



Phylogenetic analysis and expression profiling of the pattern recognition receptors: Insights into molecular recognition of invading pathogens in *Manduca sexta*

Xiufeng Zhang^{a,1}, Yan He^{a,1}, Xiaolong Cao^a, Ramesh T. Gunaratna^a, Yun-ru Chen^b, Gary Blissard^b, Michael R. Kanost^c, Haobo Jiang^{a,*}

^a Department of Entomology and Plant Pathology, Oklahoma State University, Stillwater, OK 74078, USA

^b Boyce Thompson Institute, Cornell University, Ithaca, NY 14853, USA

^c Department of Biochemistry and Molecular Biophysics, Kansas State University, Manhattan, KS 66506, USA

ARTICLE INFO

Article history:

Received 14 December 2014

Accepted 2 February 2015

Available online 18 February 2015

Keywords:

Insect immunity
Hemolymph proteins
RNA-Seq
Transcriptome
Genome annotation

ABSTRACT

Pattern recognition receptors (PRRs) detect microbial pathogens and trigger innate immune responses. Previous biochemical studies have elucidated the physiological functions of eleven PRRs in *Manduca sexta* but our understanding of the recognition process is still limited, lacking genomic perspectives. While 34 C-type lectin-domain proteins and 16 Toll-like receptors are reported in the companion papers, we present here 120 other putative PRRs identified through the genome annotation. These include 76 leucine-rich repeat (LRR) proteins, 14 peptidoglycan recognition proteins, 6 EGF/Nim-domain proteins, 5 β -1,3-glucanase-related proteins, 4 galectins, 4 fibrinogen-related proteins, 3 thioester proteins, 5 immunoglobulin-domain proteins, 2 hemocytins, and 1 Reeler. Sequence alignment and phylogenetic analysis reveal the evolution history of a diverse repertoire of proteins for pathogen recognition. While functions of insect LRR proteins are mostly unknown, their structure diversification is phenomenal: In addition to the Toll homologs, 22 LRR proteins with a signal peptide are expected to be secreted; 18 LRR proteins lacking signal peptides may be cytoplasmic; 36 LRRs with a signal peptide and a transmembrane segment may be non-Toll receptors on the surface of cells. Expression profiles of the 120 genes in 52 tissue samples reflect complex regulation in various developmental stages and physiological states, including some likely by Rel family transcription factors via κ B motifs in the promoter regions. This collection of information is expected to facilitate future biochemical studies detailing their respective roles in this model insect.

© 2015 Elsevier Ltd. All rights reserved.

Abbreviations: PG, peptidoglycan; LPS, lipopolysaccharide; LTA, lipoteichoic acid; PRR, pattern recognition receptor; PGRP, peptidoglycan recognition protein; TIR, Toll/interleukin-1 receptor; β GRP, β -1,3-glucanase related protein; MBP, microbe binding protein; GNBP, Gram-negative bacteria-binding protein; proPO, prophenoloxidase; ELRRP and ILRRP, extra- and intracellular leucine-rich repeat protein; TMP, transmembrane protein; TEP, thioester protein; FREP, fibrinogen-related protein; Dscam, Down syndrome cell adhesion molecule; EGF, epidermal growth factor; Nim, Nimrod; CF, control fat body; CH, control hemocytes; IF, induced fat body; IH, induced hemocytes.

* Corresponding author. Tel.: +1 405 744 9400; fax: +1 405 744 6039.

E-mail address: haobo.jiang@okstate.edu (H. Jiang).

¹ These authors contribute equally to this work.

1. Introduction

Pathogen recognition is the essential first step of effective immune responses. Since insects possess only innate immunity, they must rely on a repertoire of germline encoded proteins to recognize diverse groups of invading pathogens from the environment (Gillespie et al., 1997; Lemaitre and Hoffmann, 2007). After pathogenic bacteria, fungi, protozoa, and metazoa cross the external or internal barrier lining insect body (e.g. cuticle, trachea, midgut) and enter the hemocoel, they encounter soluble or membrane-bound receptor proteins that may bind and trigger humoral and cellular immune responses. Humoral responses involve the activation of a serine protease system to generate phenoloxidase for melanization and cytokines for intracellular signaling to induce the synthesis of antimicrobial proteins and other immune factors (Jiang et al., 2010).

Hemocytes can sense the pathogen presence through receptors on their surface, leading to phagocytosis, nodulation and encapsulation responses that eliminate or sequester the invading organisms (Strand et al., 2008). Hemocytes and fat body produce receptors, mediators, regulators and effectors of the innate immune system. The receptor proteins recognize conserved molecular patterns of microbes (e.g. peptidoglycans (PGs), lipopolysaccharide (LPS), lipoteichoic acid (LTA), β -1,3-glucan), therefore, termed pattern recognition receptors (PRRs) (Pal and Wu, 2009; Yu et al., 2002). In comparison, the recognition of protozoan and metazoan parasites or parasitoid wasps is less understood, but binding to glycolipids and glycoproteins is considered to be critical (McGuinness et al., 2003; Schmidt et al., 2001). The PRR structures, functions, and mechanisms have been investigated in various insects. In *Manduca sexta*, hemolin, peptidoglycan recognition protein-1 (PGRP1), β -1,3-glucanase-related proteins (β GRPs 1–3), microbe binding protein (MBP), immunlectin-1–4, and leureptin-1 were found to be PRRs (Jiang et al., 2010; Zhu et al., 2010; Wang et al., 2011; Rao et al., 2014). In other arthropods, PGRPs, β GRPs, C-type lectins, galectins, Leu-rich repeat (LRR) proteins, Nimrods, fibrinogen-related proteins (FREPs), thioester proteins (TEPs), hemocytins, Dscam, and Reeler may recognize pathogens or parasites (Pal and Wu, 2009; Yassine and Osta, 2010; Wang et al., 2005; Estévez-Lao and Hillyer, 2014). Based on the experimental data, genome analyses have uncovered putative PRR genes in *Drosophila melanogaster*, *Anopheles gambiae*, *Apis mellifera*, *Tribolium castaneum*, and *Bombyx mori* (Christophides et al., 2002; Evans et al., 2006; Zou et al., 2007; Tanaka et al., 2008).

As a model for insect biochemical research, *M. sexta* has substantially contributed to our understanding of insect immunity including pathogen recognition. This species also represents a model for serious lepidopteran agricultural pests that can be controlled by entomopathogens or parasitoid wasps. Therefore, knowledge on how PRRs recognize infectious agents may be useful in the development of biological control measures. To establish a solid foundation for future *M. sexta* PRR research, we took advantage of its recently determined genome sequence and RNA-Seq datasets (<http://www.ncbi.nlm.nih.gov/nucore/384358938>) to identify putative PRR genes, study evolutionary relationships with their homologs from other insects, and examine their expression profiles in 52 tissue samples taken at various developmental stages. We also explored putative immune responsive elements in the promoter regions of these genes and checked whether presence of these elements correlates with the mRNA and protein level changes in larval hemolymph before and after immune challenge (Zhang et al., 2011, 2014). In summary, we have generated a collection of basic information on these putative PRRs to facilitate the elucidation of their physiological functions in the future.

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 Set (OGS) 1.0 were downloaded from *Manduca* Base (<ftp://ftp.bioinformatics.ksu.edu/pub/Manduca/>) (X et al., 2015). The PRR sequences (Gunaratna and Jiang, 2013) were used as queries to search *M. sexta* Cufflinks Assembly 1.0b (http://darwin.biochem.okstate.edu/blast/blast_links.html) (Cao and Jiang, 2015) and OGS 1.0 using the TBLASTN algorithm with default settings. Hits with aligned regions longer than 30 residues and identity over 40% were retained for retrieving corresponding cDNA sequences. Correct open reading frames in the retrieved sequences were identified using ORF Finder (<http://www.ncbi.nlm.nih.gov/gorf/>

<http://www.ncbi.nlm.nih.gov/gorf.html>). Errors resulting from problematic regions (e.g. NNN ...) in the genome assembly were corrected after BLASTN search of *Manduca* Oases and Trinity Assemblies 3.0 of RNA-Seq data (http://darwin.biochem.okstate.edu/blast/blast_links.html). The two genome-independent RNA-Seq assemblies (Cao and Jiang, 2015) were developed to cross gaps between genome scaffolds/contigs and detect errors in the gene models. After the manual improvement, all sequences were further validated by BLASTP homolog search of GenBank (<http://www.ncbi.nlm.nih.gov/>) and then incorporated into OGS 2.0. Conserved domains and transmembrane (TM) segments were identified using SMART (http://smart.embl-heidelberg.de/smart/set_mode.cgi) and InterProScan (<http://www.ebi.ac.uk/Tools/pfa/ipscan/>). The domain architectures were plotted using DOG 2.0 (<http://dog.biocuckoo.org/>). Signal peptides were predicted using SignalP4.1 (<http://www.cbs.dtu.dk/services/SignalP/>) (Petersen et al., 2011).

2.2. Multiple sequence alignment and phylogenetic analysis

Multiple sequence alignments of PRRs from *M. sexta* and other insects (<http://www.ncbi.nlm.nih.gov/>) were performed using MUSCLE, a module of MEGA 6.0 (<http://www.megasoftware.net/>) (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 and promoter analysis

The 52 *M. sexta* cDNA libraries, representing mRNA samples from whole insects, organs or tissues at various life stages, were constructed and sequenced by Illumina technology (*Manduca sexta* genome and transcriptome project; <http://www.ncbi.nlm.nih.gov/bioproject/PRJNA81039>). Reads from the individual RNA-Seq datasets were trimmed to 50 bp and mapped to the updated OGS 1.0 using Bowtie (0.12.8) (Langmead et al., 2009). Numbers of the mapped reads were used to calculate FPKM (fragments per kilobase of exon per million fragments mapped) by RSEM (1.2.12) (Li and Dewey, 2011) for interlibrary comparisons. Hierarchical clustering of the $\log_2(\text{FPKM}+1)$ 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 changes after immune challenge, the entire set of PRR sequences were used as queries 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). Potential transcription factor binding sites in the 1000 bp region before the translation initiation site were searched using MacVector Sequence Analysis Software (Oxford Molecular Ltd.). Sequences, positions, and strand polarities of the perfectly matched GATA (WGATAR), R1 (KKGNNTTTY), and CATTW boxes were documented. NF- κ B motifs (GGGRAYYYYY) with 0, 1 or 2 mismatches were also identified.

