	K	27.8	40.5	26	AGG	R	9.8	10.9	13	G
G	GTT	V	13.3	14.9	24	GCT	A	17.5	23.3	33	GAT	D	25.2	23.8	6	GGT	G	13.6	19.7	1	T
	GTC	V	14.1	17.8	0	GCC	A	17.8	24.8	0	GAC	D	29.8	32.5	54	GGC	G	21.0	23.3	32	C
	GTA	V	12.1	12.8	12	GCA	A	14.3	14.3	23	GAA	E	34.2	34.1	33	GGA	G	14.7	17.1	18	A
	GTG	V	25.3	24.7	24	GCG	A	23.0	17.5	22	GAG	E	31.0	32.4	28	GGG	G	7.9	7.0	3	G

*1, freq. G: frequency per thousand number based on transcripts CDS sequence. *2, Freq. T: frequency based on RNA-seq data. *3, number of tRNA genes with corresponding anticodon. *4, - stands for stop codon. *5, Seleno-Cysteine tRNA gene. *6, suppressor tRNA gene.

Table S2 Umapped reads with blastn match in each group

	Sum of unmapped reads with blastn match	**Percentage (%)**
rRNA	76,843,832	80.23
phage	6,582,176	6.87
Other	5,271,964	5.50
mitochondrion	1,714,581	1.79
E. coli	1,605,683	1.68
M. sexta	2,675,232	2.79
Oryza	1,080,110	1.12
Sum	95,773,578	100

Figures

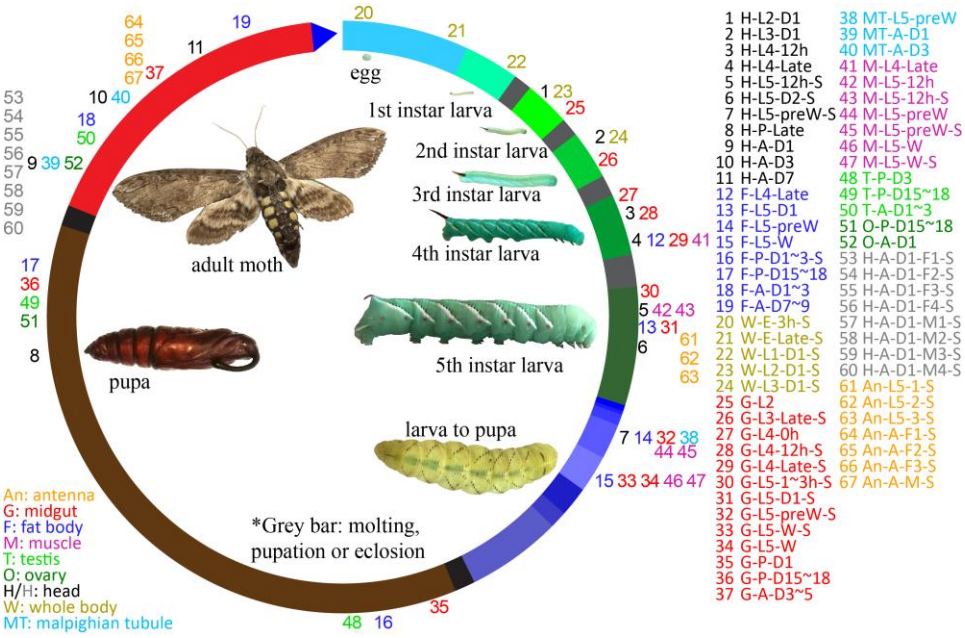


Fig. 1. **Life cycle and public available RNA-seq data sets of *M. sexta***. Bars in the circle represent different developing stages of M. sexta and are proportional to time length of the insect raised with artificial food described in previous publication (Reinecke et al., 1980). For library names, the first part indicates the type of tissue that the libraries are from, and were labeled in different color shown in the figure. The second part, L for larvae, P for pupa, and A for adult. The third part, D for day, h for hour, preW for pre-wandering stage, W for wandering stage, M for male, and F for female. The –S in the end represents that reads were single-end, otherwise paired-end. The cDNA libraries represent the following tissues and stages: head [**1**. 2nd (instar) L (larvae), d1 (day 1); **2**. 3rd L, d1; **3**. 4th L, 12h (hour); **4**. 4th L, late; **5**. 5th L, d0.5; **6**. 5th L, d2; **7**. 5th L, pre-W (pre-wandering); **8**. P (pupae), late; **9**. A (adults), d1; **10**. A, d3; **11**. A, d7], fat body (**12**. 4th L, late; **13**. 5th L, d1; **14**. 5th L, pre-W; **15**. 5th L, W; **16**. P, d1-3; **17**. P, d15-18; **18**. A, d1-3; **19**. A, d7-9), whole animals [**20**. E (embryos), 3h; **21**. E, late; **22**. 1st L; **23**. 2nd L; **24**. 3rd L), midgut (**25**. 2nd L; **26**. 3rd L; **27**. 4th L, 0h; **28**. 4th L, 12h; **29**. 4th L, late; **30**. 5th L, 1-3h; **31**. 5th L, 24h; **32**. 5th L, pre-W; **33-34**. 5th L, W; **35**. P, d1; **36**. P, d15-18; **37**. A, d3-5;), MT (**38**. 5th L, pre-W; **39**. A, d1; **40**. A, d3), muscle (**41**. 4th L, late; **42-43**. 5th L, 12h; **44-45**. 5th L, pre-W; **46-47**. 5th L, W), testes (**48**. P, d3; **49**. P, d15-18; **50**. A, d1-3), and ovaries (**51**. P, d15-18; **52**. A, d1), head [**53-56**. A, d1, F (Female); **57-60**, A, d1, M (male)], antenna (**61-63**, 5th L; **64-66**, A, F; **67**, A, M).

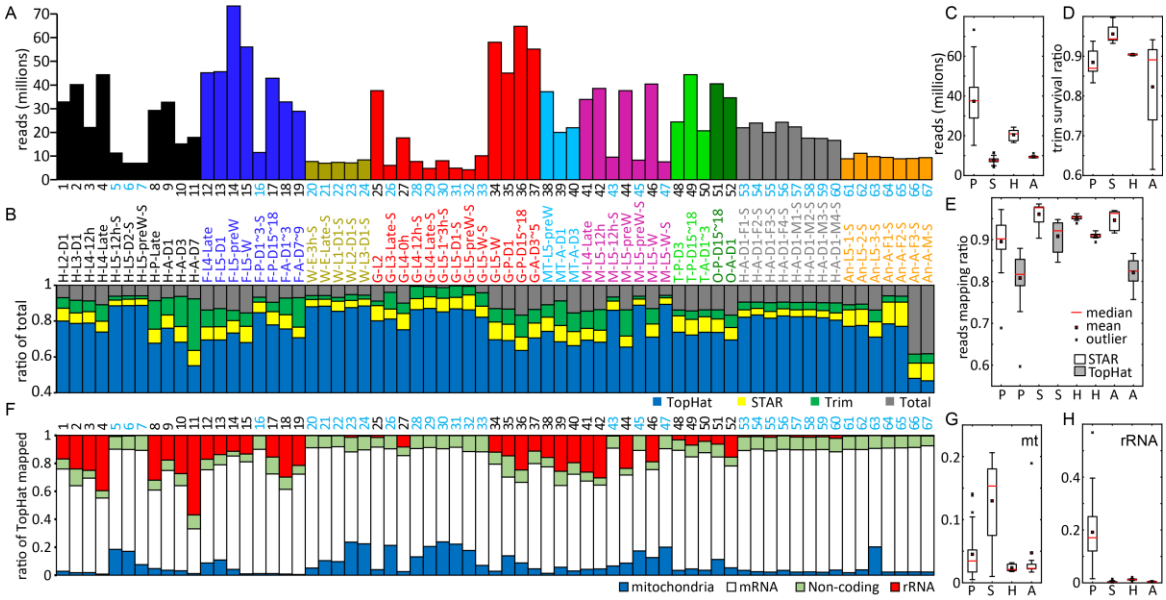


Fig. 2. **Overview of 67 cDNA libraries**. **A**, reads number in each library. Color represent tissues source of mRNA, and 1 to 67 represent library number (black for paired end reads, cyan for single end reads), the same as in Fig. 1. **B**, up-boundaries represent ratio of reads after different treatment, including trimming, mapping with STAR and TopHat. Total reads in each library were set to 1. **C**, **D**, box-plot of reads number and trimming survival rate in each library categories, respectively. **E**, mapping rates of trimming survival reads by STAR and TopHat. **F**, ratio of TopHat mapped reads mapping to mitochondria, mRNA, non-coding and rRNA genes. **G**, **H**, box-plot of ratio of mapping ration to mitochondria and rRNA of different library groups. For library categories, P for 33 paired-end of 52 libraries sequenced together with genome project, S for 19 single-end of 52 libraries, H for 8 libraries from head and A for 7 libraries from antenna in two different individual studies. Library names were the same as in Fig. 1.

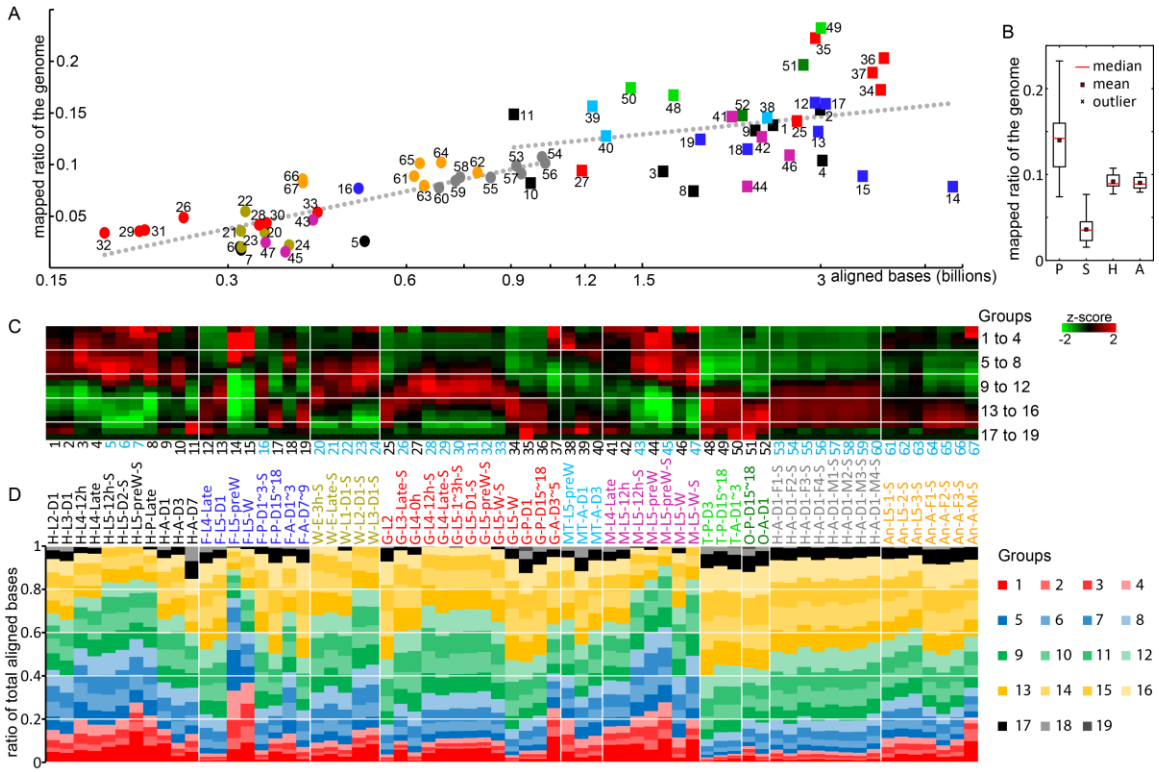


Fig. 3. **Reads aligned to genome**. **A**, relation between mapping ratio of genome and aligned bases by TopHat. Each symbol in the figure represents one library, with their library numbers labeled. Square for paired-end libraries, Circle for single-end libraries. Two lines were linear regression of paired-end libraries and single-end libraries. **B**, box-plot of mapping ratio of genome in different library groups. P for paired, S for single, H for head, A for antenna, as described in results. **C**, heatmap of z-score in each base groups. Bases in the genome were sorted based on BPKM value first. Group 1 to 19 are top 400 bases, 400×2^{n} to $400 \times 2^{n+1}$ where n equals 0 to 17. Heatmap is colored based on the z-score of average BPKM in each group. **D**, ratio of total aligned bases in each base groups.

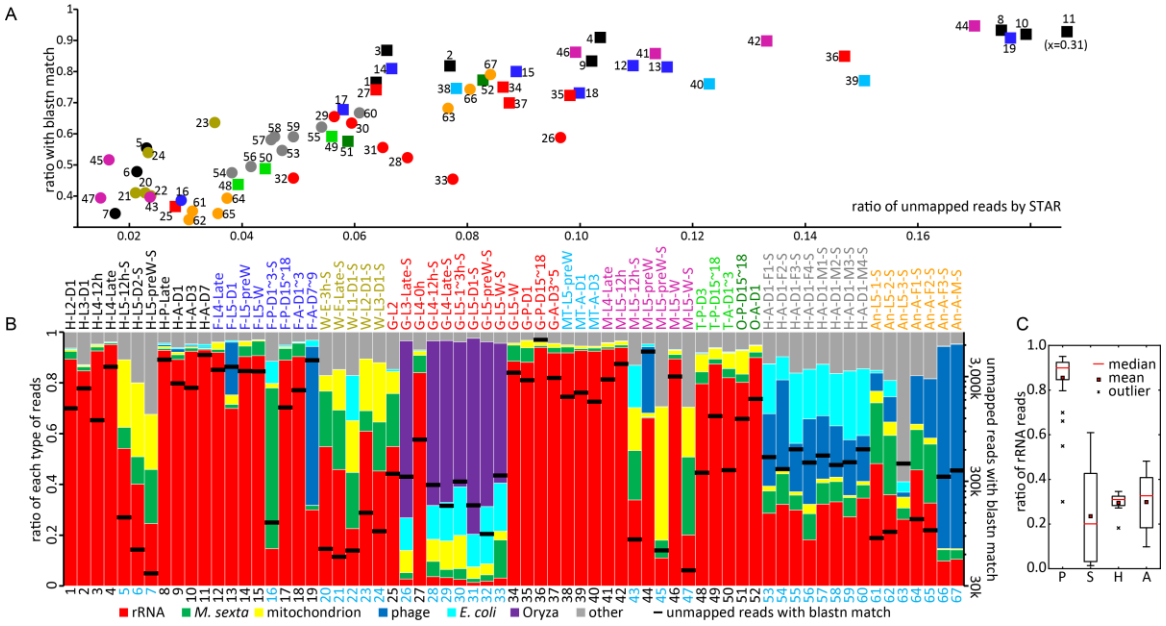


Fig. 4. **Unmapped Reads**. **A**, relationship between ratio of unmapped reads with blastn match and ration of unmapped reads by STAR. **B**, distribution of unmapped reads with blastn match. Different colors of bars represent different types of reads. Lines in the figure show the number of unmapped reads with blastn match. **C**, box-plot of ratio of rRNA-reads in unmapped reads with blastn match. Number in the figure represents library number, the same as in Fig. 1. Square for paired-end, circle for single-end. Color indicating the library type, the same as in Fig. 1 and Fig. 3.

