Insect Biochemistry and Molecular Biology 62 (2015) 38–50

Contents lists available at ScienceDirect

Insect Biochemistry and Molecular Biology

journal homepage: www.elsevier.com/locate/ibmb

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 κB motifs in the promoter regions. This collection of information is expected to facilitate future biochemical studies detailing their respective roles in this model insect.

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).

Abbreviations: PG, peptidoglycan; LPS, lipopolysaccharide; LTA, lipoteichoic acid; PRR, pattern recognition receptor; PGRP, peptidoglycan recognition protein; TIR, Toll/interleukin-1 receptor; βGRP, 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.

http://dx.doi.org/10.1016/j.ibmb.2015.02.001
0965-1748/

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 (βGRP1–3), microbe binding protein (MBP), immulectin-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, βGRPs, 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/nuccore/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/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/iprscan/). 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 $\log_2$(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 (KKGNNCTTTY), and CATTW boxes were documented. NF-κB motifs (GGGRAYYYYY) with 0, 1 or 2 mismatches were also identified.