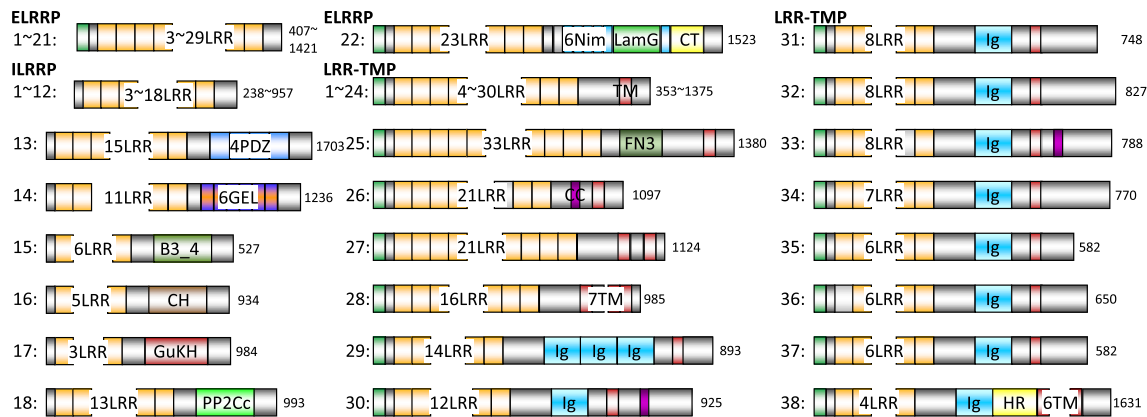


Fig. 1. Domain architectures of the *M. sexta* LRRPs. Signal peptide, LRR and TM region are in green, yellow and red, respectively. Low complexity (LC) regions within the gray areas are not shown. Protein sizes or size ranges are indicated at the end of each bar. The domain, repeat and protein sizes are not in proportion. Other domains including Nimrod (Nim), laminin G, CT (C-terminal cystine knot-like), PDZ (PSD95-Dlg-ZO1/2), GEL (Ca²⁺ and actin-binding gelsolin-like), B3_4 (RNA binding), CH (actin-binding calponin homolog), GuKH (guanylate kinase homolog), PP2Cc (Ser/Thr phosphatase, family 2C, catalytic), FN3 (fibronectin type III), CC (coiled coil), Ig (immunoglobulin), and HR (hormone receptor) domains are in different colors. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3. Results and discussion

3.1. Leucine-rich repeat proteins (LRRPs)

LRRPs constitute one of the largest protein families in insects. Of the 92 LRRP genes identified in the *M. sexta* genome, 16 encode Toll-like receptors, LRR transmembrane proteins (TMPs) with an intracellular Toll/interleukin-1 receptor (TIR) homology domain (Cao et al., 2015a). We divided the other 76 LRRPs into three groups based on their predicted cellular locations: 22 LRRPs containing a signal peptide are expected to be soluble extracellular proteins (E), 18 lack a signal peptide and are likely intracellular (I), and 36 LRR-TMPs with a putative transmembrane segment but lacking a TIR domain are expected to be integral membrane proteins with the LRRs on the cell exterior (Fig. 1). In addition to LRRs, some of these proteins contain other domains. For instance, ELRRP1–21 encompass 407 to 1421 residues, with 3–28 LRRs, some separated by regions that do not contain known sequence motifs. ELRRP2–4 (i.e. leureptin-2–4) are closely related to *M. sexta* leureptin-1 (i.e. ELRRP1), a PRR that recognizes LPS (Zhu et al., 2010). ELRRP22 (1523 residues long) has 23 LRRs, 7 Nimrod domains, 1 laminin G domain, and 1 CT domain (Fig. 1). The eighteen ILRRPs lacking a secretion signal peptide may reside in cytosol. ILRRP1–12 (238–957 residues) are simple, with 3–18 LRRs; ILRRP13–18 (527–1703 residues) are more complex, containing 3 to 15 LRRs and other structural units such as PDZ, GEL, B3_4, CH, GuKH, and PP2Cc domains. Likewise, LRR-TMP1–24 (353–1375 residues) have a simple domain organization: 4 to 30 LRRs followed by a TM region; LRR-TMP25–38 (582–1631 residues) contain other domains/regions (e.g. Ig, coiled coil, FN3, HR). Instead of having a single TM region, LRR-TMP27, 38 and 28 have 2, 6 and 7 TM segments, respectively (Fig. 1). LRR-TMP38 has a hormone receptor (HR) domain.

To study the phylogenetic relationships of the *Manduca* LRRP sequences with homologous proteins from other insect species, we searched FlyBase, VectorBase, BeetleBase and SilkBase using the *M. sexta* LRRP sequences as queries and identified 73, 130, 90 and 45 homologs in *D. melanogaster*, *A. gambiae*, *T. castaneum* and *B. mori*, respectively. Sequence alignment of the 416 LRRPs (data not shown) revealed major lineage-specific expansions of the gene family mostly in *A. gambiae* and *T. castaneum*, consistent with the fact that many families of PRRs, extracellular signal mediators/modulators (e.g. serine protease-like proteins, serpins), and

effectors (e.g. antimicrobial proteins) have undergone gene duplication and sequence divergence (d'Alencón et al., 2013). After removing branches that are not closely related to the *M. sexta* ones, we generated a simplified tree of 248 sequences (Fig. S1), showing that *M. sexta* ELRRP, ILRRP, and LRR-TMP genes form three main groups. Separation of these groups and further removal of loosely connected branches yielded three trees exhibiting strong phylogenetic relationships (Fig. 2).

M. sexta ELRRP1–4 (leureptin-1–4, containing 10 to 11 LRRs) are results of a lineage-specific family expansion. Their genes form a cluster located on Scaffold 00089. ELRRP5, 6, 8, 13, 18 and 19 (with 7–9 LRRs) arose after the radiation of the insect orders but before the divergence of *M. sexta* and *B. mori* (Fig. 2). ELRRP7, 10, 16/17, 20/21, 22 have orthologs in the beetle, fly, mosquito and silkworm, indicating divergence due to speciation. Speciation may also account for the orthologous groups of ILRR8, 10–14, LRR-TMP5, 9, 11, 12, 18, 22, 23, 27, 34, 35/37, and 36. Alignment of LRRPs can be misleading due to sequence hypervariation in the non-Leu positions of repeats (data not shown). To further investigate the observed orthologous relationships, we also examined the existence of predicted signal peptide, TM region, other domains, and the number and distribution of LRRs in members of each group. We found in nearly all cases the results support the orthology described above. Five gene pairs likely representing recent duplications (ELRRP20–21, LRR-TMP5–13, 20–26, 24–25, and 35–37) are located on the same scaffolds and have high sequence identity. Function of the members of each orthologous group is probably conserved. Interestingly, our analysis indicated that LRR-TMP14 and 22 contain a TIR domain whose score was below the SMART cutoff value (data not shown). We think, in light of possible function conservation, it is worth testing their roles as Toll-like receptors.

LRRPs, present in animals, plants, and certain microbes (Kobe and Kajava, 2001), participate in molecular binding, signal transduction, and cell adhesion. In plants, ILRRPs with nucleotide binding domains and LRR-TMPs are critical for pathogen recognition and intracellular signaling (Padmanabhan et al., 2009). In animals including insects, Toll-like receptors are the best known LRRPs involved in immune responses (Leulier and Lemaitre, 2008; Pal and Wu, 2009). The heterodimer of *A. gambiae* APL1C and LRIM1 directs TEP1 deposition and subsequent attack on the malaria parasites (Povelones et al., 2009). These two LRRPs arose from lineage-specific family expansion, and no orthologs were identified in *M. sexta* (data not shown). *M. sexta* leureptin-1 binds to LPS and