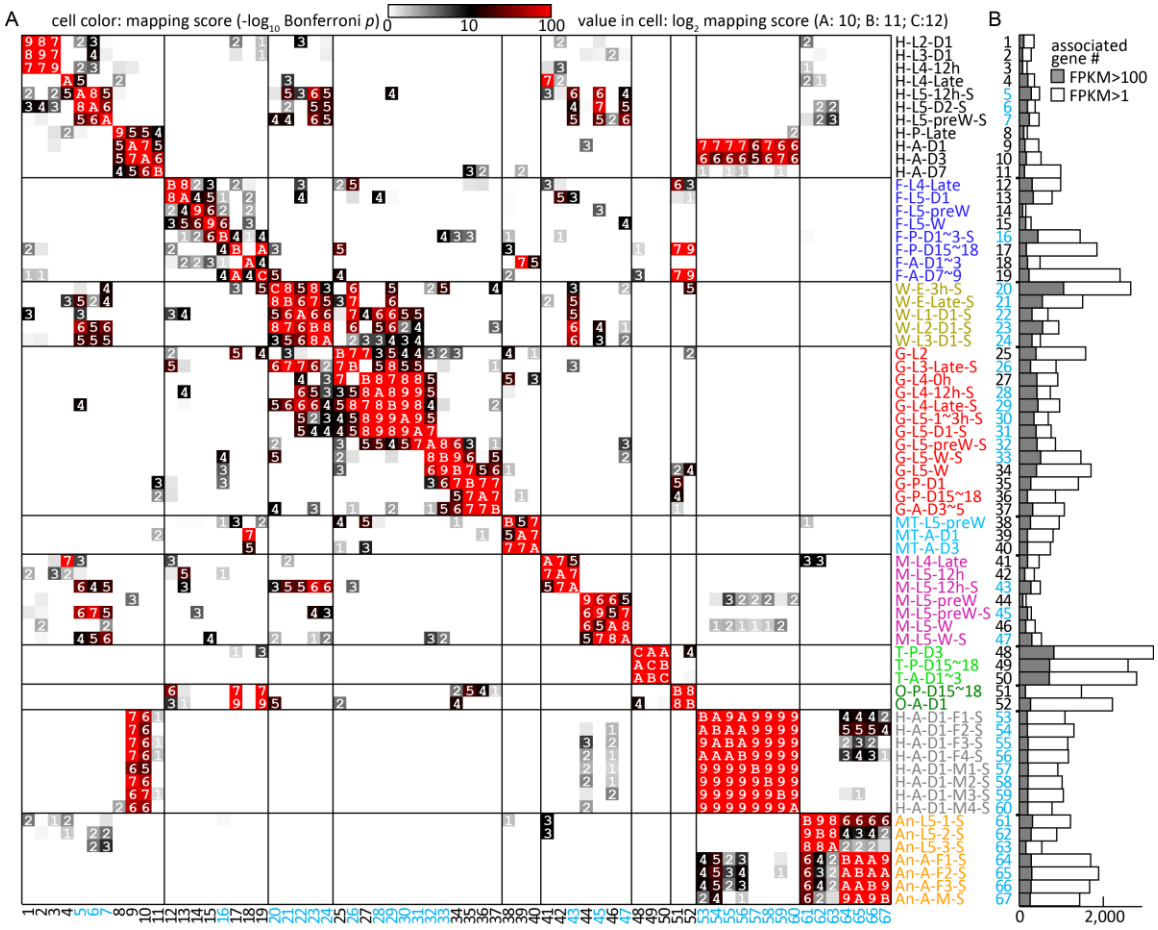


Fig. 5. **Library-associated genes and comparison of different libraries**. **A**, mapping scores of different library pairs. Values in the cells were $\log_2$(mapping score). A value greater than 4 (mapping score greater than 16) means two libraries were very dependent. **B**, number of associated genes in each library. Grey bars were associated genes with FPKM value greater than 100.

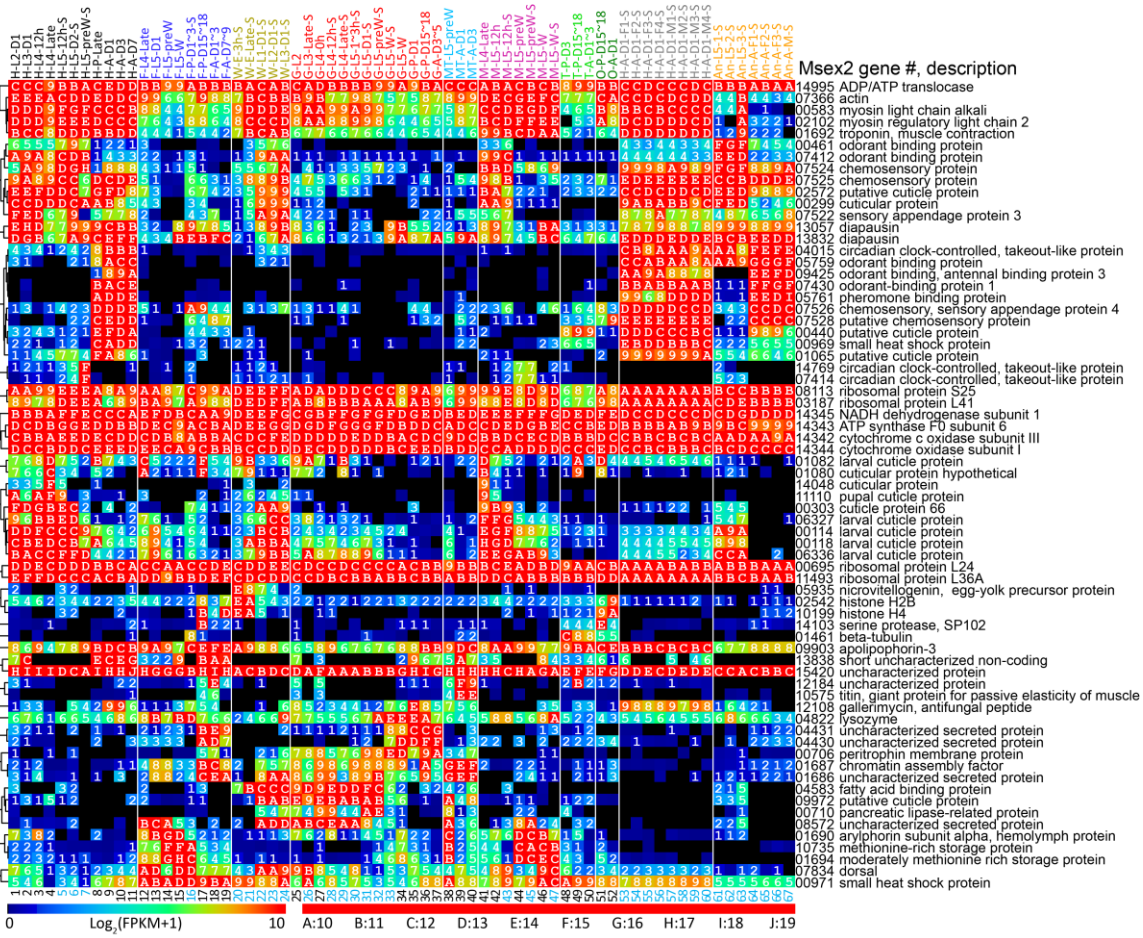


Fig. 6. **Expression profile of 69 highly expressed genes in 67 libraries**. Top 3 expressed genes in each library were included, and their mRNA levels were represented by $\log_2$(FPKM+1) values, are shown in the rainbow gradient color in heat map. Library names and descriptions were the same as in Fig. 1.

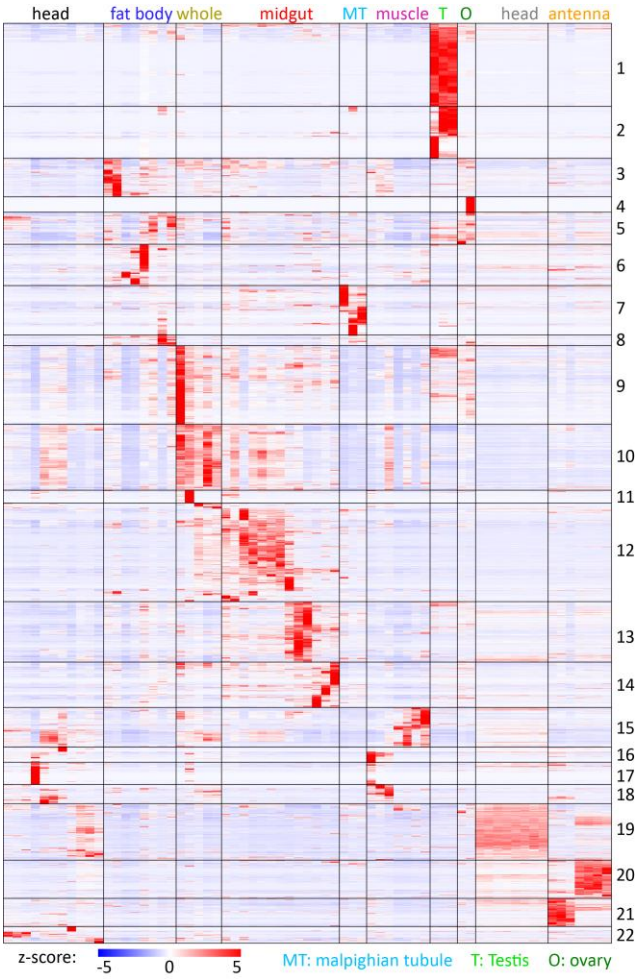


Fig. 7. **Library-specific expression of different genes in OGS 2.0**. Z-scores for high expressed genes were calculated from FPKM values. Genes were clustered based on Z-score and divided to different groups manually based on the expression pattern.

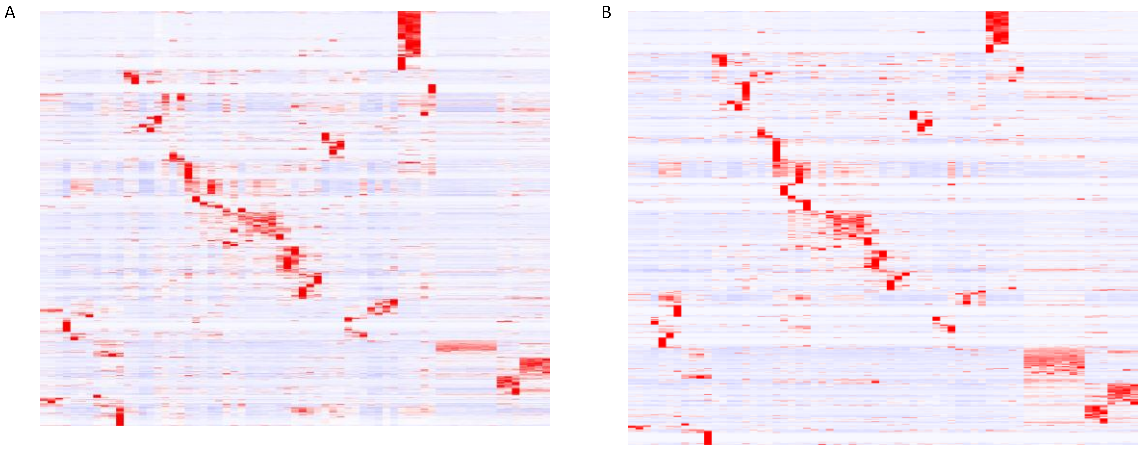


Fig. S1. **Library-specific expression of MCOT 1.0-specific and non-coding genes**. **A**), MCOT 1.0 specific genes. **B**) Non-coding genes.

Reference

Al Souhail, Q., Hiromasa, Y., Rahnamaeian, M., Giraldo, M.C., Takahashi, D., Valent, B., Vilcinskas, A., Kanost, M.R., 2016. Characterization and regulation of expression of an antifungal peptide from hemolymph of an insect, Manduca sexta. Developmental and comparative immunology.

Arrese, E., Soulages, J., 2010. Insect fat body: energy, metabolism, and regulation. Annual review of entomology 55, 207-225.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England) 30, 2114-2120.

Brown, J.B., Boley, N., Eisman, R., May, G.E., Stoiber, M.H., Duff, M.O., Booth, B.W., Wen, J., Park, S., Suzuki, A., 2014. Diversity and dynamics of the Drosophila transcriptome. Nature.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC bioinformatics 10, 421.

Cao, X., He, Y., Hu, Y., Wang, Y., Chen, Y.-R.R., Bryant, B., Clem, R.J., Schwartz, L.M., Blissard, G., Jiang, H., 2015a. The immune signaling pathways of Manduca sexta. Insect biochemistry and molecular biology 62, 64-74.

Cao, X., He, Y., Hu, Y., Zhang, X., Wang, Y., Zou, Z., Chen, Y., Blissard, G.W., Kanost, M.R., Jiang, H., 2015b. Sequence conservation, phylogenetic relationships, and expression profiles of nondigestive serine proteases and serine protease homologs in Manduca sexta. Insect biochemistry and molecular biology 62, 51-63.

Cao, X., Jiang, H., 2015. Integrated modeling of protein-coding genes in the Manduca sexta genome using RNA-Seq data from the biochemical model insect. Insect biochemistry and molecular biology 62, 2-10.

Dittmer, N.T., Tetreau, G., Cao, X., Jiang, H., Wang, P., Kanost, M.R., 2015. Annotation and expression analysis of cuticular proteins from the tobacco hornworm, Manduca sexta. Insect biochemistry and molecular biology 62, 100-113.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T.R., 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15-21.

Götz, S., García-Gómez, J.M., Terol, J., Williams, T.D., Nagaraj, S.H., Nueda, M.J.J., Robles, M., Talón, M., Dopazo, J., Conesa, A., 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic acids research 36, 3420-3435.

Gunaratna, R.T., Jiang, H., 2013. A comprehensive analysis of the Manduca sexta immunotranscriptome. Developmental and comparative immunology 39, 388-398.

Hangauer, M.J., Vaughn, I.W., McManus, M.T., 2013. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. PLoS genetics 9.

He, Y., Cao, X., Li, K., Hu, Y., Chen, Y.R., Blissard, G., Kanost, M.R., Jiang, H., 2015. A genome-wide analysis of antimicrobial effector genes and their transcription patterns in Manduca sexta. Insect biochemistry and molecular biology 62, 23-37.

Hopkins, T.L., Krchma, L.J., Ahmad, S.A., Kramer, K.J., 2000. Pupal cuticle proteins of Manduca sexta: characterization and profiles during sclerotization. Insect biochemistry and molecular biology 30, 19-27.

Jiang, H., Vilcinskas, A., Kanost, M.R., 2010. Immunity in lepidopteran insects, Invertebrate Immunity. Springer, pp. 181-204.

Kanost, M.R., Arrese, E.L., Cao, X., Chen, Y.-R.R., Chellapilla, S., Goldsmith, M.R., Grosse-Wilde, E., Heckel, D.G., Herndon, N., Jiang, H., Papanicolaou, A., Qu, J., Soulages, J.L., Vogel, H., Walters, J., Waterhouse, R.M., Ahn, S.-J.J., Almeida, F.C., An, C., Aqrawi, P., Bretschneider, A., Bryant, W.B., Bucks, S., Chao, H., Chevignon, G., Christen, J.M., Clarke, D.F., Dittmer, N.T., Ferguson, L.C., Garavelou, S., Gordon, K.H., Gunaratna, R.T., Han, Y., Hauser, F., He, Y., Heidel-Fischer, H., Hirsh, A., Hu, Y., Jiang, H., Kalra, D., Klinner, C., König, C., Kovar, C., Kroll, A.R., Kuwar, S.S., Lee, S.L., Lehman, R., Li, K., Li, Z., Liang, H., Lovelace, S., Lu, Z., Mansfield, J.H., McCulloch, K.J., Mathew, T., Morton, B., Muzny, D.M., Neunemann, D., Ongeri, F., Pauchet, Y., Pu, L.-L.L., Pyrousis, I., Rao, X.-J.J., Redding, A., Roesel, C., Sanchez-Gracia, A., Schaack, S., Shukla, A., Tetreau, G., Wang, Y., Xiong, G.-H.H., Traut, W., Walsh, T.K., Worley, K.C., Wu, D., Wu, W., Wu, Y.-Q.Q., Zhang, X., Zou, Z., Zucker, H., Briscoe, A.D., Burmester, T., Clem, R.J., Feyereisen, R., Grimmelikhuijzen, C.J., Hamodrakas, S.J., Hansson, B.S., Huguet, E., Jermiin, L.S., Lan, Q., Lehman, H.K., Lorenzen, M., Merzendorfer, H., Michalopoulos, I., Morton, D.B., Muthukrishnan, S., Oakeshott, J.G., Palmer, W., Park, Y., Passarelli, A.L., 2016. Multifaceted biological insights from a draft genome sequence of the tobacco hornworm moth, Manduca sexta. Insect biochemistry and molecular biology.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S.L., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome biology 14, R36.

Koenig, C., Hirsh, A., Bucks, S., Klinner, C., Vogel, H., Shukla, A., Mansfield, J.H., Morton, B., Hansson, B.S., Grosse-Wilde, E., 2015. A reference gene set for chemosensory receptor genes of Manduca sexta. Insect biochemistry and molecular biology 66, 51-63.

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357-359.

Li, B., Dewey, C., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC bioinformatics 12, 323.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., Genome Project Data Processing, S., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England) 25, 2078-2079.

Li, J.J., Huang, H., Bickel, P.J., Brenner, S.E., 2014. Comparison of D. melanogaster and C. elegans developmental stages, tissues, and cells by modENCODE RNA-seq data. Genome research.

Liu, S., Vijayendran, D., Bonning, B.C., 2011. Next generation sequencing technologies for insect virus discovery. Viruses 3, 1849-1869.

Lowe, T.M., Eddy, S.R., 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic acids research 25, 955-964.

Numata, H., Miyazaki, Y., Ikeno, T., 2015. Common features in diverse insect clocks. Zoological letters 1, 10.

Rao, X.-J.J., Cao, X., He, Y., Hu, Y., Zhang, X., Chen, Y.-R.R., Blissard, G., Kanost, M.R., Yu, X.-Q.Q., Jiang, H., 2015. Structural features, evolutionary relationships, and transcriptional regulation of C-type lectin-domain proteins in Manduca sexta. Insect biochemistry and molecular biology 62, 75-85.

Reinecke, J.P., Buckner, J., Grugel, S., 1980. Life cycle of laboratory-reared tobacco hornworms, Manduca sexta, a study of development and behavior, using time-lapse cinematography. The Biological Bulletin 158, 129-140.

Riddiford, L.M., Hiruma, K., Zhou, X., Nelson, C.A., 2003. Insights into the molecular basis of the hormonal control of molting and metamorphosis from Manduca sexta and Drosophila melanogaster. Insect biochemistry and molecular biology 33, 1327-1338.

Shields, V.D., Hildebrand, J.G., 2001. Recent advances in insect olfaction, specifically regarding the morphology and sensory physiology of antennal sensilla of the female sphinx moth Manduca sexta. Microscopy research and technique 55, 307-329.

Smith, G., Chen, Y.-R.R., Blissard, G.W., Briscoe, A.D., 2014. Complete dosage compensation and sex-biased gene expression in the moth Manduca sexta. Genome biology and evolution.

Tanaka, H., Sato, K., Saito, Y., Yamashita, T., Agoh, M., Okunishi, J., Tachikawa, E., Suzuki, K., 2003. Insect diapause-specific peptide from the leaf beetle has consensus with a putative iridovirus peptide. Peptides 24, 1327-1333.

Tetreau, G., Cao, X., Chen, Y.R., Muthukrishnan, S., Jiang, H., Blissard, G.W., Kanost, M.R., Wang, P., 2015a. Overview of chitin metabolism enzymes in Manduca sexta: Identification, domain organization, phylogenetic analysis and gene expression. Insect biochemistry and molecular biology 62, 114-126.

Tetreau, G., Dittmer, N.T., Cao, X., Agrawal, S., Chen, Y.-R.R., Muthukrishnan, S., Haobo, J., Blissard, G.W., Kanost, M.R., Wang, P., 2015b. Analysis of chitin-binding

proteins from Manduca sexta provides new insights into evolution of peritrophin A-type chitin-binding domains in insects. Insect biochemistry and molecular biology 62, 127-141.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., Pachter, L., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols 7, 562-578.

Xiong, G.H., Xing, L.S., Lin, Z., Saha, T.T., Wang, C., Jiang, H., Zou, Z., 2015. High throughput profiling of the cotton bollworm Helicoverpa armigera immunotranscriptome during the fungal and bacterial infections. BMC Genomics 16, 321.

Zhang, S., Gunaratna, R.T., Zhang, X., Najar, F., Wang, Y., Roe, B., Jiang, H., 2011. Pyrosequencing-based expression profiling and identification of differentially regulated genes from Manduca sexta, a lepidopteran model insect. Insect biochemistry and molecular biology 41, 733-746.

Zhang, X., He, Y., Cao, X., Gunaratna, R.T., Chen, Y.-r.R., Blissard, G., Kanost, M.R., Jiang, H., 2015a. Phylogenetic analysis and expression profiling of the pattern recognition receptors: Insights into molecular recognition of invading pathogens in Manduca sexta. Insect biochemistry and molecular biology 62, 38-50.

Zhang, X., Zheng, Y., Cao, X., Ren, R., Yu, X.-Q.Q., Jiang, H., 2015b. Identification and profiling of Manduca sexta microRNAs and their possible roles in regulating specific transcripts in fat body, hemocytes, and midgut. Insect biochemistry and molecular biology 62, 11-22.

CHAPTER IV

FUNCTIONAL STUDY OF STRESS RESPONSIVE PEPTIDES IN IMMUNITY AND OTHER BIOLOGICAL PROCESSES OF MANDUCA SEXTA

Abstract

Cytokines are important regulators of biological process. Stress responsive peptides (SRPs) are a conserved group of insect cytokine. Eleven SRPs were identified in *Manduca sexta*, including uENF1, uENF2, PP, and SRP 1 to 8. Among them, SRP5, 7 and 8 are lowly expressed in 52 RNA-seq libraries. PAP1 and PAP3 are proved capable of activating proSRP1 and proSRP2. MALDI-MS suggested that the predicted activation site is right. uENF1, uENF2, PP, SRP1 and SRP2 can induce expression of some AMPs, while SRP6 can block the feeding and growth of larvae. The functions of different SRPs are diverse, and this work helps elucidate the role of cytokines in insect immunity and development.

1. Introduction

As main agricultural pests and vectors of many human or animal diseases (e.g., malaria, Dengue fever), insects cause losses up to one fifth of agricultural production in the world and threaten people's life and health. In 2007, $4.3 were spent on insecticides (https://www.epa.gov/sites/production/files/2015-10/documents/market_estimates2007.pdf) and, according to WHO, 438,000 people died of malaria carried by mosquitos in 2015 (http://apps.who.int/gho/data/node.main.A1368?lang=en). Immune system plays key role in defense against pathogens and the study of insect immune system may help us reduce agricultural loss and control disease transmission.

With five larval instars, large body size and hemolymph volume, the tobacco hornworm *M. sexta* is widely used as a model organism to study various insect physiological processes, especially immune-related proteins in the hemolymph (Jiang et al., 2010). Among various immune responses, antimicrobial peptides (AMPs) are key components of the innate immune system, which are evolutionarily conserved weapons against bacteria, fungi and viruses, and widely used throughout the plant and animal kingdoms (Diamond et al., 2009; Imler, 2013; Izadpanah and Gallo, 2005; Pasupuleti et al., 2012). After years of study in *Drosophila melanogaster* and other insects, the Toll and IMD pathways are found to be two main cellular signaling cascades that control the expression of AMPs and other immune responses (Kleino and Silverman, 2013). Other important immune responses include phenoloxidase (PO) activation which induces melanization and kills a wide range of pathogens, and cellular response, such as swallowing of bacteria by plasmatocytes (Eleftherianos et al., 2009; Isaac and Alex, 2012).

In humans, the signaling and development of different immune cells are tightly regulated by various immune-related cytokines, but only a few cytokines were identified and studied in insects. One of the most studied insect cytokines is paralytic peptide (PP), which induces AMP expression (Tsuzuki et al., 2012) and regulates melanization (Ninomiya and Hayakawa, 2007), causes paralysis, blocks larval growth, and induces the plasmatocyte (Yang Wang, 1999). Interestingly, PP can be translated from a tricistronic transcript, which encodes two other proteins, named uENF1 and uENF2, whose function remains unclear. In 2011, a new cytokine, named stress responsive peptide (SRP), was identified in *Spodoptera litura*. SRP was induced under stress conditions, including heat, cold, injury, and infection by microbes and parasites. Similar to PP, it inhibits feeding activity, retards larval growth, and causes plasmatocyte spreading. SRP is more highly expressed in hemocytes and brain than in fat body(Yamaguchi et al., 2012). All these data suggest that SRP and PP are key signaling molecules for humoral and neural regulation of immune or stress responses.

By searching the *M. sexta* genome, we identified eight SRP genes in different regions of the genome. Together with uENF1, uENF2 and PP, 11 cytokines were identified in *M. sexta*. With remarkable differences, the SRPs, uENFs, and PP share some common features in amino acid sequences, and are potential regulators of AMP expression or other innate immune pathways.

2. Literature review

Cytokines can be loosely defined as small signaling proteins which are usually released to influence behavior of surrounding cells through receptors. Nearly all biological processes are regulated by cytokines, including embryonic development, stem cell differentiation, specific or non-specific immune responses, and the aging process (Vilcek and Feldmann, 2004). Several types of best known cytokines, including interferons, interleukins and chemokines, are well-studied for their vital roles in the mammalian immune response, such as cell-to-cell communication between macrophage, neutrophil cells, mast cell, T-cells, and B-cells, and immune cell activation, differentiation and proliferation, and other immune responses (Stenken and Poschenrieder, 2015).

Over 100 different kinds of cytokines have been identified and studied (Arango Duque and Descoteaux, 2014), while only a few of cytokines were studied in any single insect species. Many insect cytokines were first studied in *Drosophila melanogaster*, the most widely used model insect, including Dpp, dawdle, Eiger, GBP, Spätzle, Udp3, and Vago (Clark et al., 2011; Safia et al., 2008; Tsuzuki et al., 2012). Other than those cytokines, a group of bioactive peptides with similar functions were identified in at least six orders of insects (Matsumoto et al., 2012). The first member of these peptides was firstly identified in a wasp parasite lepidopteran species, *Mythimna separata*, and named growth blocking peptide (GBP) for its function in blocking the development of host larva (Hayakawa, 1990). Later on, this GBP peptides were identified and studied in other lepidopteran species, and they were named growth blocking, paralytic, plasmatocyte spreading, or stress responsive peptide (GBP/PP/PSP/SRP) for their diverse functions (Clark et al., 1997; Skinner et al.,

1991; Yamaguchi et al., 2012). Even though they were grouped as one group of peptides, their sequences and functions can be very different. We will use the term stress responsive peptides (SRPs) to represent this group of cytokines as they are activated under certain stress.

2.1 Sequence features of SRPs

The active SRP peptides are usually 22 to 32 residues, follow this loose formula (R/K)-$X_{1/15}$-C-$X_{7/9}$-G-$X_{1/2}$-C-$X_{1/15}$, where R/K indicating the putative activation cleavage site and X represents amino acids other than cysteine, and the two cysteine residues form an intramolecular disulfide bond which determine the basic loop structure of the peptides (Matsumoto et al., 2012). The precursor proteins, typically 60 to 150 residues long, usually have a signal peptide (15 to 23 amino acids), a pro-region with no or few cysteine residues and a relative unique cleavage site such as R/K or R/K-$X_{1/2}$-R/K, and a functional peptide region, a typical structure of pre-pro-proteins. The signal peptide of pre-pro-SRPs indicates that these proteins can be secreted outside the cell. The pro-SRPs are secreted waiting for activation through specific cleavage by extracellular proteases, or further cleaved by intracellular processing enzymes and stored as active peptides in secretory vesicles inside the cell (Hayakawa et al., 1995; Nakatogawa et al., 2009; Wang et al., 1999).

2.2 Paralytic peptides (PPs)

The first SRP peptide in a lepidopteran species, *M. separata*, and named GBP (Hayakawa, 1990). A group of seven peptides purified from hemolymph of other lepidopteran insects, *including M. sexta*, *Spodoptera exigua*, and *Heliothis virescens*, were named paralytic peptides (PPs) for the rigid paralysis behavior observed of larvae injected with those

peptides (Skinner et al., 1991). Later on, a new peptide was identified in another lepidopteran insect, Pseudoplusia includens, and named plasmatocyte-spreading peptide (PSP) for its function in inducing the spreading of plasmatocyte, a key class of hemocytes involved in cellular immunity of insects (Clark et al., 1997). All these peptides turn out from the same group, all start with Glu-Asn-Phe (ENF) with highly conserved sequences, and were named ENF peptides collectively (Strand et al., 2000). We will use PP to represent this group of cytokines.