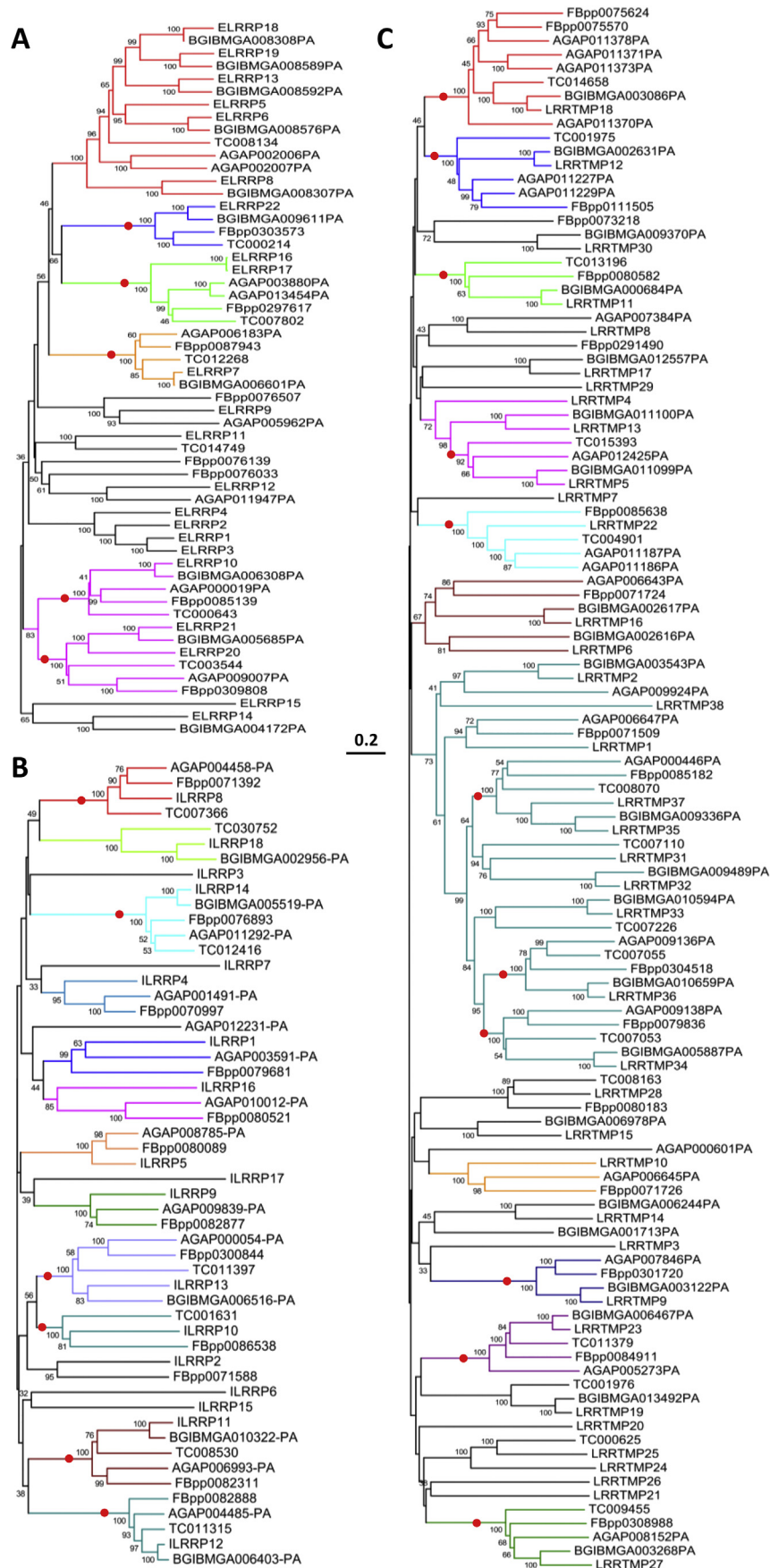


Fig. 2. Phylogenetic relationships of *M. sexta* ELRRPs (A), ILRRPs (B), LRR-TMPs (C) and their close homologs in the other insects. Based on the preliminary analysis (Fig. S1), the three groups of *M. sexta* LRR sequences were separately compared with their homologs from *D. melanogaster* (FB), *A. gambiae* (Ag), *B. mori* (BGIBM), and *T. castaneum* (Tc). Branches in each closely related group (bootstrap value >700 in 1000 trials, ≥ 3 members) are shown in the same color. Based on stricter standards, each orthologous group (denoted by red spheres) has to exceed the bootstrap limit of 900 and contain one or two species in Coleoptera, Lepidoptera and Diptera. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

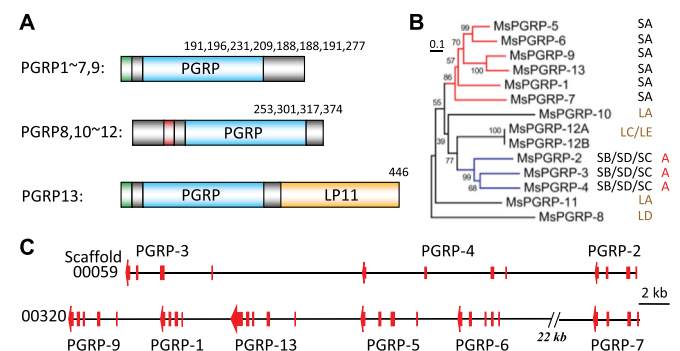


Fig. 3. Domain organization (A), evolutionary relationships (B) and gene clusters (C) of *M. sexta* PGRPs. Signal peptide, TM region, PGRP domain and lepidopteran low M_r lipoprotein 11 domain are in green, red, cyan and yellow, respectively. Protein sizes are indicated at the end of each bar. Domain and protein sizes are not in proportion. The orthologous *D. melanogaster* PGRPs are indicated. *M. sexta* PGRP2–4 contain the key residues (H, Y, H, T, C) for the amidase (“A”) activity. The exons are shown as red bars; gene orientation is marked on the last exon. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

certain hemocytes to perhaps enhance cellular responses (Zhu et al., 2010). These results indicate the roles of LRRPs in insect immunity and other physiological processes are poorly understood but worth exploring. The orthology and lineage-specific expansion provide leads for their functional analyses in the future.

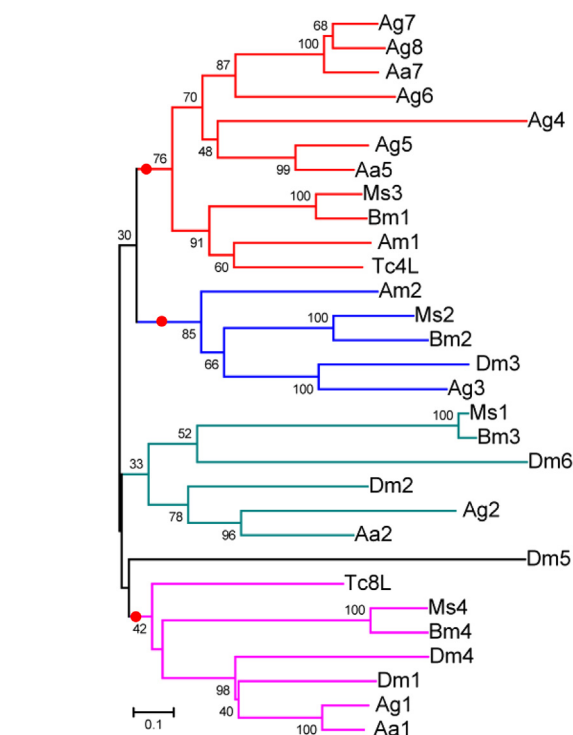


Fig. 4. Phylogenetic relationships of the insect galectins. *M. sexta* galectin-1–4 (Ms1–Ms4), *A. aegypti* galectin-1 (Aa1, GI|157136839), 2 (Aa2, GI|157131540), 5 (Aa5, GI|157138456), 7 (Aa7, GI|157109508), *D. melanogaster* galectin-1 (Dm1, CG11372), 2 (Dm2, CG5335), 3 (Dm3, CG18565), 4 (Dm4, CG11374), 5 (Dm5, CG14879), 6 (Dm6, CG13950), *B. mori* galectin-1–4 (Bm1–Bm4), *A. mellifera* galectin-1 (Am1), 2 (Am2), *T. castaneum* galectin-4 like (Tc4L), 8 like (Tc8L) sequences are aligned to generate the tree. Branches in each closely related group (bootstrap value >700 in 1000 trials, ≥ 3 members) are shown in the same color. Based on the bootstrap value and domain architecture, three orthologous groups (denoted by red spheres) are identified, each containing one or more species in each of three insect orders. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

3.2. Peptidoglycan recognition proteins (PGRPs)

PGRPs are a family of homologous molecules that recognize bacterial peptidoglycans. Based on the *Drosophila* naming system (Werner et al., 2000), the short (S) PGRPs have a classic signal peptide and are secreted, whereas the long (L) PGRPs lack a signal peptide and are associated with cell membrane or reside in cytosol. We identified 13 PGRP genes in the *M. sexta* genome (Fig. 3A). PGRP1, 5–7, 9, 13 are similar to *D. melanogaster* SA, PGRP2–4 to SB/SC/SD, PGRP8 to LD, PGRP10/11 to LA, and PGRP12 A/B to LC/LE (Fig. 3B). The SA homologs appear to be products of a family expansion resulting in a cluster of six genes (PGRP9, 1, 13, 5–7) on Scaffold 00320 and another cluster of three genes (PGRP2–4) on Scaffold 00059 (Fig. 3C). PGRP13 contains a lepidopteran low M_r lipoprotein-11 domain, which has not been found in any PGRP before. PGRP12A and 12B are derived by alternative splicing from the same gene with five exons. The exon 3 is removed during RNA splicing in PGRP12A production, whereas its inclusion yields PGRP12B with two instead of one TM region. Consequently, the C-terminal PGRP domain of 12B may be located inside rather than outside the cell (as for 12A). Some PGRPs are simple PG binding proteins, others may have a catalytic activity to cleave the lactylamide bond between *N*-acetylmuramic acid residues on the glycan strand and *L*-alanine in the crosslinking peptides (Mellroth et al., 2003). Thus, we examined and found PGRP2–4 contain His, Tyr, His, Thr, and Cys residues at conserved positions, which are essential for the amidase activity. PGRP2 and PGRP3 are expressed in midgut (see below).

Our previous study indicated that *M. sexta* PGRP1 specifically binds to Lys-type PG from *Micrococcus luteus* and meso-diaminopimelic acid-PGs from *Escherichia coli* or *Bacillus megaterium* (Sumathipala and Jiang, 2010). The recognition triggers a serine

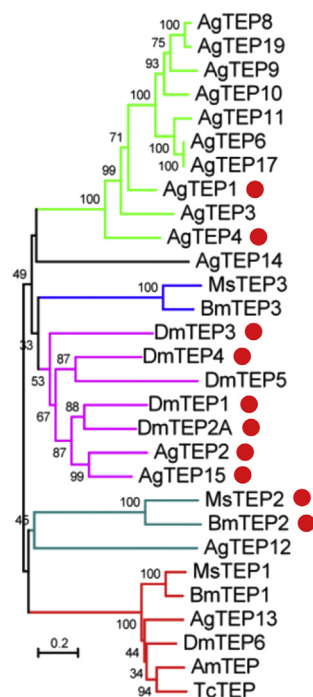


Fig. 5. Phylogenetic relationships of the insect thioester proteins. *M. sexta* TEP1–3, *B. mori* TEP1–3, *D. melanogaster* TEP1, 2A, 3–6, *A. gambiae* TEP1–4, 6, 8–15, 17, 19, *A. mellifera* TEP, and *T. castaneum* TEP sequences are aligned to build the tree. Branches in each closely related group (bootstrap value >450 in 1000 trials) are shown in the same color. The ones with conserved GCCEQ are marked by a red sphere. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