These lepidopteran PPs have diverse functions. First, they can block the growth of insects, retarding larval body weight gaining and delaying pupal formation. In the first PP-related study in *M. separata*, the parasitic wasp, *Apantales kariyai*, can induce the expression of PP, which somehow would reduce the activity of juvenile hormone (JH) esterase, and thus decrease the level of JH which is important in larva's preparation for pupation (Hayakawa, 1990). The elongated larval stage was beneficial for the growth and development of parasites. Similar growth blocking function of *M. separata* PP were also observed in *Bombyx mori*, by injecting different amount of *M. separata* PP to *B. mori* larvae, but the efficiency was lower comparing with *B. mori* PP. This indicated that while different PPs might work across species for their conservation, the receptors in each species were co-evolved with their PPs and thus more sensitive to their own PPs (Hayakawa and Yasuhara, 1993; Miura et al., 2002).

Secondly, PPs induce paralysis behavior of larvae after injected into hemocoel. This effect was observed in several different insects, including *B. mori*, *M. sexta* and *M. separata* (Ha et al., 1999; Wang et al., 1999). The contraction of muscle is somehow influenced by the

peptides in a dose-dependent manner, and whether this works through the neuron system remains unknown (Ishii et al., 2015).

Third, PPs influence the behavior of plasmatocytes and other hemocytes. Plasmatocytes have similar function as mammalian monocyte or macrophage cells, which involve in phagocytosis and encapsulation of pathogens (Williams, 2007). Plasmatocytes account for 90% of hemocytes in *D. melanogaster*, while in most other studied insects, the most abundant hemocytes are granulocytes, which are characterized by the presence of granules in the cytoplasm, and are also capable of adhering to foreign molecules or pathogens. PP can increase spreading and attaching speed of plasmatocytes in *M. separata*, *P. includens*, *M. sexta*, and *S. litura*. This may explain the loss of plasmatocytes after injection of PP (Wang et al., 1999). Spreading and attaching of plasmatocytes help clot formation to stop bleeding after the insect became injured and promote wound healing. Plasmatocytes become more active after injection of PP, and become more capable to phagocytose co-injected bacteria, *Staphylococcus aureus*, in *B. mori* (Ishii et al., 2010).

Fourth, PPs regulate immune-related gene expression. Antimicrobial peptides (AMPs) are hugely induced after immune challenge, attacking invading pathogens directly and play vital role in insect immunity and (Lemaitre and Hoffmann, 2007). Upon infection, insects also generate reactive oxygen/nitrogen species (ROS/RNS) including NO, hydroxyl and peroxides to kill microbes. Injection of PP increased expression of several AMPs and nitric oxide synthase (NOS) expression in *B. mori* (Ishii et al., 2013; Ishii et al., 2010). Tetraspanin, a hemocytes surfaced protein involved in encapsulation, was also upregulated by PP, which may enhance phagocytosis by plasmatocytes (Ishii et al., 2010).

Finally, PPs have mitogenic activity. PPs from *B. mori* and *P. separata* enhanced nucleotide consumption by cultured cells in vitro, a sign of increased mitogenic activity (Hayakawa and Ohnishi, 1998; Tsuzuki et al., 2012). Surprisingly, *in vivo*, injection of 1 or 10 pmol of PP to *M. separata* larvae has the opposite function, one promoting and one inhibiting the body weight gain, respectively. The mechanism behind this controversial function is still unknown.

2.3 uENF peptides

Most eukaryotic mRNAs are monocistronic and only a few are dicistronic. Surprisingly, the PP of *M. separata* is encoded by two different transcripts, one monocistronic and one tricistronic. The shorter one codes PP only, and the tricistronic one codes PP and two other upstream proteins, named uENF1 and uENF2 (Kanamori et al., 2010). A similar tricistronic transcript is found in each of the well-sequenced lepidopteran species, and is a conserved gene in lepidopteran species. In vitro experiments suggested that all three proteins can be translated, and the mechanism was context-dependent leaky-scanning of ribosome. The amount of the longer form transcript was higher in embryo, while the shorter form dominated in most other developmental stages. Both uENF1 and uENF2 sequences are conserved among different lepidopteran species, and preliminary research shows that uENF1 promotes while uENF2 inhibits spreading of plasmatocytes. More study is needed to elucidate their functions.

2.4 SRP peptides

The first cytokine formally named SRP was discovered in *Spodoptera litura*. Similar to of PP, this *S. litura* SRP was also identified in moth larva parasitized by a wasp (Yamaguchi

et al., 2012). This study shows that *S. litura* SRP have similar functions as PPs, including blocking growth of larvae and inducing spreading of hemocytes. The expression of this SRP can be induced by parasitization, injury, and heat-treatment. The first cloned SRP was from *Hyphantria cunea*. By injecting bacteria to the hemocoel of larva, several inducible genes were found and cloned, one of which was named Hdd23, the homolog of *S. litura* SRP (Shin et al., 1998). Similar induction of SRP by bacteria infection was also found in *Helicoverpa armigera*. Knocking down of *H. armigera* SRP by injection of double-stranded RNA would decrease nodule formation and transcription of prophenoloxidase gene, and increase *E. coli* survival rate in hemolymph (Qiao et al., 2014). Despite functional similarity of SRPs with PPs, SRPs, with obvious distinct sequence features, widely exist in lepidopteran species and evolve independent of PPs.

2.5 SRPs of non-lepidopteran insects

SRPs were also identified in other insect species, including *D. melanogaster*, *Tribolium castaneum* and *Lucilia cuprina*, and were usually named GBP in these species. Their sequences are already very different from PPs and SRPs of lepidopteran species. GBPs of *D. melanogaster* and *L. cuprina*, like PP, can induce the expression of AMPs in a Toll/Imd-independent way, spreading of plasmatocytes of *D. melanogaster*, and also have mitogenic activity. Similar to *S. litura* SRP, GBP of *D. melanogaster* can also be induced by heat-treatment and bacterial infection (Tsuzuki et al., 2012).

3 Materials and Method

3.1 Insect rearing and injection, total RNA preparation, and cDNA synthesis

Eggs of *M. sexta* were purchased from Carolina Biological Supply, and larvae were reared on an artificial diet mainly consisted of wheat germ. Different tissue samples were collected from naïve day 2 fifth instar larvae, and stored in TRIZOL reagent (Thermo Fisher Scientific). Day 2 fifth instar larvae were injected with PBS, 4 μg synthetic SRP peptides in 40 μl PBS, or a mixture of pathogens, including *Escherichia coli* (2×10^7 cells), *Micrococcus luteus* (20 μg) (Sigma-Aldrich), and curdlan (20 μg, insoluble β-1,3-glucan from *Alcaligenes faecalis*) (Sigma-Aldrich) in 40 μl H_2O. Hemolymph and fat body samples were collected at 6 or 24 hours after challenge. Total RNA was extracted according the manual of the manufacture. The cDNA was synthesized from total RNA with iScript cDNA synthesis kit for qRT-PCR (bio-rad).

3.2 Quantitative real-time polymerase chain reaction (qRT-PCR)

The final 10 μl reaction system consists of 200 ng cDNA samples, 1× iTag Universal SYBR Green Supermix (Bio-Rad), and specific primers (0.5 μM each) in triplicate. The primers were: j037 (5' CATGATCCACTCCGGTGACC) and j038 (CGGGAGCATGATTTTGACCTTAA) for rpS3; j1070 (GCAGGCGACGACAAGAAC) and j1071 (ATGCGTGTTGGTAAGA GTAGC) for attacin; j1072 (CCGTGTTTTATTCTTCGTCTTC) and j1073 (AATCCTTTGACCTGCACCC) for cecropin-6; j1827 (GCTGTTGATCTGCGTGACAT) and j1828 (TC CTCCTTTGAATCCACGTC) for

defensin-2; j1074 (GCAAGTCGGCAACAATGG) and j1075 (ACCCTGTCCTGTCAGTTTG) for gloverin; j1076 (GTGTGCCTCGTGGAGAATG) and j1077 (ATGCCTTGGTGATGTCGTC) for lysozyme; j1078 (TGCTTTCTTTAACCTTTGTCCTC) and j1079 (TATTCTAACACAGCCTATAATGCG) for moricin-1; j1819 (TGCTCGTGCCTATACTCGTG) and j1834 (TACCTTGGCTACACGCACTG) for uENF1; j1821 (GGACGCGAAATTTGTGCTAT) and j1822 (TTTGTCTGCAGTTCCCCAAC) for uENF2; j1823 (GCGTGGTGTGGGAAAGTTAT) and j1824 (AACCCCCTGCAAAGTTTTCT) for PP; j1066 (GCCGAGGGTATCGTT) and j1067 (TCAGGCTTTGGCGTT) for SRP1; j1068 (GCCGAGGGCATCACC) and j1069 (CGGATGAGTTCTTCGTTTA) for SRP2; j1832 (TGGTGGATGTGAACCTCAAA) and j1833 (TACATAGCCTTTCGGGCATC) for SRP3. The thermal cycling conditions were: denaturation at 95 ℃ for 3 minutes followed by 40 cycles of 95 ℃ for 10 seconds and 60 ℃ for 30 seconds. After the PCR was complete on a CFX Connect Real-Time PCR Detection System (Bio-Rad), melt curves of the PCR products were examined to ensure proper shape and TM values. Amplification efficiencies (E), measured by amplifying a series dilution of cDNA sample diluted using the different primer pairs under the same conditions, were 93.8%% (rpS3), 105.3% (attacin), 90.0% (cecropin-6), 104.7% (defensin-2), 71.2% (gloverin), 88.1% (lebocin D), 121.0% (lysozyme-1), 88.9% (moricin-1), 91.2% (uENF1), 93.4% (uENF2), 100.4% (PP), 103.2% (SRP1), 105.3% (SRP2), 129.3% (SRP3). Relative mRNA levels were calculated as: $(1 + E_{rpS3})^{Ct,\ rpS3}/(1 + E_x)^{Ct,\ x}$ (Rieu and Powers, 2009).

3.3 SRP genes identification in M. sexta and arthropods

Proteins of all arthropods were downloaded from the European Bioinformatics Institute website (ftp://ftp.ebi.ac.uk/pub/databases/fastafiles/uniprot/uniprotkb_arthropoda.gz). A python script was written to find proteins less than 250 amino acids, ending with K/R-X_{1-15}-C-X_{3-10}-G-$X_{1/2}$-C-X_{1-15}, where X represents any amino acid residue other than cysteine, from the MCOT 1.0 gene models of *M. sexta* (Cao and Jiang, 2015) and arthropod protein sequences. This motif is modified from the SRP consensus sequence of C-X_2-G-$X_{4/6}$-G-$X_{1/2}$-C-K/R (Matsumoto et al., 2012). The reported SRP protein sequences were also used for tblastn search against Transcriptome Shortgun Assembly (TSA) of NCBI, setting the organism to arthropoda. The transcripts from tblastn search were translated to protein sequences. Protein sequences from both searches were combined and manually examined to identified putative SRPs, which are usually shorter than 250 residues, with a signal peptide and a simple pro-region with few or no cysteine.

3.4 Sequence alignment and phylogenetic analysis

Multiple sequence alignments of different groups of SRPs from *M. sexta* and other insects were performed using MUSCLE, a module of MEGA 7.0 (Tamura et al., 2013) at the following settings: refining alignment, gap opening penalty (–2.9), gap extension penalty (0), hydrophobicity multiplier (1.2), maximal iterations (100), UPGMB clustering (for iterations 1 and 2) and maximum diagonal length (24). The aligned sequences were used to construct neighbor-joining trees by MEGA 7.0 with bootstrap method for the phylogeny test (1000 replications, Poisson model, uniform rates, and complete deletion of gaps or missing data).

3.5 Expression and purification of pro-SRPs from E. coli

DNA fragments of SRPs were obtained from cDNA pool of 5th instar *M. sexta* fat body or hemocytes by PCR, with primers j1055 (GAATTCATATGGCGCCGACCTTAATTCAAGA) and j1057 (GTCGACCTATTACTGCCAAGGCTGCCTGCA) for uENF1, j1056 (GAATTCATATGGGTGTGGTTTTTAATTTTCA) and j1058 (GTCGACCTATTAAAACCTTTGTCTGCAGT) for uENF2, j1059 (GAATTCATATGAAGACCAAAGAGTTCCCGTTAC) and j1060 (GTCGACTTACTCGAGGTAGTCGTCATCRGG) for SRP1 and SRP2, and j1081 (GAATTCCGTGTGATCGACTCGAC) and j1082 (CTCGAGTTAATAGTCATAATCC) for SRP3. Primer j1060 is a mixture, where base R stands for A/G. The recovered PCR fragments were ligated with pGEM-T vector (Promega) and transformed to *E. coli* cell line JM109 (Promega). After verification by sequencing, the fragments were sub-cloned to pSKB3, an expression vector modified from pET-28a vector.

Proteins were expressed in *E. coli* BL21 gold (DE3) cells (Stratagene). Bacteria carrying the plasmid were grown in 800 ml LB medium at 37 ℃ until optical density at wavelength of 600 nm (OD600) reached 0.4. Bacteria were cultured for another 6 hours at 37 ℃ after adding IPTG (isopropyl-β-D-thiogalactopyranoside) to 1 mM final concentration. The cells were harvested by centrifugation at 4 ℃ and homogenized by sonication on ice. Cell lysates were centrifuge at 12,000 rpm for 30 minutes at 4 ℃. The pellets were dissolved in lysis buffer B (0.1 M NaH_2PO_4, 0.01 M Tris-HCl, 8 M urea, pH 8.0) at room temperature and centrifuged at 30,000 rpm for one hour at 25 ℃ to remove the insoluble pellets. The supernatants were loaded to a Ni-NTA column (Qiagen) and binding recombinant protein

was eluted with buffer B with pH 6.3, 5.9 and 4.5. The purified denatured protein was recovered by dialysis against buffer for renaturing the protein (20 mM Tris, pH 7.5, 100 mM NaCl, 2 mM reduced glutathione, 0.2 mM oxidized glutathione, 5% glycerol) with 4 M, 2 M, 1 M and 0 M urea for 12 hours each. After centrifugation to remove pellets, protein was concentrated and buffer was changed with spin columns (cutoff 3,500 daltons).