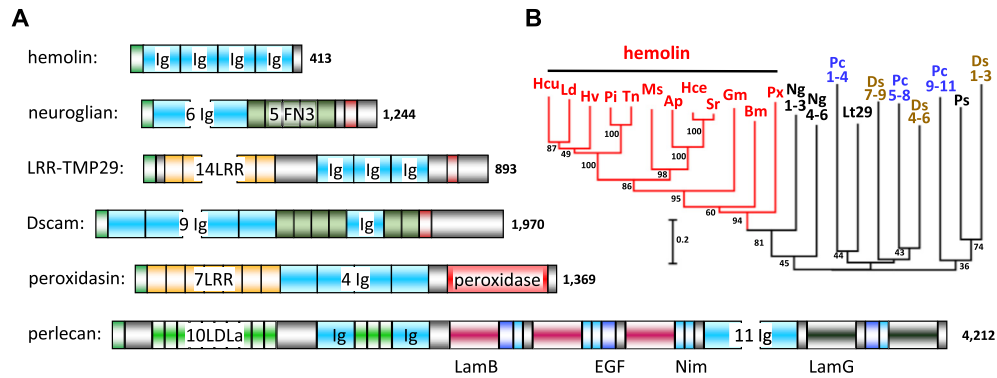


Fig. 6. Domain organization (A) and evolutionary relationships (B) of proteins containing multiple Ig domains. (A) Signal peptide, TM region, Ig domains, and Leu-rich repeats are in green, red, cyan and yellow, respectively. Other domains including peroxidase, FN3, Nim, EGF, laminins B and G, and LDLa are shown in different colors. Protein sizes are indicated at the end of each bar. Domain and protein sizes are not in proportion. (B) Sequences of hemolins from *Antheraea pernyi* (Ap), *B. mori* (Bm), *Galleria mellonella* (Gm), *H. cecropia* (Hce), *Hyphantria cunea* (Hcu), *Heliothis virescens* (Hv), *Lymantria dispar* (Ld), *M. sexta* (Ms), *Pseudoplusia includens* (Pi), *Plutella xylostella* (Px), *Samia ricini* (Sr), *Trichoplusia ni* (Tn) are aligned with Dscam Ig domains 1–3 (Ds1–3), 4–6 (Ds4–6), 7–9 (Ds7–9), LRR-TMP29 Ig domains 1–3 (Lt29), neuroglian Ig domains 1–3 (Ng1–3), 4–6 (Ng4–6), perlecan Ig domains 1–4 (Pc1–4), 5–8 (Pc5–8), 9–11 (Pc9–11) and peroxidasin Ig domains 1–4 (Ps), all from *M. sexta*. All hemolins form a monophyletic group. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

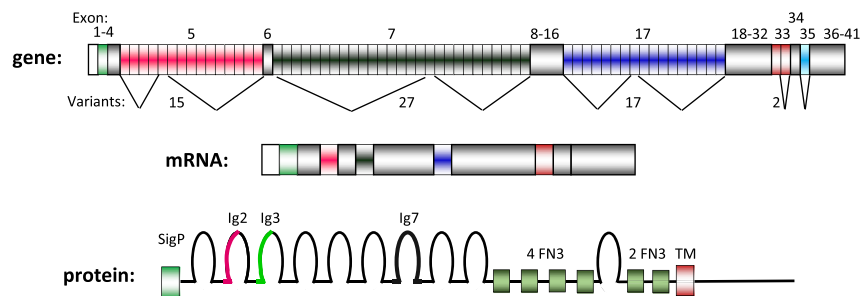


Fig. 7. Gene structure, pre-mRNA splicing, and protein domain organization of *M. sexta* Dscam. Exons of the gene are shown as (fused) vertical bars, introns are omitted. Exons 5 (pink), 7 (dark green), 17 (blue) and 33 (red) have 15, 27, 17 and 2 variants, respectively. Exon 35 is optional. Mutually exclusive alternative splicing generates a population of mature mRNAs with different sequences. The hypervariable regions in the Ig domains 2, 3, and 7 are encoded by exons 5, 7 and 17 while signal peptide, Ig8-Ig9-4FN3-Ig10-2FN3 and TM regions are encoded by exons 2, 18–32 and 33, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

protease cascade that activates prophenoloxidase (proPO). *M. sexta* PGRP2 and 3 also recognize PGs and modulate melanization (Zhang et al., unpublished data). *B. mori* PGRP, as the first insect PRR characterized, binds PG and elicits the melanization cascade

(Yoshida et al., 1996). Systematic studies of *D. melanogaster* PGRPs reveal their implications in various immune responses, such as phagocytosis (PGRP-LC, SC1), negative regulation (PGRP-SB1, SC1, LB, LF), melanization (PGRP-SA, LE), and activation of Toll, IMD and

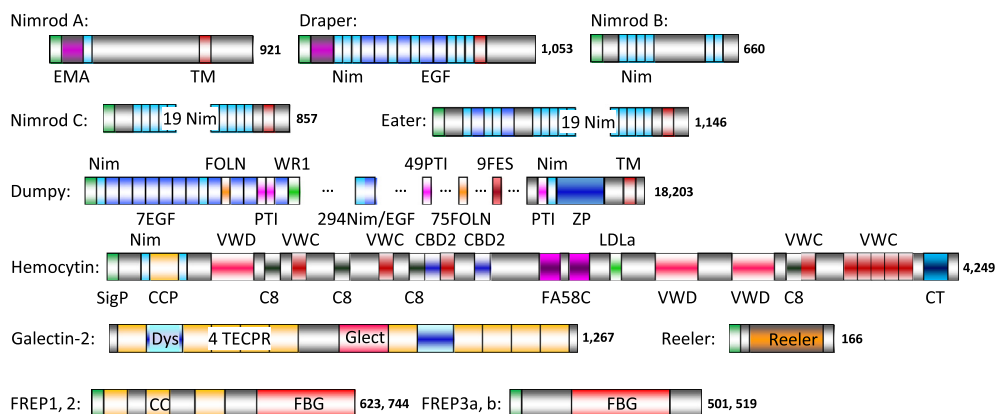
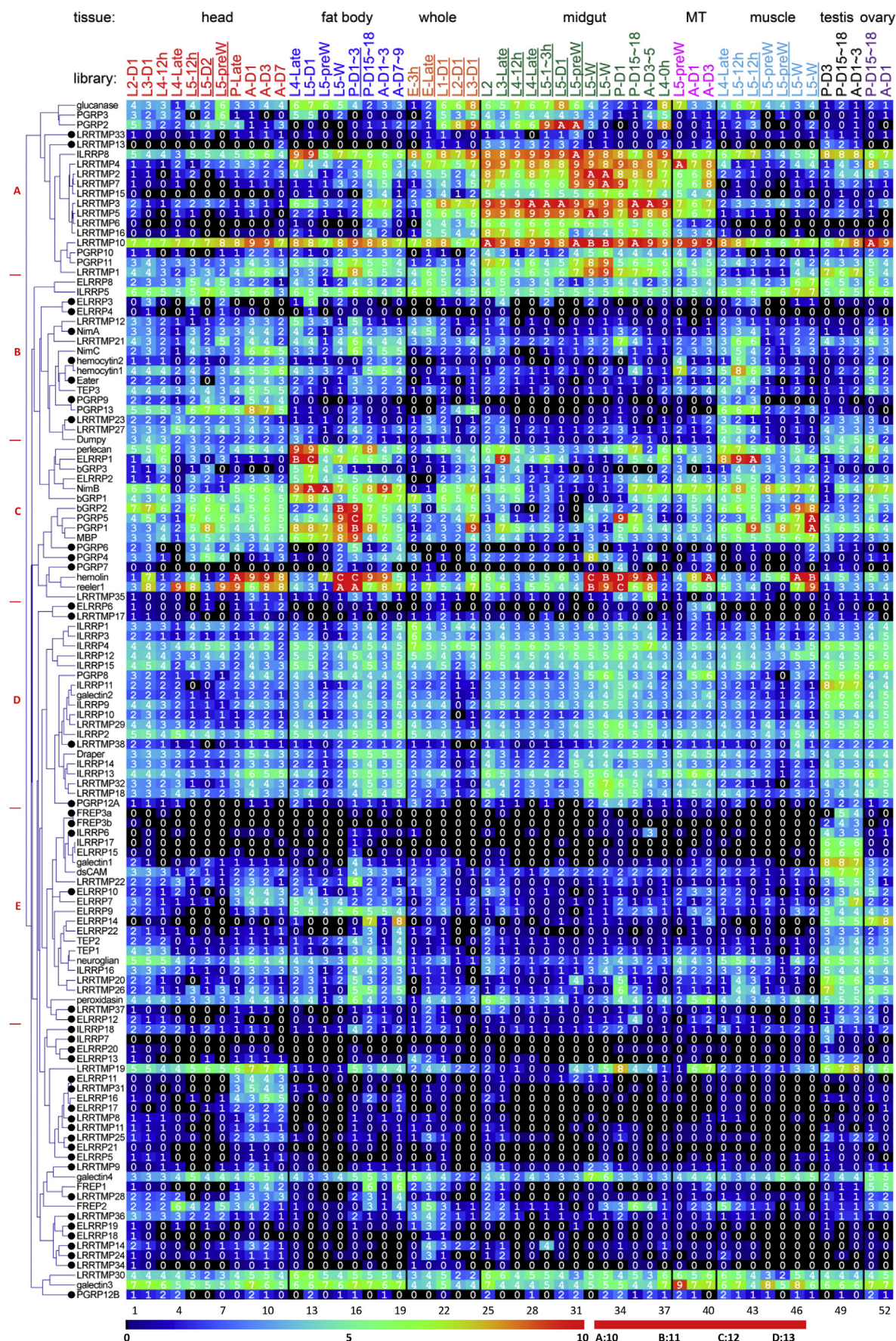


Fig. 8. Domain architectures of other putative PRRs in *M. sexta*. Signal peptide, Nim, EGF and TM region are in green, cyan, blue and red, respectively. Low complexity (LC) regions within the gray areas are not shown. Protein sizes are indicated at the end of each bar. The domain, repeat and protein sizes are not in proportion. Other domains are in different colors, including FOLN (follistatin N-terminal), PTI (plant trypsin inhibitor), WR1 (worm-specific repeat-1), FES (Fe–S binding), ZP (Zona pellucida), CCP (complement control protein), VWC/VWD (von Willebrand factor C/D), C8 (eight Cys), CBD2 (chitin binding-2), FA58C (coagulation factor 5/8 C-terminal), LDLa (low-density lipoprotein receptor repeat-a), CT (C-terminal cystine knot-like), Dys (dysferlin N-terminal), TECPR (teconin β -propeller repeats), Glect (galectin), Reeler, CC (coiled coil), FBG (fibrinogen β and γ chains) domains. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



JNK pathways (PGRP-SA, SD, LE, LC) (Charroux et al., 2009; Ligoxygakis, 2013). Those studies shed light on future exploration of biochemical functions of *M. sexta* PGRPs.