3.6 ProSPRs cleavage by hemolymph, hemolymph fractions and PAPs

Bar stage hemolymph fractions were collected from a hydroxyapatite (HT) column as described previously (Wang et al., 2014). In each tube, 2 µl fraction, 500 ng recombinant pro-SRP and tris buffer (20 mM Tris, 100 mM NaCl, pH8.0) were mixed together (totally 12 µl) and reacted at room temperature for 30 minutes. 20 µl of 100% saturated ammonium sulfate was added to 20 µl induced or pupa day 1 hemolymph to 50% saturation. Pellets were collected after centrifugation to remove storage protein and re-dissolved in 40 µl Tris buffer. 10 µg *Micrococcus luteus* was added to activate the hemolymph. In each tube, 10 µl reaction system contained 2 µl activated hemolymph resuspension, 500 ng pro-SRP and Tris buffer, and reacted at room temperature for 30 minutes. PAP3 was kindly provided by Yang Wang and was purified from *M. sexta* hemolymph. In reaction with PAP3, the 10 ul reaction system included about 50 ng PAP3 and 500 ng proSRP, and reacted at room temperature for 30 minutes. ProPAP3, kindly provided by Yingxia Hu, was expressed and purified in insect Sf9 cell lines. In reaction with proPAP3, 10 ul reaction system included about 5 ng PAP3, 200 ng proPAP3 and 100 ng proSPR3. Tricine loading buffer was added and samples were boiled 100 ℃ for 5min. Western blot was done with mouse anit-His ×6 as primary antibody and alkaline-phosphatase linked goat anti-mouse IgG as secondary antibody.

3.7 Cleavage site identification by MALDI-MS

The reaction system contained about 5 ng PAP3, 200 ng proPAP3 and 5 μg recombinant proSRP2 and reacted at room temperature for 1 hour. Samples with PAP3 and proPAP3 only or proSPR2 only were used as negative control. Matrix-assisted laser desorption-ionization time-of-flight mass spectrometry (MALDI-TOF MS) was performed in the core facility of Oklahoma State University to detect newly generated peptide.

4. Results

4.1 Overview of SRPs in insects

We use SRP to represent this group of cytokines including PP, uENF1, uENF2 and SRPs. 19 PP, 5 uENF1 and 5 uENF2 genes were found in NCBI GenBank database directly by simple blast search (Fig. 1A, B, C). The uENF1 and uENF2 sequences of *Plutella xylostella* were manually identified by translating nucleotide sequences to proteins. As one of the first identified and most well-studied insect cytokines, PPs are very conserved in different lepidopteran species, usually a 23 amino acid peptides with the common sequence of ENF-A/S/R-GGC-A/L/V-A/T-GY/F-M/Q-RT-A/S-DGRCKPTF. Several residues in the sequences, including the first three residues, ENF, and the last 8 residues, DGRCKPTF are very conserved across all species. Deletion of the N-terminal E would eliminated plasmatocyte spreading activity without influence mitogenic activity, while deletion of C-terminal F would eliminate mitogenic activity without influence plasmatocyte spreading activity in *P. separata* (Aizawa et al., 2001). Alanine-scanning mutagenesis of *P. includens* found that E1A mutation increased plasmatocyte activity while F3A had no activity,

indicating critical role of F3 (Clark et al., 2001). However, PP of *P. xylostella* has M instead of F at third position. Encoded by the same tricistronic mRNA, uENF1 have less conserved residues than PP and uENF2. uENF1, uENF2 and PP of *P. xylostella* are more different from other species, which can be partially explained by the fact that it is a more ancient species compared to others (Mutanen et al., 2010).

As mentioned previously, the first cytokine named SRP was identified in *S. litura* (Slit1 in Fig. 1D). By tblastn search against TSA, we identified 11 peptides highly similar to SRP of S. litura, including SRP1 of *M. sexta* (Fig. 1D), and 31 peptides less similar, including M. sexta SRP2 and another two SRP from *S. litura* (Slit2 and Slit3 in Fig. 1E). The first group has sequence like H-G/N-IRVG-T/A-CP-L/S/A-GY-T/V/S-R/K-RGGFCFPDDDY, while the second group has less conserved residues. Surprisingly, the species *Lygus hesperus* of Hemiptera order also has a SRP very similar to Slit2, suggesting the possible conserved role of SRPs in insects other than Lepidoptera.

Tree of Drosophila GBP (Fig. 1F) shows that even though these sequences are from species belong to the same genus, the conservation of them are lower compared to PPs. These sequences are very different from SRPs in lepidopteran species. The low conservation of SRPs in insects makes it hard for evolutionary analysis.

4.2 SRPs in M. sexta

Totally 11 cytokines were found in M. sexta. SRP1 to 5, 7 and 8 were grouped together in the tree (Fig. 2), with a D-rich C-terminus usually ended with Y. The predicted cutting site for SRP1, 2 and 3 is between R and F in the sequences, and between R and N for SRP6 (only F/N shown in Fig. 2). Comparing with other SRPs, SRP1 is more similar with SRP

of S. litura, and thus may play similar roles. SRP2 and SRP3 have very similar sequence with SRP1 and may also have similar functions with SRP1, while SRP6 is already very different than other sequences, and may have its unique roles.

4.3 Expression of SRPs of M. sexta by RNA-seq

52 cDNA libraries were sequenced together with the genome of *M. sexta* (Kanost et al., 2016), and are very excellent resources for preliminary checking of gene expression in different tissues of different developmental stages. As shown in Table 1 and Fig. 3, SRP5, 7 and 8 are almost not expressed in all libraries. Translated from the same mRNA, uENF1 and uENF2 have almost identical FPKM values in Fig. 3. The genes coding uENF1, uENF2 and PP have two forms, the longer one coding all three and shorter one coding PP only. The ratio of two different transcripts can be estimated from FPKM values of uENFs and PPs. In most libraries, the shorter one dominated at high value, and surprisingly, starting from pre-wandering stage, the longer form became the major one and expressed at high level in midgut. SRP2, 3 and 4 are generally higher expressed than SRP1 in most libraries, though SRP1 is more similar to the first studied SRP from S. litura. Surprisingly, SRP6 became extremely highly expressed in pre-wandering and wandering gut, indicating its potential role in influence the behavior of midgut at these stages.

4.4 PAP1 and PAP3 are enzymes activating proSRPs in hemolymph

Similar to PP, SRPs were secreted in the form of pro-protein and activated by limited cleavage at the activation site. Since PP can be cleaved in hemolymph (Yang Wang, 1999), we first check the cleavage of recombinant proSRPs in induced hemolymph. As shown in Fig. 4A, proSRP1, 2 and 3 can be cleaved, while pro-uENF1 and 2 cannot.

In order to identify the enzyme which activates proSRP1 and proSPR2, we used the bar stage hemolymph fractions to react with recombinant proteins. As shown in Fig. 4B, the first few tubes have higher activity in cleavage of both proSRP1 and 2. PAP1 and PAP3 were eluted in first in HT column and their cutting site is between R and F (Wang et al., 2014), which are the predicted activation site of proSRP1, 2 and 3. Thus we checked whether PAP1 and PAP3 did cleave proSRP1, 2 and 3. With induced (IH) and pupa hemolymph (PH) as positive control, PAP1 and PAP3 can cut proSRP1 and 2 (Fig. 4C), while it is not clear whether proSRP3 can be cleaved or not (data not shown). Different from proSRP1 and 2, the pro-region of proSRP3 might be further cleaved by enzyme, which explains why his antibody cannot detect it well. Due to the limited amount of PAP1 and PAP3 purified from hemolymph, we then used recombinant proPAP3, which can be self-activated by active PAP3 (Wang et al., 2014), to cut proSRP3. ProSRP3 can be totally cleaved by proPAP3. However, proSRP3 is not well cleaved in IH and PH, indicating possible other activating enzymes.

4.5 Activation site of proSRP2

The active *S. litura* SRP peptides were activated by cleavage between R and H, and the predicted activation site for *M. sexta* SRP1, 2 and 3 is between R and F. Without purifying active SRPs from the hemolymph, we decided to check the activation site with recombinant protein *in vitro*. The result of MALDI-MS supported our hypothesis (Fig. 5). The signal of full-length recombinant proSRP2 can be detected in proSRP2 only sample. After mixing with PAP3 and proPAP3, there is a very strong peak from 2992 to 3000 Da, the position of activated recombinant peptide FGVKDGKCPSGRVRRLGICVPDDDYLE (2994.5 Da). This result showed PAP3 specifically cleaved between R and F.

4.6 SRP6 blocks growth of M. sexta larvae

S. litura SRP peptides could block the growth of larvae after injection (Yamaguchi et al., 2012). We first checked similar function of SRP1 and SRP2, the mostly similar peptides to *S. litura* SRP (Fig. 6A). There is no significant difference of larvae injected with SRP1 or 2 and PBS. Noticing that SRP6 is extremely expressed in pre-wandering and wandering midgut, we also tested function of SRP6 (Fig. 6B). We observed that, after injection of SRP6, in the first 6 hours, the larvae stopped feeding, and had more feces compared to control, which was why the body weight decreased after 6 hours. After 24 hours, the body weight was significantly lower than control group. Blast search showed that there was a group of proteins very similar to SRP6 in lepidopteran species (data not shown). This result indicates that SRP6 is an important regulator of larva feeding behavior, and plays important role in pre-wandering and wandering stage.

4.7 qRT-PCR analysis of expression of SRPs in different tissues and after different treatments

PP and *S. litura* SRP were both identified in parasitized moth larvae, and are possible regulator of immune response. We did qRT-PCR checking expression of SRPs under immune challenge and in different tissues to investigate their roles in immunity and to verify the expression by RNA-seq data (Fig. 7 and 8). Similar expression pattern was observed for uENF1 and uENF2 in different cDNA libraries, as they are from the same mRNA. PP is much higher expressed than uENF1 and uENF2, and mostly highly expressed in fat body. mRNA level of PP decreased 6 hours after immune challenge, and recovered after 24 hours. Generally, SRP2 was more highly expressed than SRP1 and SRP3, and

mostly highly expressed in hemocytes, but was not induced after immune challenge. One the other hand, both SRP1 and SRP3 can be significantly induced after 6 hours, and SRP1 even reached a higher level than SRP2. However, the expression of SRP3 is much less compared to SRP1 and SRP2. Heat treatment 42 ℃ for 1 hour did not influence expression of these cytokines. Overall, SRP1 was mostly likely the ortholog of *S. litura* SRP, though we did not observe the growth blocking function of it (Fig. 6A).

4.8 Injection of SRPs induces expression of AMPs

A dramatic phenomenon of immune response of insects is the induction of many AMP genes. Injection of PP induced expression several AMPs and other immune-related genes in *B. mori* (Ishii et al., 2013; Ishii et al., 2010). We used qRT-PCR to check expression of AMP gene expression in fat body after injection synthetic peptides for 6 hours (Fig. 9). PP significantly induced the expression of cecropin 6, gloverin, attacin, lysozyme and moricin. uENF1 and uENF2 had a little stronger effect than PP. SRP1 and SRP2 also induced expression of several AMPs, while SRP3 did not induce expression of AMPs. Overall, PP, uENF1, uENF2, SRP1 and SRP2 could induce the expression of AMPs, but efficiency is much lower compared to pathogens (data not shown).

5 Discussion

The mRNA level of PP is the highest and of the 11 cytokines it is the only one identified in the proteome peptidome study of hemolymph of *M. sexta* (He et al., 2016; Zhang et al., 2014), indicating higher protein level in hemolymph of 5^{th} instar larva. This may be why PP is the first identified and well-studied immune-related cytokine. ProPP is stored in

hemolymph and can be activated upon immune challenge. However, the upstream enzyme of proPP remains undiscovered.

Based on the RNA-seq data, the longer transcript coding uENF1, uENF2 and PP is lowly transcribed in fat body, the major source of hemolymph proteins, and is only about 10% of the shorter form coding PP only. It becomes highly transcribed and seems the only form in midgut after pre-wandering stage. While the ratio of uENF1, uENF2 and PP proteins translated from this transcript remains unknown, it may play a more import role in midgut.

SRP1 is most similar to *S. litura* SRP, with similar sequence and both being highly induced after immune challenge. *S. litura* SRP was reported with similar function like PP. However, SRP1 behaves like AMPs, whose expression levels increase 6 hours after challenge and decrease after 24 hours. It takes time to accumulate enough pro-SRP1 in hemolymph, and at that time, PAP1 or PAP3, activator of proSRP1, may be already deactivated. Surprisingly, the expression of PP drops 6 hours after challenge. May be proSRP1 is stored for next immune response.

SRP2 is very similar to SRP1, both with PDDDY in the c-terminus and both can be activated by PAP1 and PAP3. SRP2 is much higher expressed in hemocytes than in fat body. Beginning from wandering stage, SRP2 expression increase and peaks at pupa 1 to 3 days in fat body, a sign of immune-related genes (He et al., 2015). It is unknown how SRP1 and SRP2 are differently regulated to regulate immune response.

There is a family of SRP6 like cytokines in lepidopteran species. *M. sexta* SRP6 become very highly expressed in pre-wandering and wandering midgut, and likely the important regulator of behavior of midgut in wandering stage. Proteins expressed in midgut are

usually not secreted to hemocoel, and how injection of SRP6 works remains unknown. It will be interesting to check activation and working mechanisms of SRP6, which may help control pest in the future.

Acknowledgement

This study was supported by NIH grant GM58634. This work was supported in part under project OKLO2450 (to H. Jiang). Computation for this project was performed at OSU High Performance Computing Center supported in part through NSF grant OCI-1126330.