3.3. Galectins

Insect galectins are involved in the regulation of immunity and development (Pace and Baum, 2004). They contain at least one conserved carbohydrate recognition domain (CRD) that specifically binds β -galactosides independent of Ca^{2+} . Evidence exists that *Drosophila* and sand fly galectins regulate development and defense against bacterial and parasitic pathogens (Pace and Baum, 2004; Kamhawi et al., 2004). We identified 4 galectins in *M. sexta* and 26 in five other insects, all of which lack a classic signal peptide except for galectin1 from *A. gambiae* and *Aedes aegypti*. *M. sexta* (Ms) galectin 4 and its orthologs have two CRDs (Fig. 4). MsGalectin1 and BmGalectin3 have one CRD but are most similar to DmCG13950, DmCG5335, AgGalectin2 and AaGalectin2, all having two CRDs. The orthologous group containing MsGalectin3, with a single CRD, expanded in mosquitoes. MsGalectin2 and its orthologs contain a single CRD, 10 teconin β -propeller repeats and 2 dysferlin N-terminal fragments (Fig. 8). Genetic analysis of DmCG18565 may reveal conserved function of these interesting mosaic proteins.

3.4. Thioester proteins (TEPs)

Thioester proteins include some complement factors and α_2 -macroglobulin in vertebrates and TEPs in invertebrates (Blandin and Levashina, 2004). Several *D. melanogaster* TEPs are up-regulated in larvae upon bacterial immune challenge. *A. gambiae* TEP1 opsonizes bacteria for phagocytosis through covalent binding (Levashina et al., 2001), and this protein also localizes to the ookinete surface and participates in destruction and sequestration of the parasite (Blandin et al., 2004). The thioester bond between Cys and Gln side chains in GCGEQ may react with an amino or hydroxyl group on its target. We have identified three TEPs in *M. sexta*. Only TEP2 contains the GCGEQ motif. Phylogenetic analysis indicates that the *M. sexta* and *B. mori* TEPs form three orthologous pairs while *D. melanogaster* and *A. gambiae* TEPs underwent major lineage-specific expansions (Fig. 5). *B. mori* TEP1 has a C-terminal receptor binding domain which may promote hemocyte phagocytosis (Tanaka et al., 2008). Perhaps, *M. sexta* and *B. mori* are useful as biochemical models to explore the mechanisms of insect TEPs in pathogen recognition and opsonization, proteolytic activation, and association with proteins involved in phagocytosis and melanization.

3.5. β -1,3-Glucanase related proteins (β GRPs)

β -1,3-glucan is a fungal cell wall component that triggers insect immune responses. Several studies have revealed functions of proteins in *M. sexta* that bind to this polysaccharide (Ma and Kanost, 2000; Jiang et al., 2004; Wang et al., 2011; Rao, 2014;

Takahashi, 2014). β GRP1, β GRP2, and microbe binding protein (MBP, closely related to Gram-negative bacteria-binding proteins (GNBPs) from other insects) contain two domains, but β GRP3 has one. Their C-terminal β -1,3-glucanase-like domains lack the catalytic residues needed for glucan hydrolysis. The N-terminal CRD of β GRP1, β GRP2 and MBP may adopt a β -sandwich structure similar to an Ig domain (Takahashi et al., 2014). The mRNA levels of β GRP1 and β GRP3 in larval fat body did not change after an injection of killed bacteria, but those of β GRP2 and MBP increased considerably. These glucanase-like proteins bind specifically to β -1,3-glucan and other surface molecules to agglutinate microbes and elicit targeted proPO activation in the larval hemolymph. Intriguingly, *M. sexta* β GRP3 inhibits *Bacillus cereus* growth (Rao et al., 2014). The ratio of β GRP2 and laminarin in their complex affects its ability to trigger melanization (Takahashi et al., 2014). In the *M. sexta* genome, there is a β -1,3-glucanase gene, whose expression in the midgut indicates its involvement in digestion (see below). We also detected low, induced expression of the β -1,3-glucanase in hemocytes. This raises an interesting possibility that the enzyme may hydrolyze β -1,3-glucan to dampen the antifungal response. Last but not the least, the evolutionary history of insect β GRPs has been investigated in previous studies (Jiang et al., 2004; Wang et al., 2011; Rao et al., 2014).

3.6. Hemolin and other immunoglobulin (Ig) domain proteins

As more insect genomes are sequenced, many proteins containing one or more Ig domains are identified, but their possible roles in pathogen recognition are mostly unexplored. Ig domain proteins often function in cell adhesion or immunity (or both). Here, we report six proteins with tandem Ig domains from *M. sexta* (Fig. 6A). Hemolin, first isolated from hemolymph of *Hyalophora cecropia* and *M. sexta*, is highly inducible after exposure to microorganisms, and is composed of four Ig domains (Sun et al., 1990; Ladendorff and Kanost, 1990; 1991). Hemolins have only been found in Lepidoptera. Other proteins with tandem Ig domains in *M. sexta* include neuroglian, Dscam, LRR-TMP29, peroxidase, and perlecan. Sequence alignment indicates *M. sexta* hemolin is closer to its orthologs from other lepidopteran species than to other *M. sexta* Ig domain proteins (Fig. 6B). Hemolin recognizes bacteria by binding to LPS or LTA and associates with hemocytes to modulate phagocytosis and nodule formation (Jiang et al., 2010). RNAi knockdown of hemolin expression in *M. sexta* larvae caused a reduction in cellular responses against *E. coli* infection (e.g. phagocytosis and nodulation) (Eleftherianos et al., 2007). The role of hemolin in antiviral response is unclear but its activity is inhibited by a bracovirus protein CcV1 that co-localizes with hemolin on cell membranes (Labropoulou et al., 2008). *M. sexta* neuroglian resides on the surface of glial and neuronal cells, developing embryonic prothoracic gland, and a subgroup of plasmatocytes (Nardi et al., 2006). This transmembrane protein affects hemocyte attachment to other hemocytes and to foreign surfaces. *M. sexta* LRR-TMP29, peroxidase and perlecan contain 3, 4 and 13 I-

Fig. 9. Transcript profiles of the putative PRR genes in the fifty-two tissue samples. The mRNA levels, as represented by $\log_2(\text{FPKM}+1)$ 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, 11.50–12.49 and 12.50–13.49 are labeled as 0, 1, 2 ... 9, A, B, C and D, respectively. The 52 cDNA libraries (1 through 52) are constructed from the following tissues and stages: **head** [1. 2nd (instar) L (larvae), d1 (day 1); 2. 3rd L, d1; 3. 4th L, d0.5; 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), 3 h; 21. E, late; 22. 1st L; 23. 2nd L; 24. 3rd L], **midgut** [25. 2nd L; 26. 3rd L; 27. 4th L, 12 h; 28. 4th L, late; 29. 5th L, 1–3 h; 30. 5th L, 24 h; 31. 5th L, pre-W; 32–33. 5th L, W; 34. P, d1; 35. P, d15–18; 36. A, d3–5; 37. 4th L, 0 h], **Malpighian tubules (MT)** [38. 5th L, pre-W; 39. A, d1; 40. A, d3], **muscle** [41. 4th L, late; 42–43. 5th L, 12 h; 44–45. 5th L, pre-W; 46–47. 5th L, W], **testis** [48. P, d3; 49. P, d15–18; 50. A, d1–3], and **ovary** [51. P, d15–18; 52. 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. Cluster analysis has revealed 5 distinct groups (A–E), as shown on the left. The group F genes (marked by black dot) are expressed at low levels [$\log_2(\text{FPKM}+1)$: 0 to 4] in nearly all the 52 samples. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

set Ig domains (Fig. 6A), respectively. As an atypical ELRRP, peroxidase also includes a peroxidase domain and, like its orthologs in *Drosophila* (Nelson et al., 1994), may participate in extracellular matrix consolidation, phagocytosis, and defense. Perlecan, also known as Trol or basement membrane-specific heparan sulfate proteoglycan core protein, controls neuroblast proliferation (Voigt et al., 2002). Dscam is a mosaic protein with a conserved domain structure: signal peptide, nine Ig domains, four FN3, one Ig domain, two FN3, one TM region, and a cytosolic tail (Fig. 7), which can form a very complex set of isoforms by alternative exon splicing (Neves et al., 2004). The Ig domains 2, 3 and 7 are partly or completely encoded by exons 5, 7 and 17, each having multiple variants (15, 27 and 17 variants, respectively). Since exon 33 has two variants and exon 35 is optional, this gene has the potential to encode, through mutually exclusive exon splicing, $15 \times 27 \times 17 \times 2 \times 2$ or 27,540 isoforms, which might provide many binding specificities. *Drosophila* Dscam is essential in neuronal wiring (Schmucker et al., 2007) as well as hemocyte phagocytosis (Watson et al., 2005; Dong et al., 2006).

3.7. Others

We have identified seven EGF/Nim-domain proteins in *M. sexta*. Nimrod A and Draper contain a signal peptide, an EMI domain with a CCXG(Y/W) motif, a Nim domain with 6 Cys, 0 or 14 EGF domains with 8 Cys, and a TM region (Fig. 8). Nimrod B and ELRRP22 (Fig. 1) are secreted proteins with 6 and 7 Nim domains, respectively. *Holotrichia diomphalia* Nimrod B acts as a PRR for LPS in hemolymph (Ju et al., 2006). *M. sexta* Nimrod C (857 residues) is composed of a signal peptide, 19 Nim domains, and a TM region, whereas another form of Nimrod C (1146 residues) has an N-terminal extension of a signal peptide, a Nim, an EGF, two Nim and another EGF, located before the 19 Nim domains. The longer variant is similar to *Drosophila* Eater in domain structure. Some Nimrod A and C-type proteins mediate phagocytosis (Somogyi et al., 2008). The former tend to bind apoptotic cells via phosphatidylserine while Eater binds bacteria directly (Tung et al., 2013). *M. sexta* Dumpy is a 1934 kDa protein with 450 domains spanning 18,203 residues (Fig. 8). Its 308 EGF/Nim modules are interspersed with 52 PTI domains, 78 follistatin N-terminal domains, 9 Fe–S binding domains, and a worm-specific repeat. *Drosophila* Dumpy maintains tension at epidermal-cuticle attachment sites (Wilkin et al., 2000). Like Dumpy, *M. sexta* hemocytin-1 and 2 are large proteins (4249 residues, 467 kDa) with multiple structural modules: 2 Nim, 1 Sushi/CCP, 3 VWD, 9 VWC, 4C8, 2 CBD, 2 FA58C, 1 LDLA and 1 CT domains (Fig. 8). The two proteins (99.1% identical; 99.7% similar) are encoded by two genes that are 97.3% identical in the coding region. Alternative splicing may generate C-terminally truncated forms (data not shown). As a homolog of von Willebrand factor in mammals, *B. mori* hemocytin is stored in granules of granulocytes and released upon bacteria exposure to form sticky fibrous structure trapping hemocytes and microbes (Arai et al., 2013). *M. sexta* Reeler is a bacteria-induced protein with a single reeler domain following a signal peptide (Zhu et al., 2003). *B. mori* Reeler1 is expressed in hemocytes, fat body, and midgut in response to bacterial injection (Bao et al., 2011). RNA interference and injection of the recombinant protein affected nodule formation and proPO activation. In addition, there are three fibrinogen-related protein (FREPs) genes in the *M. sexta* genome. FREP1 and FREP2 contain a signal peptide, three coiled coil regions, and one FBG domain (Fig. 8). Through alternate exon usage, FREP3 gene encodes two isoforms with a signal peptide and an FBG domain. In *A. gambiae*, several FREPs exhibit high binding affinity to bacteria, fungi or parasites and certain FREPs are essential to innate immune

hemostasis (Dong and Dimopoulos, 2009). Specific recognition of glycoconjugates by the FBG domains is responsible for the binding.

3.8. Expression profiles

We further examined expression profiles of the putative PRRs in the 52 RNA-Seq libraries constructed from tissue samples collected at different developmental stages. Cluster analysis of the expression profiles revealed six groups (Fig. 9). Group A includes β -1,3-glucanase, PGRP2, 3, 10, 11, ILRRP8, LRR-TMP1–7, 10, 15, and 16. Their mRNA levels in midgut are higher than Malpighian tubules and fat body. While β -1,3-glucanase, PGRP2 and PGRP3 may digest fungal and bacterial cell walls, it is surprising that the LRR-TMP mRNAs are highly abundant in midgut. We speculate that their potential binding to commensal bacteria in the gut may diminish host immune responses against non-pathogenic bacteria in this compartment. Transcripts of the Group B genes (ELRRP8, ILRRP5, LRR-TMP12, 21, 27, Nimrod C, hemocytin-1, TEP3, PGRP13, Dumpy) in head and muscle in feeding larvae are low (2–6) but higher than fat body and other tissues. Group C (perlecan, leureptin/ELRRP1, ELRRP2, β GRP1–3, MBP, PGRP1, PGRP5, Nimrod B, hemolin, Reeler) is highly expressed (5–13) in fat body and muscle. Their transcripts are also present in other parts of the insect at different stages but the levels are lower. The mRNAs of Group D genes (ILRRP1–4, 9–15, LRR-TMP18, 19, 29, 30, 32, 38, galectin2–4, PGRP8, Draper, neuroglian, peroxidase, and FREP2) are fairly low and evenly distributed. Their levels in testis and ovary are somewhat higher. This difference is more pronounced in Group E (ELRR7, 9, 14, 15, 22, ILRRP16, 17, LRR-TMP20, 22, 26, TEP1, TEP2, galectin1, Dscam), since the mRNA levels in other tissues are lower. Transcripts of the 43 Group F genes are hardly detected (0–4) throughout the 52 libraries.

Our previous expression analyses of non-digestive serine protease-related protein, C-type lectin-domain protein, and immune effector genes revealed an interesting trend that transcription of most immunity-related genes are up-regulated in fat body and midgut from pre-wandering to early pupal stage (Cao et al., 2015b; Rao et al., 2015; He et al., 2015). Since immune function for most of the genes reported herein, especially LRRPs, lacks experimental support, we tried to use their expression patterns to predict this feature. Among the 122 genes analyzed, 38 and 53 had higher mRNA levels in fat body and midgut, respectively, during the wandering larval/pupal period compared with the feeding stage 5th instar larva. Twenty-eight of them were up-regulated in both tissues, including PGRP1, 5, 6, 11, β GRP2, MBP, β -1,3-glucanase, galectin 3, 4, hemolin, neuroglian, peroxidase, perlecan, Reeler, Nimrod C, Draper, ELRRP2, ILRRP5, 8, 13, 18, LRR-TMP1, 10, 19–21, 26 and 30. While the role of pattern recognition is established for some of these proteins (e.g. PGRP1, β GRP2, MBP, hemolin), that function remains to be experimentally examined for the others.

The examination of changes in mRNA and polypeptide levels in immune challenged *M. sexta* larvae (Zhang et al., 2011, 2014; Gunaratna and Jiang, 2013) yielded an overview of the immune system constituents. In light of the genome annotation, we reanalyzed the results on putative PRRs (Table 1) in conjunction with a search for κ B, GATA, CATTW, R1 motifs in the 1000 bp region upstream of the longest transcript. There are 57 genes possessing CATTW, 36 with 1 mismatch in the κ B motif and 57 with 2 mismatches. LRR-TMP3, 13, 23, 28, and Reeler genes have a perfect match of the κ B element, but only Reeler mRNA and protein show significant induction. Some genes containing the κ B element(s) with 1 or 2 mismatches are not up-regulated after immune challenge, while others without such element(s) are significantly induced. Similar results were obtained from the analysis of C-type lectin-domain protein genes and antimicrobial peptide genes (Rao et al., 2015; He et al., 2015). These results indicate that even the

Table 1Features of the putative *M. sexta* PRR genes, transcripts, and polypeptides.

name	length ^a	GATA	CATTW	κB -0	κB -1	κB -2	κB Σ	R1 ^b	IF/CF RNA	IH/CH RNA	IH/CH peptide	p-value peptide
ELLRP1		2	5			8	8		2.21	5.94	N.A.	N.A.
ELLRP2		5	8			3	3		0.71	0.00	N.A.	N.A.
ELLRP3		5	8		1	10	11					
ELLRP4		5	10		1	12	13					
ELLRP5	647	1	13		1	4	5	1				
ELLRP6		2	10		1	10	11	1				
ELLRP7		2	8		1	6	7					
ELLRP8		3	9		3	8	11					
ELLRP9		1	7		1	3	4	1	2.04	N.A.	0.00	0.37
ELLRP10		2	8		2	8	10					
ELLRP11		2	9		5	9	14					
ELLRP12		0	6			8	8					
ELLRP13		4	14		1	10	11		N.A.	N.A.	0.00	0.37
ELRRP15		4	8			3	3					
ELRRP16		4	7			10	10					
ELRRP17		2	6			13	13					
ELRRP19		2	11		1	8	9					
ILRRP1		6	6			12	12					
ILRRP2		6	16		2	9	11					
ILRRP4		2	9		1	9	10					
ILRRP5		7	6		1	3	4					
ILRRP6		1	10			12	12					
ILRRP7		6	10		1	8	9					
ILRRP8		2	11			9	9					
ILRRP10		3	8		7	12	19					
ILRRP11		3	8		1	15	16	1				
ILRRP12		1	8		2	10	12					
ILRRP13		4	8			7	7					
ILRRP14		3	9		1	13	14					
ILRRP15	85	0	1				0					
ILRRP16		2	8			7	7					
ILRRP17	401	2	4		1	5	6					
ILRRP18		3	15			5	5					
LRR-TMP1		1	8		3	10	13					
LRR-TMP2	628	2	3		3	5	8					
LRR-TMP3		6	9	1	5	10	16	1				
LRR-TMP4		2	6		2	10	12	1				
LRR-TMP5		3	11		1	17	18					
LRR-TMP6		2	14			8	8					
LRR-TMP7		3	4			5	5					
LRR-TMP8		1	7			10	10					
LRR-TMP9		4	8			5	5					
LRR-TMP10		3	11		1	8	9					
LRR-TMP11		4	4			7	7					
LRR-TMP12		8	7		1	12	13	1				
LRR-TMP13		5	6	1		15	16					
LRR-TMP14		1	4			7	7		N.A.	N.A.	0.00	0.37
LRR-TMP15		7	1			10	10					
LRR-TMP16		2	8			10	10					
LRR-TMP17		1	6		1	5	6					
LRR-TMP18		2	9			11	11					
LRR-TMP19		7	9		1	6	7					
LRR-TMP20		0	6		2	10	12					
LRR-TMP21		3	8		6	12	18		0.25	1.59	N.A.	N.A.
LRR-TMP23		5	9	1	1	13	15	2				
LRR-TMP24		2	7		1	7	8	1				
LRR-TMP25		3	8			8	8					