Tables

Table 1. Expression of SRPs in fat body and midgut libraries

	fat body								midgut											
FPKM	L4-Late	L5-D1	L5-preW	L5-W	P-D1~3	P-	A-D1~3	A-D7~9	L2	L3-Late	L4-12h	L4-Late	L5-1~3h	L5-D1	L5-preW	L5-W	L5-W	P-D1	P-	A-D3~5
SRP1	0	0	0	2	6	2	2	7	14	2	1	0	2	2	1	61	17	6	0	0
SRP2	14	29	8	191	784	27	69	198	11	12	1	2	6	2	6	313	205	909	35	5
SRP3	5	1	1	2	12	63	10	27	1	12	0	8	9	1	1	8	1	4	1	0
SRP4	3	4	2	21	199	17	4	16	15	5	4	5	7	2	3	108	16	47	2	2
SRP5	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0
SRP6	1	1	0	0	1	158	31	0	110	194	144	142	187	324	**1947**	**8548**	**1200**	126	426	152
SRP7	0	0	0	0	3	4	0	1	0	0	0	0	0	1	0	0	1	0	0	0
SRP8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
uENF1	62	94	5	104	65	35	59	10	17	44	20	27	62	60	30	316	303	567	144	245
uENF2	74	104	7	103	192	39	55	11	23	88	26	41	88	91	62	441	261	520	166	199
PP	762	865	274	776	740	190	493	261	34	**1418**	333	141	214	155	76	699	174	445	229	211

Figures

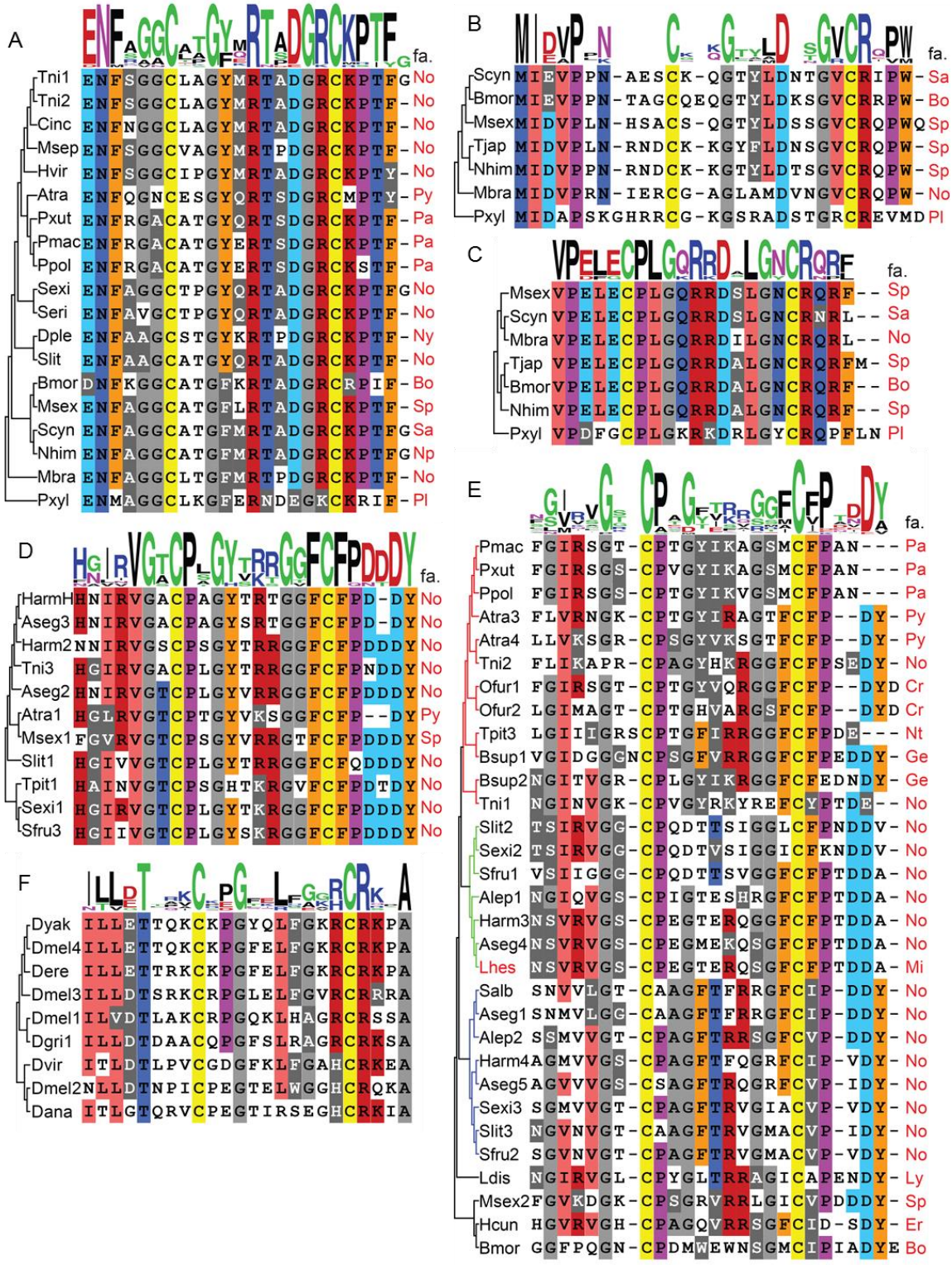

A
fa.
Tni1 ENFSGGCLAGYMRTADGRCKPTFG No
Tni2 ENFSGGCLAGYMRTADGRCKPTF- No
Cinc ENFNGGCLAGYMRTADGRCKPTF- No
Msep ENFSGGCVAGYMRTPDGRCKPTF- No
Hvir ENFSGGCIPGYMRTADGRCKPTY- No
Atra ENFQGNCESGYQRTSDGRCMPTY- Py
Pxut ENFRGACATGYQRTSDGRCKPTF- Pa
Pmac ENFRGACATGYERTSDGRCKPTF- Pa
Ppol ENFRGACATGYERTSDGRCKSTF- Pa
Sexi ENFAGGCTPGYQRTADGRCKPTFG No
Seri ENFAVGCTPGYQRTADGRCKPTF- No
Dple ENFAAGCSTGYKRTPDGRCKPTF- Ny
Slit ENFAAGCATGYQRTADGRCKPTF- No
Bmor DNFKGGCATGFKRTADGRCRPIF- Bo
Msex ENFAGGCATGFLRTADGRCKPTF- Sp
Scyn ENFAGGCATGFMRTADGRCKPTFG Sa
Nhim ENFAGGCATGFMRTADGRCKPTFG Np
Mbra ENFAGGCLTGFMRTPDGRCKPTF- No
Pxyl ENMAGGCLKGFERNDEGKCKRIF- Pl
B
Scyn MIEVPPN-AESCK-QGTYLDNTGVCRIPW- Sa
Bmor MIEVPPN-TAGCQEQGTYLDKSGVCRRPW- Bo
Msex MIDVPLN-HSACS-QGTYLDSSGVCRQPWQ Sp
Tjap MIDVPLN-RNDCK-KGYFLDNSGVCRQPW- Sp
Nhim MIDVPPN-RNDCK-KGTYLDTSGVCRQPW- Sp
Mbra MIDVPRN-IERCG-AGLAMDVNGVCRQPW- No
Pxyl MIDAPSKGHRRCG-KGSRADSTGRCREVMD Pl
C
fa.
Msex VPELECPLGQRRDSLGNCRQRF-- Sp
Scyn VPELECPLGQRRDSLGNCRNRL-- Sa
Mbra VPELECPLGQRRDILGNCRQRL-- No
Tjap VPELECPLGQRRDALGNCRQRFM- Sp
Bmor VPELECPLGQRRDALGNCRQRF-- Bo
Nhim VPELECPLGQRRDALGNCRQRF-- Sp
Pxyl VPDFGCPLGKRKDRLGYCRQPFLN Pl
D
fa.
HarmH HNIRVGACPAGYTRTGGFCFPD-DY No
Aseg3 HNIRVGACPAGYSRTGGFCFPD-DY No
Harm2 NNIRVGSCPSGYTRRGGFCFPDDDY No
Tni3 HGIRVGACPLGYTRRGGFCFPNDDY No
Aseg2 HNIRVGTCPLGYVRRGGFCFPDDDY No
Atra1 HGLRVGTCPTGYVKSGGFCFP--DY Py
Msex1 FGVRVGTCPSGYVRRGTFCFPDDDY Sp
Slit1 HGIVVGTCPLGYTRRGGFCFQDDDY No
Tpit1 HAINVGTCPSGHTKRGVFCFPDTDY No
Sexi1 HGIRVGTCPLGYTKRGGFCFPDDDY No
Sfru3 HGIIVGTCPLGYSKRGGFCFPDDDY No
E
fa.
Pmac FGIRSGT-CPTGYIKAGSMCFPAN--- Pa
Pxut FGIRSGS-CPVGYIKAGSMCFPAN--- Pa
Ppol FGIRSGS-CPTGYIKVGSMCFPAN--- Pa
Atra3 FLVRNGK-CPTGYIRAGTFCFP--DY- Py
Atra4 LLVKSGR-CPSGYVKSGTFCFP--DY- Py
Tni2 FLIKAPR-CPAGYHKRGGFCFPSEDY- No
Ofur1 FGIRSGT-CPTGYVQRGGFCFP--DYD Cr
Ofur2 LGIMAGT-CPTGHVARGSFCFP--DYD Cr
Tpit3 LGIIIGRSCPTGFIRRGGFCFPDE--- Nt
Bsup1 VGIDGGGNCPSGFVRRGGFCFPEDDY- Ge
Bsup2 NGITVGR-CPLGYIKRGGFCFEDNDY- Ge
Tni1 NGINVGK-CPVGYRKYREFCYPTDE-- No
Slit2 TSIRVGG-CPQDTTSIGGLCFPNDDV- No
Sexi2 TSIRVGG-CPQDTVSIGGICFKNDDV- No
Sfru1 VSIIGGG-CPQDTTSVGGFCFPTDDA- No
Alep1 NGIQVGS-CPIGTESHRGFCFPTDDA- No
Harm3 NSVRVGS-CPEGTERQGGFCFPTDDA- No
Aseg4 NSVRVGS-CPEGMEKQSGFCFPTDDA- No
Lhes NSVRVGS-CPEGTERQSGFCFPTDDA- Mi
Salb SNVVLGT-CAAGFTFRRGFCIP-DDY- No
Aseg1 SNMVLGG-CAAGFTFRRGFCIP-DDY- No
Alep2 SSMVVGT-CPAGFTRRSGFCVP-DDY- No
Harm4 AGMVVGS-CPAGFTFQGRFCIP-VDY- No
Aseg5 AGVVVGS-CSAGFTRQGRFCVP-IDY- No
Sexi3 SGMVVGT-CPAGFTRVGIACVP-VDY- No
Slit3 NGVNVGT-CAAGFTRVGMACVP-IDY- No
Sfru2 SGVNVGT-CPAGFTRVGMACVP-VDY- No
Ldis NGIRVGL-CPYGLTRRAGICAPENDY- Ly
Msex2 FGVKDGK-CPSGRVRRLGICVPDDDY- Sp
Hcun HGVRVGH-CPAGQVRRSGFCID-SDY- Er
Bmor GGFPQGN-CPDMWEWNSGMCIPIADYE Bo
F
Dyak ILLETTQKCKPGYQLFGKRCRKPA
Dmel4 ILLETTQKCKPGFELFGKRCRKPA
Dere ILLETTRKCKPGFELFGKRCRKPA
Dmel3 ILLDTSRKCRPGLELFGVRCRRRA
Dmel1 ILVDTLAKCRPGQKLHAGRCRSSA
Dgri1 ILLDTDAACQPGFSLRAGRCRKSA
Dvir ITLDTLPVCGDGFKLFGAHCRKEA
Dmel2 NLLDTNPICPEGTELWGGHCRQKA
Dana ITLGTQRVCPEGTIRSEGHCRKIA

Fig. 1. **Multiple sequence alignment and sequence logo of SRPs**. **A**, PP. **B**, uENF1. **C**, uENF2. **D**, SRPs most similar to *S. litura* SRP. **E**, SRPs less similar to S. litura SRP. **F**, GBP of Drosophila. Species: Aech, *Acromyrmex echinatior*; Aseg, *Agrotis segetum*; Atra, *Amyelois transitella*; Alep, *Athetis lepigone*; Bsup, *Biston suppressaria*; Bmor, *Bombyx mori*; Cinc, *Chrysodeixis includens*; Dple, *Danaus plexippus*; Dana, *Drosophila ananassae*; Dere, *Drosophila erecta*; Dgri, *Drosophila grimshawi*; Dmel, *Drosophila melanogaster*; Dvir, *Drosophila virilis*; Dyak, *Drosophila yakuba*; Harm, *Helicoverpa armigera*; Hass, *Helicoverpa assulta*; Hvir, *Heliothis virescens*; Hcun, *Hyphantria cunea*; Lhes, *Lygus hesperus*; Ldis, *Lymantria dispar*; Mbra, *Mamestra brassicae*; Msex, *Manduca sexta*; Msep, *Mythimna separata*; Nhim, *Neogurelca himachala sangaica*; Ofur, *Ostrinia furnacalis*; Pmac, *Papilio maChaon*; Ppol, *Papilio polytes*; Pxut, *Papilio xuthus*; Pxyl, *Plutella xylostella*; Scyn, *Samia cynthia pryeri*; Sric, *Samia ricini*; Seri, *Spodoptera eridania*; Sexi, *Spodoptera exigua*; Sfru, *Spodoptera frugiperda*; Slit, *Spodoptera litura*; Salb, *Striacosta albicosta*; Tpit, *Thaumetopoea pityocampa*; Tjap, *Theretra japonica*; Tni, *Trichoplusia ni*. Families (fa.): Bo, Bombycidae; Er, Erebidae; Ge, Geometridae; No, Noctuidae; Nt, Notodontidae; Ny, Nymphalidae; Pa, Papilionidae; Pl, Plutellidae; Py, Pyralidae; Sa, Saturniidae; Sp, Sphingidae. All families before are from the order of Lepidoptera. Mi, Miridae, order of Hemiptera.

Fig. 2. **Sequence alignment of SRPs of *M. sexta***.

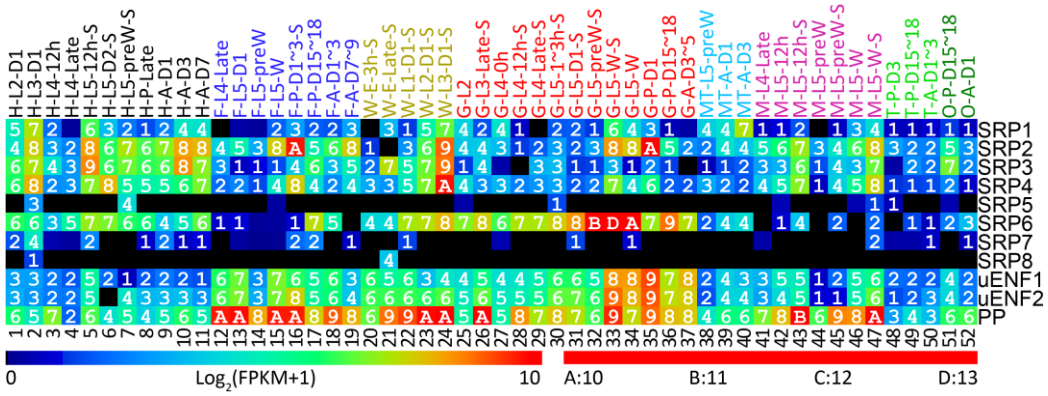


Fig. 3. **Expression of SRPs in different cDNA libraries**. The cDNA libraries represent the following tissues and stages: head [**1**. 2nd (instar) L (larvae), d1 (day 1); **2**. 3rd L, d1; **3**. 4th L, 12h (hour); **4**. 4th L, late; **5**. 5th L, d0.5; **6**. 5th L, d2; **7**. 5th L, pre-W (pre-wandering); **8**. P (pupae), late; **9**. A (adults), d1; **10**. A, d3; **11**. A, d7], fat body (**12**. 4th L, late; **13**. 5th L, d1; **14**. 5th L, pre-W; **15**. 5th L, W; **16**. P, d1-3; **17**. P, d15-18; **18**. A, d1-3; **19**. A, d7-9), whole animals [**20**. E (embryos), 3h; **21**. E, late; **22**. 1st L; **23**. 2nd L; **24**. 3rd L), midgut (**25**. 2nd L; **26**. 3rd L; **27**. 4th L, 0h; **28**. 4th L, 12h; **29**. 4th L, late; **30**. 5th L, 1-3h; **31**. 5th L, 24h; **32**. 5th L, pre-W; **33-34**. 5th L, W; **35**. P, d1; **36**. P, d15-18; **37**. A, d3-5;), MT (**38**. 5th L, pre-W; **39**. A, d1; **40**. A, d3), muscle (**41**. 4th L, late; **42-43**. 5th L, 12h; **44-45**. 5th L, pre-W; **46-47**. 5th L, W), testes (**48**. P, d3; **49**. P, d15-18; **50**. A, d1-3), and ovaries (**51**. P, d15-18; **52**. A, d1).