name	length	GATA	CATTW	κB 0	κB -1	κB -2	κB Σ	R1	IF/CF RNA	IH/CH RNA	IH/CH peptide	p-value peptide
LRR-TMP26		6	5		1	10	11	1				
LRR-TMP27		4	8				4	4				
LRR-TMP28		2	9	1		7	8		N.A.	N.A.	0.00	0.37
LRR-TMP29		1	10		2	17	19					
LRR-TMP30		2	5			6	6		0.68	1.82	0.00	0.37
LRR-TMP31		1	5			11	11					
LRR-TMP32		1	8		1	7	8					
LRR-TMP33		5	7			17	17					
LRR-TMP35		3	6		1	15	16					
LRR-TMP36		1	7			9	9	2				
LRR-TMP37		0	8			6	6					
LRR-TMP38		0	7			9	9					
PGRP1		5	9			18	18		5.10	64.16	1.02	0.79
PGRP2		1	4			9	9		1818.46	11.66	N.A.	N.A.
PGRP3		4	12			7	7	1	44.02	14.58	13.88	0.37
PGRP4		4	8		1	11	12	1				
PGRP5		6	7		6	18	24		54.37	10.09	10.82	0.00
PGRP6		2	10		1	8	9					
PGRP7		3	4		2	14	16					
PGRP8		4	7		1	7	8	1				
PGRP9		3	8		1	9	10	1				
PGRP10		9	8			7	7		N.A.	N.A.	0.00	0.37
PGRP11		9	8			7	7					
PGRP12		3	11			10	10	1				
PGRP13		0	9		1	5	6					
Galectin2		6	7		1	9	10					
Galectin3	919	10	14			9	9		1.02	1.71	N.A.	N.A.
Galectin4		1	16		1	17	18					
TEP2		3	10		2	10	12					
TEP3		3	9		3	11	14					
βGRP1		3	8		1	10	11		0.96	11.66	0.68	0.43
βGRP2									9.69	9.23	N.A.	N.A.
βGRP3		3	9		1	10	11	1	0.98	N.A.	N.A.	N.A.
MBP	889	3	8		2	6	8		2.05	7.12	0.97	0.66
β1,3-glucanase	621	1	4		1	4	5	1	1.27	3.56	0.40	0.02
Hemolin		2	7		3	17	20	1	748.52	3.96	5.43	0.00
Peroxidasin		5	8			14	14		1.02	0.89	N.A.	N.A.
Perlecan		5	7			8	8		0.64	1.01	10.70	0.01
Draper		3	7			5	5		5.61	1.69	0.92	0.96
Dumpy		6	7		1	10	11		N.A.	N.A.	0.00	0.37
NimA		2	14			13	13					
NimB		3	10			8	8		1.02	1.02	0.75	0.05
NimC		1	9		2	20	22					
Dscam		5	9			5	5					
reeler1		3	6	1	1	12	14		1159.95	3.02	487.70	0.00
FREP1		1	5		2	8	10	1				
FREP2		3	4		1	13	14					
FREP3a		2	5			7	7	1				
FREP3b		4	9		1	10	11	1				
Hemocytin1		5	8			6	6		0.04	0.44	N.A.	N.A.
Hemocytin2		5	8			6	6		0.04	0.44	0.00	0.37

^a length (bp) of the analyzed region before the longest transcript in Cufflink1.0: 1000 bp if unspecified.

^b numbers of the κB motif (GGGRAYYYY) with 0, 1, and 2 mismatches. R1 site, KKGNNNTTTY; GATA box, WGATAR. ^c relative abundances of the mRNA in fat body (IF/CF) and hemocytes (IH/CH) (Zhang et al., 2011). N.A., unavailable. ^d I/C ratio from the peptidome (Zhang et al., 2014) and proteome (He et al., unpublished data). *p*-value from the Student's t-test of normalized spectral counts.

loosely defined κB motif (GGGRAYYYY) from *Drosophila* may not be useful for predicting regulation of immunity-related genes in *M. sexta*. Perhaps, the responsive elements have undergone changes in this lineage and co-evolution of the DNA-binding domains allows the transcription factors respond to immune challenge.

3.9. Conclusions

Aided by the RNA-Seq data, genome assembly and domain search, we identified and organized 122 putative PRRs in *M. sexta*. Phylogenetic analysis of LRRPs and their close homologs in other species revealed orthologous relationships and lineage-specific

family expansions, but roles for most of the LRRPs in innate immunity are not yet established. Examination of other protein families (*i.e.* PGRP, β GRP, galectin, Nimrod, multi-Ig domain protein, TEP, FREP) and comparison with homologs from other insects provided clues for future functional studies of these molecules. Based on the RNA-Seq data and upstream sequences, we explored their expression profiles, immune relatedness, and transcription regulation. It is likely that some of these gene products have synergistic or antagonistic effects on the initiation of host defense responses in hemocoel, midgut, and other tissues of the insect. Together, this collection of information constitutes a solid foundation for future studies of *M. sexta* PRRs.

Acknowledgments

This work was supported by NIH grants GM58634 (to H. Jiang) and GM041247 (to M. Kanost), and DARPA/NSF (IOS-1354421) grants (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.

Appendix A. Supplementary data

Supplementary data related to this article can be found at <http://dx.doi.org/10.1016/j.ibmb.2015.02.001>.