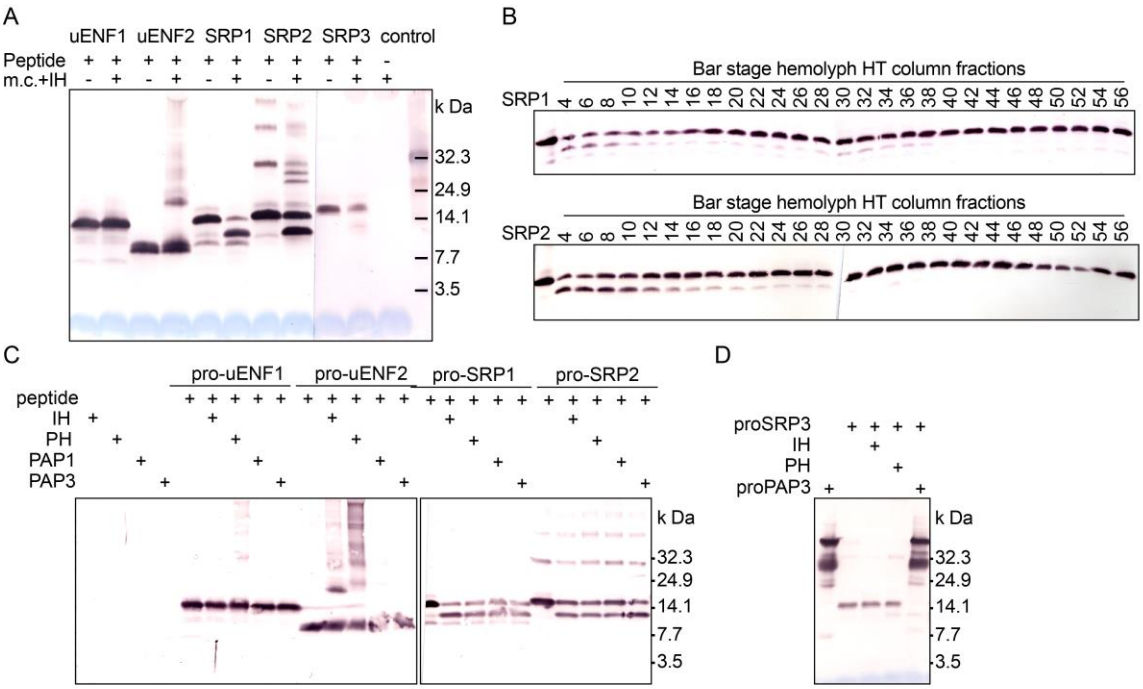


Fig. 4. **In vitro activation of SRPs**. **A**, cleavage of pro-SRPs in induced hemolymph (IH). **B**, activation of pro-SRP1/2 in different hydroxyapatite (HT) column fractions of bar stage hemolymph. **C**, cleavage of pro-uENF1, pro-uENF2, pro-SRP1 and pro-SRP2 by IH and PH (day 1 pupa hemolymph). **D**, cleavage of pro-SRP3 by activated PAP3. PAP3 purified from hemolymph was added to the system to induce the auto-cleavage of proPAP3.

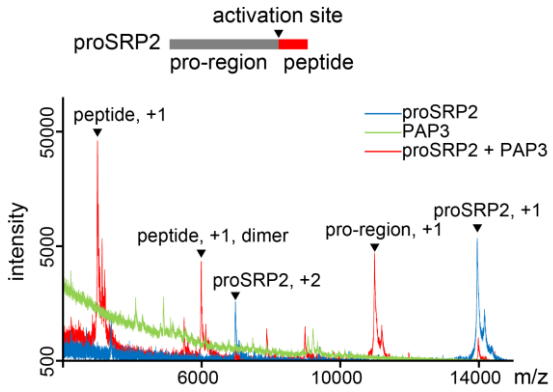


Fig. 5. **Identification of proSRP2 cleavage site by MALDI-MS**. PAP3 and proSRP2 were mixed together for one hour at room temperature. MALDI-MS were performed for three samples. m/z, molecular weight/charge. +1, +2, charge of fragments.

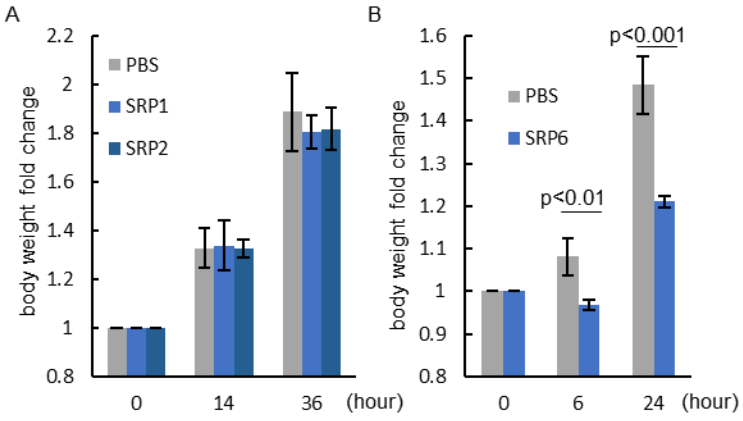


Fig. 6. **Effects of SRPs on growth rate**. **A**, body weight fold change of 5th instar larvae injected with SRP1 and SRP2. PBS as control. **B**, body weight fold change of 5th instar larvae injected with SRP6. 5th instar day 1 larvae were injected with 40 µl PBS or PBS with 4 µg peptides. Body weight was measured different time after injection. Time 0 body weight was set to 1. Body weight fold change was calculated for each larva. Each value represents the mean ± S.D. for 4 to 6 different larvae.

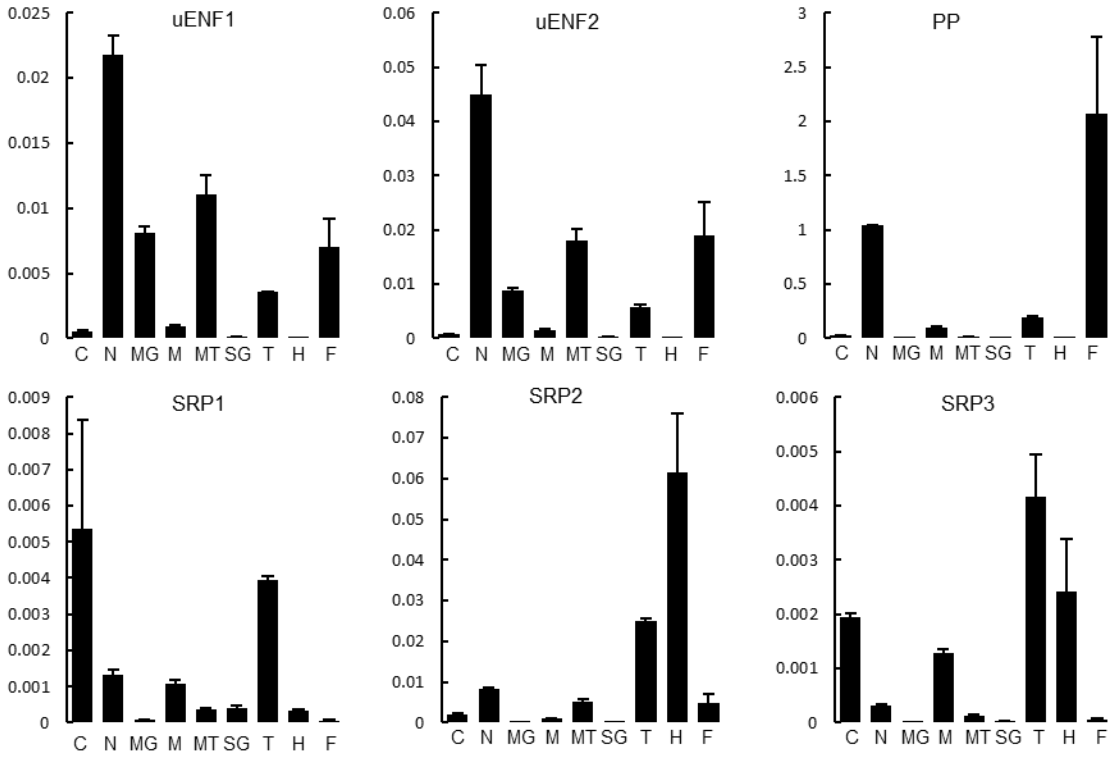


Fig. 7. **Expression of SRPs in different tissues**. Relative amount of SRPs to rpS3 by qRT-PCR. cDNA libraries: C, cuticle; N, nerve; MG, midgut; M, muscle, MT, malpighian tubule; SG, salivary gland; T, trachea; H, hemocytes; F, fat body.

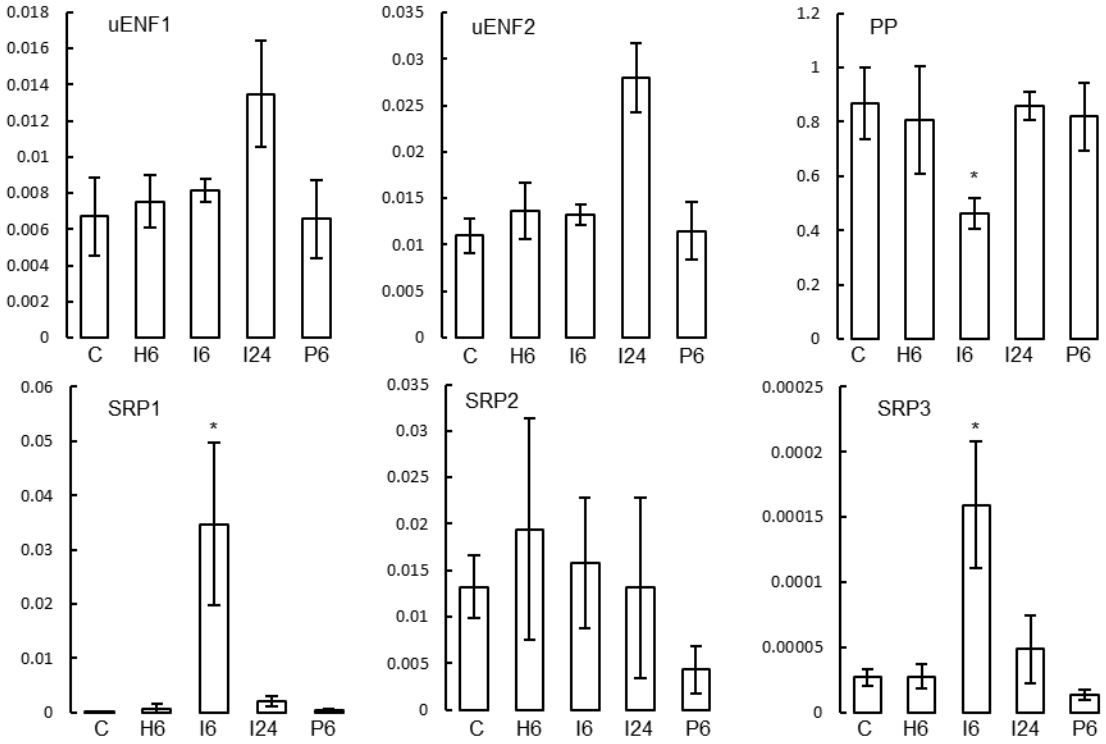


Fig. 8. **Expression of SRPs with heat treatment or bacteria challenge**. Relative amount of SRPs to rpS3 by qRT-PCR. cDNA libraries (all from fat body): C, control, healthy 5th instar day 2 larvae; H6, 6 hours after one hour of 42℃ heat treatment; I6, 6 hours after injecting mixture of pathogens; I24, 24 hours after injecting mixture; P6, 6 hours after injecting PBS. *, $p < 0.05$ compared with library P6.

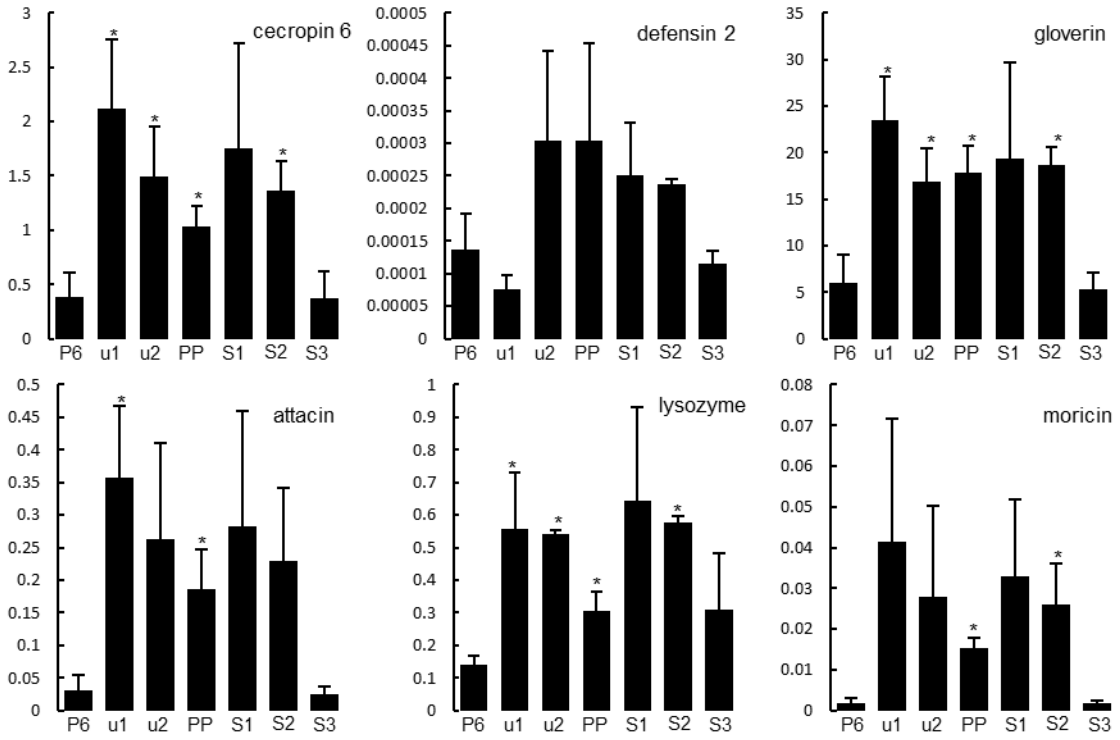


Fig. 9. **Expression of AMPs after injection of SRPs**. Relative amount of AMPs to rpS3 by qRT-PCR. Each library with 3 biological replicates. 5th instar day 2 larvae were injected with 40 ul 100ng/ul synthetic peptides in PBS, and after 6 hours, fat bodies were collected. Libraries: P6, PBS; u1, uENF1; u2, uENF2; PP, paralytic peptide; S1, SRP1; S2, SRP2; S3, SRP3. *, p <0.05 compared with library P6.

References

Aizawa, T., Hayakawa, Y., Ohnishi, A., Fujitani, N., Clark, K.D., Strand, M.R., Miura, K., Koganesawa, N., Kumaki, Y., Demura, M., Nitta, K., Kawano, K., 2001. Structure and activity of the insect cytokine growth-blocking peptide. Essential regions for mitogenic and hemocyte-stimulating activities are separate. The Journal of biological chemistry 276, 31813-31818.

Arango Duque, G., Descoteaux, A., 2014. Macrophage cytokines: involvement in immunity and infectious diseases. Front Immunol 5, 491.

Cao, X., Jiang, H., 2015. Integrated modeling of protein-coding genes in the Manduca sexta genome using RNA-Seq data from the biochemical model insect. Insect biochemistry and molecular biology 62, 2-10.

Clark, K.D., Pech, L.L., Strand, M.R., 1997. Isolation and identification of a plasmatocyte-spreading peptide from the hemolymph of the lepidopteran insect Pseudoplusia includens. The Journal of biological chemistry 272, 23440-23447.

Clark, K.D., Volkman, B.F., Thoetkiattikul, H., King, D., Hayakawa, Y., Strand, M.R., 2001. Alanine-scanning mutagenesis of plasmatocyte spreading peptide identifies critical residues for biological activity. The Journal of biological chemistry 276, 18491-18496.

Clark, R.I., Woodcock, K.J., Geissmann, F., Trouillet, C., Dionne, M.S., 2011. Multiple TGF-beta superfamily signals modulate the adult Drosophila immune response. Curr Biol 21, 1672-1677.

Diamond, G., Beckloff, N., Weinberg, A., Kisich, K., 2009. The roles of antimicrobial peptides in innate host defense. Current pharmaceutical design 15, 2377-2392.

Eleftherianos, I., Xu, M., Yadi, H., Ffrench-Constant, R., Reynolds, S., 2009. Plasmatocyte-spreading peptide (PSP) plays a central role in insect cellular immune defenses against bacterial infection. The Journal of experimental biology 212, 1840-1848.

Ha, S.D., Nagata, S., Suzuki, A., Kataoka, H., 1999. Isolation and structure determination of a paralytic peptide from the hemolymph of the silkworm, Bombyx mori. Peptides 20, 561-568.

Hayakawa, Y., 1990. Juvenile hormone esterase activity repressive factor in the plasma of parasitized insect larvae. The Journal of biological chemistry 265, 10813-10816.

Hayakawa, Y., Ohnishi, A., 1998. Cell growth activity of growth-blocking peptide. Biochemical and biophysical research communications 250, 194-199.

Hayakawa, Y., Ohnishi, A., Yamanaka, A., Izumi, S., Tomino, S., 1995. Molecular cloning and characterization of cDNA for insect biogenic peptide, growth-blocking peptide. FEBS Lett 376, 185-189.

Hayakawa, Y., Yasuhara, Y., 1993. Growth-Blocking Peptide or Polydnavirus Effects on the Last Instar Larvae of Some Insect Species. Insect biochemistry and molecular biology 23, 225-231.

He, Y., Cao, X., Li, K., Hu, Y., Chen, Y.R., Blissard, G., Kanost, M.R., Jiang, H., 2015. A genome-wide analysis of antimicrobial effector genes and their transcription patterns in Manduca sexta. Insect biochemistry and molecular biology 62, 23-37.

He, Y., Cao, X., Zhang, S., Rogers, J., Hartson, S., Jiang, H., 2016. Changes in the Plasma Proteome of Manduca sexta Larvae in Relation to the Transcriptome Variations after an Immune Challenge: Evidence for High Molecular Weight Immune Complex Formation. Molecular & cellular proteomics : MCP 15, 1176-1187.

Imler, J.-L., 2013. Overview of Drosophila immunity: a historical perspective. Developmental and comparative immunology.

Isaac, G.-S., Alex, C.-A., 2012. Phenoloxidase: a key component of the insect immune system. Entomologia Experimentalis et Applicata 142.

Ishii, K., Adachi, T., Hamamoto, H., Oonishi, T., Kamimura, M., Imamura, K., Sekimizu, K., 2013. Insect cytokine paralytic peptide activates innate immunity via nitric oxide production in the silkworm Bombyx mori. Developmental and comparative immunology 39, 147-153.

Ishii, K., Hamamoto, H., Kamimura, M., Nakamura, Y., Noda, H., Imamura, K., Mita, K., Sekimizu, K., 2010. Insect cytokine paralytic peptide (PP) induces cellular and humoral immune responses in the silkworm Bombyx mori. The Journal of biological chemistry 285, 28635-28642.

Ishii, K., Hamamoto, H., Sekimizu, K., 2015. Paralytic Peptide: An Insect Cytokine That Mediates Innate Immunity. Arch Insect Biochem 88, 18-30.

Izadpanah, A., Gallo, R.L., 2005. Antimicrobial peptides. J Am Acad Dermatol 52, 381-390; quiz 391-382.

Jiang, H., Vilcinskas, A., Kanost, M., 2010. Immunity in lepidopteran insects. Advances in experimental medicine and biology 708, 181-204.

Kanamori, Y., Hayakawa, Y., Matsumoto, H., Yasukochi, Y., Shimura, S., Nakahara, Y., Kiuchi, M., Kamimura, M., 2010. A eukaryotic (insect) tricistronic mRNA encodes three proteins selected by context-dependent scanning. The Journal of biological chemistry 285, 36933-36944.

Kanost, M.R., Arrese, E.L., Cao, X., Chen, Y.-R.R., Chellapilla, S., Goldsmith, M.R., Grosse-Wilde, E., Heckel, D.G., Herndon, N., Jiang, H., Papanicolaou, A., Qu, J., Soulages, J.L., Vogel, H., Walters, J., Waterhouse, R.M., Ahn, S.-J.J., Almeida, F.C., An, C., Aqrawi, P., Bretschneider, A., Bryant, W.B., Bucks, S., Chao, H., Chevignon, G., Christen, J.M., Clarke, D.F., Dittmer, N.T., Ferguson, L.C., Garavelou, S., Gordon, K.H., Gunaratna, R.T., Han, Y., Hauser, F., He, Y., Heidel-Fischer, H., Hirsh, A., Hu, Y., Jiang,

H., Kalra, D., Klinner, C., König, C., Kovar, C., Kroll, A.R., Kuwar, S.S., Lee, S.L., Lehman, R., Li, K., Li, Z., Liang, H., Lovelace, S., Lu, Z., Mansfield, J.H., McCulloch, K.J., Mathew, T., Morton, B., Muzny, D.M., Neunemann, D., Ongeri, F., Pauchet, Y., Pu, L.-L.L., Pyrousis, I., Rao, X.-J.J., Redding, A., Roesel, C., Sanchez-Gracia, A., Schaack, S., Shukla, A., Tetreau, G., Wang, Y., Xiong, G.-H.H., Traut, W., Walsh, T.K., Worley, K.C., Wu, D., Wu, W., Wu, Y.-Q.Q., Zhang, X., Zou, Z., Zucker, H., Briscoe, A.D., Burmester, T., Clem, R.J., Feyereisen, R., Grimmelikhuijzen, C.J., Hamodrakas, S.J., Hansson, B.S., Huguet, E., Jermiin, L.S., Lan, Q., Lehman, H.K., Lorenzen, M., Merzendorfer, H., Michalopoulos, I., Morton, D.B., Muthukrishnan, S., Oakeshott, J.G., Palmer, W., Park, Y., Passarelli, A.L., 2016. Multifaceted biological insights from a draft genome sequence of the tobacco hornworm moth, Manduca sexta. Insect biochemistry and molecular biology.

Kleino, A., Silverman, N., 2013. The Drosophila IMD pathway in the activation of the humoral immune response. Developmental and comparative immunology.

Lemaitre, B., Hoffmann, J., 2007. The host defense of Drosophila melanogaster. Annu Rev Immunol 25, 697-743.

Matsumoto, H., Tsuzuki, S., Date-Ito, A., Ohnishi, A., Hayakawa, Y., 2012. Characteristics common to a cytokine family spanning five orders of insects. Insect biochemistry and molecular biology 42, 446-454.

Miura, K., Kamimura, M., Aizawa, T., Kiuchi, M., Hayakawa, Y., Mizuguchi, M., Kawano, K., 2002. Solution structure of paralytic peptide of silkworm, Bombyx mori. Peptides 23, 2111-2116.

Mutanen, M., Wahlberg, N., Kaila, L., 2010. Comprehensive gene and taxon coverage elucidates radiation patterns in moths and butterflies. Proceedings. Biological sciences / The Royal Society 277, 2839-2848.

Nakatogawa, S., Oda, Y., Kamiya, M., Kamijima, T., Aizawa, T., Clark, K.D., Demura, M., Kawano, K., Strand, M.R., Hayakawa, Y., 2009. A Novel Peptide Mediates Aggregation and Migration of Hemocytes from an Insect. Current Biology 19, 779-785.

Ninomiya, Y., Hayakawa, Y., 2007. Insect cytokine, growth-blocking peptide, is a primary regulator of melanin-synthesis enzymes in armyworm larval cuticle. The FEBS journal 274, 1768-1777.

Pasupuleti, M., Schmidtchen, A., Malmsten, M., 2012. Antimicrobial peptides: key components of the innate immune system. Critical reviews in biotechnology 32, 143-171.

Qiao, C., Li, J., Wei, X.-H., Wang, J.-L., Wang, Y.-F., Liu, X.-S., 2014. SRP gene is required for Helicoverpa armigera prophenoloxidase activation and nodulation response. Developmental & Comparative Immunology 44.

Rieu, I., Powers, S.J., 2009. Real-time quantitative RT-PCR: design, calculations, and statistics. The Plant cell 21, 1031-1033.

Safia, D., Nicolas, M., Aidan, B., Stefanie, M., Cordula, K., Delphine, G.-A., Catherine, D., Christophe, A., Jules, A.H., Jean-Luc, I., 2008. The DExD/H-box helicase Dicer-2 mediates the induction of antiviral activity in drosophila. Nature immunology.

Shin, S.W., Park, S.S., Park, D.S., Kim, M.G., Kim, S.C., Brey, P.T., Park, H.Y., 1998. Isolation and characterization of immune-related genes from the fall webworm, Hyphantria cunea, using PCR-based differential display and subtractive cloning. Insect biochemistry and molecular biology 28, 827-837.

Skinner, W.S., Dennis, P.A., Li, J.P., Summerfelt, R.M., Carney, R.L., Quistad, G.B., 1991. Isolation and identification of paralytic peptides from hemolymph of the lepidopteran insects Manduca sexta, Spodoptera exigua, and Heliothis virescens. The Journal of biological chemistry 266, 12873-12877.

Stenken, J.A., Poschenrieder, A.J., 2015. Bioanalytical chemistry of cytokines--a review. Anal Chim Acta 853, 95-115.

Strand, M.R., Hayakawa, Y., Clark, K.D., 2000. Plasmatocyte spreading peptide (PSP1) and growth blocking peptide (GBP) are multifunctional homologs. J Insect Physiol 46, 817-824.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Molecular biology and ….

Tsuzuki, S., Ochiai, M., Matsumoto, H., Kurata, S., Ohnishi, A., Hayakawa, Y., 2012. Drosophila growth-blocking peptide-like factor mediates acute immune reactions during infectious and non-infectious stress. Scientific reports 2, 210.

Vilcek, J., Feldmann, M., 2004. Historical review: Cytokines as therapeutics and targets of therapeutics. Trends Pharmacol Sci 25, 201-209.

Wang, Y., Jiang, H., Kanost, M.R., 1999. Biological activity of Manduca sexta paralytic and plasmatocyte spreading peptide and primary structure of its hemolymph precursor. Insect biochemistry and molecular biology 29, 1075-1086.

Wang, Y., Lu, Z., Jiang, H., 2014. Manduca sexta proprophenoloxidase activating proteinase-3 (PAP3) stimulates melanization by activating proPAP3, proSPHs, and proPOs. Insect biochemistry and molecular biology 50, 82-91.

Williams, M.J., 2007. Drosophila hemopoiesis and cellular immunity. J Immunol 178, 4711-4716.

Yamaguchi, K., Matsumoto, H., Ochiai, M., Tsuzuki, S., Hayakawa, Y., 2012. Enhanced expression of stress-responsive cytokine-like gene retards insect larval growth. Insect biochemistry and molecular biology 42, 183-192.

Yang Wang, H.J., Michael R. Kanost, 1999. Insect biochemistry and Mol Bio -Biological activity of Manduca sexta paralytic and plasmatocyte spreading peptide and primary structure of its hemolymph precursor.

Zhang, S., Cao, X., He, Y., Hartson, S., Jiang, H., 2014. Semi-quantitative analysis of changes in the plasma peptidome of Manduca sexta larvae and their correlation with the transcriptome variations upon immune challenge. Insect biochemistry and molecular biology 47, 46-54.

VITA

XIAOLONG CAO

Candidate for the Degree of

Doctor of Philosophy

Thesis: RNA-SEQ ASSISTED GENE MODELING AND ANNOTATION, TRANSCRIPTOME STUDY AND FUNCTIONAL ANALYSIS OF STRESS RESPONSIVE PEPTIDES OF *MANDUCA SEXTA*

Major Field: Biochemistry and Molecular Biology

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Biochemistry and Molecular Biology at Oklahoma State University, Stillwater, Oklahoma in December, 2016.

Completed the requirements for the Master of Science Biochemistry and Molecular Biology at Oklahoma State University, Stillwater, Oklahoma in 2015.

Completed the requirements for the Bachelor of Science in Biological Science at University of Science and Technology of China, Hefei, Anhui, China in 2011.

Experience:

Doctoral Research at Oklahoma State University from 2011 to 2016.
Undergraduate Research at University of Science and Technology of China from 2009 to 2011.

Professional Memberships:

ESA (Entomological Society of America)
OCEA (Overseas Chinese Entomologists Association)