References

- Arai, I., Ohta, M., Suzuki, A., Tanaka, S., Yoshizawa, Y., Sato, R., 2013. Immunohistochemical analysis of the role of hemocytin in nodule formation in the larvae of the silkworm, *Bombyx mori*. *J. Insect Sci.* 13, 125.
- Bao, Y.Y., Xue, J., Wu, W.J., Wang, Y., Lv, Z.Y., Zhang, C.X., 2011. An immune-induced reeler protein is involved in the *Bombyx mori* melanization cascade. *Insect Biochem. Mol. Biol.* 41, 696–706.
- Blandin, S., Levashina, E.A., 2004. Thioester-containing proteins and insect immunity. *Mol. Immunol.* 40, 903–908.
- Blandin, S., Shiao, S.H., Moita, L.F., Janse, C.J., Waters, A.P., Kafatos, F.C., Levashina, E.A., 2004. Complement-like protein TEPI is a determinant of vectorial capacity in the malaria vector *Anopheles gambiae*. *Cell* 116, 661–670.
- Cao, X., He, Y., Hu, Y., Wang, Y., Chen, Y., Bryant, B., Clem, R., Blissard, G., Jiang, H., 2015a. The immune signaling pathways of *Manduca sexta*. *Insect Biochem. Mol. Biol.* 62, 64–74.
- Cao, X., He, Y., Hu, Y., Zhang, X., Wang, Y., Zou, Z., Chen, Y., Blissard, G., 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 Biochem. Mol. Biol.* 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 Biochem. Mol. Biol.* 62, 2–10.
- Charroux, B., Rival, T., Narbonne-Reveau, K., Royet, J., 2009. Bacterial detection by *Drosophila* peptidoglycan recognition proteins. *Microbes Infect.* 11, 631–636.
- Christophides, G.K., Zdobnov, E., Barillas-Mury, C., Birney, E., Blandin, S., Blass, C., Brey, P.T., Collins, F.H., Danielli, A., Dimopoulos, G., Hetru, C., Hoa, N.T., Hoffmann, J.A., Kanzok, S.M., Letunic, I., Levashina, E.A., Loukeris, T.G., Lycett, G., Meister, S., Michel, K., Moita, L.F., Muller, H.M., Osta, M.A., Paskewitz, S.M., Reichhart, J.M., Rzhetsky, A., Troxler, L., Vernick, K.D., Vlachou, D., Volz, J., von Mering, C., Xu, J., Zheng, L., Bork, P., Kafatos, F.C., 2002. Immunity-related genes and gene families in *Anopheles gambiae*. *Science* 298, 159–165.
- d'Alençon, E., Bierre, N., Girard, P.A., Magdelenat, G., Gimenez, S., Seninet, I., Escoubas, J.M., 2013. Evolutionary history of X-tox genes in three lepidopteran species: origin, evolution of primary and secondary structure and alternative splicing, generating a repertoire of immune-related proteins. *Insect Biochem. Mol. Biol.* 43, 54–64.
- Dong, Y., Dimopoulos, G., 2009. *Anopheles* fibrinogen-related proteins provide expanded pattern recognition capacity against bacteria and malaria parasites. *J. Biol. Chem.* 284, 9835–9844.
- Dong, Y., Taylor, H.E., Dimopoulos, G., 2006. AgDscam, a hypervariable immunoglobulin domain-containing receptor of the *Anopheles gambiae* innate immune system. *PLoS Biol.* 4, e229.
- Eleftherianos, I., Gokcen, F., Felföldi, G., Millichap, P.J., Trenczek, T.E., Ffrench-Constant, R.H., Reynolds, S.E., 2007. The immunoglobulin family protein Hemolin mediates cellular immune responses to bacteria in the insect *Manduca sexta*. *Cell. Microbiol.* 9, 1137–1147.
- Estévez-Lao, T.Y., Hillyer, J.F., 2014. Involvement of the *Anopheles gambiae* Nimrod gene family in mosquito immune responses. *Insect Biochem. Mol. Biol.* 44, 12–22.
- 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 defense mechanisms in honey bees *Apis mellifera*. *Insect Mol. Biol.* 15, 645–656.
- 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.
- 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.* 62, 23–37.
- Jiang, H., Ma, C., Lu, Z.Q., Kanost, M.R., 2004. β -1,3-glucan recognition protein-2 (β GRP-2) from *Manduca sexta*: an acute-phase protein that binds β -1,3-glucan and lipoteichoic acid to aggregate fungi and bacteria and stimulate phenoloxidase activation. *Insect Biochem. Mol. Biol.* 34, 89–100.
- Jiang, H., Vilcinskis, A., Kanost, M.R., 2010. Immunity in lepidopteran insects. In: Söderhäll, K. (Ed.), *Invertebrate Immunity*, Adv. Exp. Med. Biol., vol. 708, pp. 181–204.
- Ju, J.S., Cho, M.H., Brade, L., Kim, J.H., Park, J.W., Ha, N.C., Söderhäll, I., Söderhäll, K., Brade, H., Lee, B.L., 2006. A novel 40-kDa protein containing six repeats of an epidermal growth factor-like domain functions as a pattern recognition protein for lipopolysaccharide. *J. Immunol.* 177, 1838–1845.
- Kamhawi, S., Ramalho-Ortigao, M., Pham, V.M., Kumar, S., Lawyer, P.G., Turco, S.J., Barillas-Mury, C., Sacks, D.L., Valenzuela, J.G., 2004. A role for insect galectins in parasite survival. *Cell* 119, 329–341.
- Kobe, B., Kajava, A.V., 2001. The leucine-rich repeat as a protein recognition motif. *Curr. Opin. Struct. Biol.* 11, 725–732.
- Labropoulou, V., Douris, V., Stefanou, D., Magrioti, C., Swevers, L., Iatrou, K., 2008. Endoparasitoid wasp bracovirus-mediated inhibition of hemolymph function and lepidopteran host immunosuppression. *Cell. Microbiol.* 10, 2118–2128.
- Ladendorff, N.E., Kanost, M.R., 1990. Isolation and characterization of bacteria-induced protein P4 from hemolymph of *Manduca sexta*. *Arch. Insect Biochem. Physiol.* 15, 33–41.
- Ladendorff, N.E., Kanost, M.R., 1991. Bacteria-induced protein P4 (hemolin) from *Manduca sexta*: a member of the immunoglobulin superfamily which can inhibit hemocyte aggregation. *Arch. Insect Biochem. Physiol.* 18, 285–300.
- 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.
- Lemaitre, B., Hoffmann, J., 2007. The host defense of *Drosophila melanogaster*. *Ann. Rev. Immunol.* 25, 697–743.
- Leulier, F., Lemaitre, B., 2008. Toll-like receptors—taking an evolutionary approach. *Nat. Rev. Genet.* 9, 165–178.
- Levashina, E.A., Moita, L.F., Blandin, S., Vriend, G., Laguerre, M., Kafatos, F.C., 2001. Conserved role of a complement-like protein in phagocytosis revealed by dsRNA knockdown in cultured cells of the mosquito, *Anopheles gambiae*. *Cell* 104, 709–718.
- Li, B., Dewey, C.N., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinform.* 12, 323.
- Ligoxygakis, P., 2013. Genetics of immune recognition and response in *Drosophila* host defense. *Adv. Genet.* 83, 71–97.
- Ma, C., Kanost, M.R., 2000. A β -1,3-glucan recognition protein from an insect, *Manduca sexta*, agglutinates microorganisms and activates the phenoloxidase cascade. *J. Biol. Chem.* 275, 7505–7514.
- McGuinness, D.H., Dehal, P.K., Pleass, R.J., 2003. Pattern recognition molecules and innate immunity to parasites. *Trends Parasitol.* 19, 312–319.
- Mellroth, P., Karlsson, J., Steiner, H., 2003. A scavenger function for a *Drosophila* peptidoglycan recognition protein. *J. Biol. Chem.* 278, 7059–7064.
- Nardi, J.B., Pilas, B., Bee, C.M., Zhuang, S., Garsha, K., Kanost, M.R., 2006. Neuroglial-positive plasmacytes of *Manduca sexta* and the initiation of hemocyte attachment to foreign surfaces. *Dev. Comp. Immunol.* 30, 447–462.
- Nelson, R.E., Fessler, L.L., Takagi, Y., Blumberg, B., Keene, D.R., Olson, P.F., Parker, C.G., Fessler, J.H., 1994. Peroxidase: a novel enzyme-matrix protein of *Drosophila* development. *EMBO J.* 13, 3438–3447.
- Neves, G., Zucker, J., Daly, M., Chess, A., 2004. Stochastic yet biased expression of multiple Dscam splice variants by individual cells. *Nat. Genet.* 36, 240–246.
- Pace, K.E., Baum, L.G., 2004. Insect galectins: roles in immunity and development. *Glycoconj. J.* 19, 607–614.
- Pal, S., Wu, L.P., 2009. Lessons from the fly: pattern recognition in *Drosophila melanogaster*. In: Kishore, U. (Ed.), *Target Pattern Recognition in Innate Immunity*, Adv. Exp. Med. Biol., vol. 653, pp. 162–174.
- Padmanabhan, M., Cournoyer, P., Dinesh-Kumar, S.P., 2009. The leucine-rich repeat domain in plant innate immunity: a wealth of possibilities. *Cell. Microbiol.* 11, 191–198.
- Petersen, T.N., Brunak, S., von Heijne, G., Nielsen, H., 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* 8, 785–786.
- Povelones, M., Waterhouse, R.M., Kafatos, F.C., Christophides, G.K., 2009. Leucine-rich repeat protein complex activates mosquito complement in defense against *Plasmodium* parasites. *Science* 324, 258–261.
- 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.* 62, 75–85.
- Rao, X.J., Zhong, X., Lin, X.Y., Huang, X.H., Yu, X.Q., 2014. Characterization of a novel *Manduca sexta* β -1,3-glucan recognition protein (β GRP3) with multiple functions. *Insect Biochem. Mol. Biol.* 52, 13–22.
- Schmidt, O., Theopold, U., Strand, M., 2001. Innate immunity and its evasion and suppression by hymenopteran endoparasitoids. *Bioessays* 23, 344–351.
- Schmucker, D., 2007. Molecular diversity of Dscam: recognition of molecular identity in neuronal wiring. *Nat. Rev. Neurosci.* 8, 915–920.
- Somogyi, K., Sipos, B., Péntzes, Z., Kurucz, E., Zsámboki, J., Hultmark, D., Andó, I., 2008. Evolution of genes and repeats in the Nimrod superfamily. *Mol. Biol. Evol.* 25, 2337–2347.
- Strand, M.R., 2008. The insect cellular immune response. *Insect Sci.* 15, 1–15.
- Sumathipala, N., Jiang, H., 2010. Involvement of *Manduca sexta* peptidoglycan recognition protein-1 in the recognition of bacteria and activation of prophenoloxidase system. *Insect Biochem. Mol. Biol.* 40, 485–495.
- Sun, S.C., Lindstöm, I., Boman, H.G., Faye, I., Schmidt, O., 1990. Hemolin: an insect immune protein belonging to the immunoglobulin superfamily. *Science* 250, 1729–1732.
- Takahashi, D., Dai, H., Hiromasa, Y., Krishnamoorthi, R., Kanost, M.R., 2014. Self-association of an insect β -1,3-glucan recognition protein upon binding laminarin stimulates prophenoloxidase activation as an innate immune response. *J. Biol. Chem.* 289, 28399–28410.
- Tamura, K., Stecher, G., Peterson, D., Filipski, A., Kumar, S., 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* 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.
- Tung, T.T., Nagaosa, K., Fujita, Y., Kita, A., Mori, H., Okada, R., Nonaka, S., Nakanishi, Y., 2013. Phosphatidylserine recognition and induction of apoptotic cell clearance by *Drosophila* engulfment receptor Draper. *J. Biochem.* 153, 483–491.
- Voigt, A., Pflanz, R., Schäfer, U., Jäckle, H., 2002. Perlecan participates in proliferation activation of quiescent *Drosophila* neuroblasts. *Dev. Dyn.* 224, 403–412.
- Wang, X., Zhao, Q., Christensen, B.M., 2005. Identification and characterization of the fibrinogen-like domain of fibrinogen-related proteins in the mosquito, *Anopheles gambiae*, and the fruitfly, *Drosophila melanogaster*, genomes. *BMC Genomics* 6, 114.
- Wang, Y., Sumathipala, N., Rayaprolu, S., Jiang, H., 2011. Recognition of microbial molecular patterns and stimulation of prophenoloxidase activation by a β -1,3-glucanase-related protein in *Manduca sexta* larval plasma. *Insect Biochem. Mol. Biol.* 41, 322–331.
- Watson, F.L., Püttmann-Holgado, R., Thomas, F., Lamar, D.L., Hughes, M., Kondo, M., Rebel, V.I., Schmucker, D., 2005. Extensive diversity of Ig-superfamily proteins in the immune system of insects. *Science* 309, 1874–1878.
- Werner, T., Liu, G., Kang, D., Ekengren, S., Steiner, H., Hultmark, D., 2000. A family of peptidoglycan recognition proteins in the fruit fly *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. U. S. A.* 97, 13772–13777.
- Wilkin, M.B., Becker, M.N., Mulvey, D., Phan, I., Chao, A., Cooper, K., Chung, H.J., Campbell, I.D., Baron, M., MacIntyre, R., 2000. *Drosophila* dumpy is a gigantic extracellular protein required to maintain tension at epidermal-cuticle attachment sites. *Curr. Biol.* 10, 559–567.
- X et al., 2015.
- Yassine, H., Osta, M.A., 2010. *Anopheles gambiae* innate immunity. *Cell. Microbiol.* 12, 1–9.
- Yoshida, H., Kinoshita, K., Ashida, M., 1996. Purification of a peptidoglycan recognition protein from hemolymph of the silkworm, *Bombyx mori*. *J. Biol. Chem.* 271, 13854–13860.
- Yu, X.Q., Zhu, Y.F., Ma, C., Fabrick, J.A., Kanost, M.R., 2002. Pattern recognition proteins in *Manduca sexta* plasma. *Insect Biochem. Mol. Biol.* 32, 1287–1293.
- Zhang, S.G., Gunaratna, R.T., Zhang, X.F., Najjar, 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, S.G., Cao, X.L., 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 Biochem. Mol. Biol.* 47, 46–51.
- Zhu, Y., Johnson, T.J., Myers, A.A., Kanost, M.R., 2003. Identification by subtractive suppression hybridization of bacteria-induced genes expressed in *Manduca sexta* fat body. *Insect Biochem. Mol. Biol.* 33, 541–559.
- Zhu, Y., Ragan, E.J., Kanost, M.R., 2010. Leureptin: a soluble, extracellular leucine-rich repeat protein from *Manduca sexta* that binds lipopolysaccharide. *Insect Biochem. Mol. Biol.* 40, 713–722.
- Zou, Z., Evans, J.D., Lu, Z., Zhao, P., Williams, M., Sumathipala, N., Hetru, C., Hultmark, D., Jiang, H., 2007. Comparative genomic analysis of the *Tribolium* immune system. *Genome Biol.* 8, R177.