HHS Public Access

Author manuscript

Insect Biochem Mol Biol. Author manuscript; available in PMC 2016 July 01.

Published in final edited form as:

Insect Biochem Mol Biol. 2015 July ; 62: 100–113. doi:10.1016/j.ibmb.2014.12.010.

Annotation and expression analysis of cuticular proteins from the tobacco hornworm, *Manduca sexta*

Neal T. Dittmer[a,*], **Guillaume Tetreau**[b], **Xiaolong Cao**[c], **Haobo Jiang**[d], **Ping Wang**[b], and **Michael R. Kanost**[a]

[a]Department of Biochemistry and Molecular Biophysics, 141 Chalmers Hall, Kansas State University, Manhattan, Kansas 66506, USA

[b]Department of Entomology, Cornell University, New York State Agricultural Experiment Station, Geneva, New York 14456, USA

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

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

Abstract

The insect cuticle is a unique material that covers the exterior of the animal as well as lining the foregut, hindgut, and tracheae. It offers protection from predators and desiccation, defines body shape, and serves as an attachment site for internal organs and muscle. It has demonstrated remarkable variations in hardness, flexibility and elasticity, all the while being light weight, which allows for ease of movement and flight. It is composed primarily of chitin, proteins, catecholamines, and lipids. Proteomic analyses of cuticle from different life stages and species of insects has allowed for a more detailed examination of the protein content and how it relates to cuticle mechanical properties. It is now recognized that several groups of cuticular proteins exist and that they can be classified according to conserved amino acid sequence motifs. We have annotated the genome of the tobacco hornworm, *Manduca sexta*, for genes that encode putative cuticular proteins that belong to seven different groups: proteins with a Rebers and Riddiford motif (CPR), proteins analogous to peritrophins (CPAP), proteins with a tweedle motif (CPT), proteins with a 44 amino acid motif (CPF), proteins that are CPF-like (CPFL), proteins with an 18 amino acid motif (18 aa), and proteins with two to three copies of a C-X_5-C motif (CPCFC). In total we annotated 248 genes, of which 207 belong to the CPR family, the most for any insect genome annotated to date. Additionally, we discovered new members of the CPAP family and determined that orthologous genes are present in other insects. We established orthology between the *M. sexta* and *Bombyx mori* genes and identified duplication events that occurred after separation of the two species. Finally, we utilized 52 RNAseq libraries to ascertain gene

*Corresponding author. Telephone: (785) 532-4033, ndittmer@k-state.edu.

expression profiles that revealed commonalities and differences between different tissues and developmental stages.

Keywords

cuticular proteins; Rebers & Riddiford; RNAseq; orthology; *Manduca sexta*

1. Introduction

The insect exoskeleton (cuticle) is a remarkable extracellular structure secreted by epidermal cells that serves as the outer body covering. It helps to protect the insect from environmental stresses such as predators, parasites, abrasions, desiccation, and UV radiation. It also functions as an attachment site for internal muscles and organs, and is instrumental to locomotion and flight. The bulk of the cuticle is made primarily of a network of chitin embedded in a proteinaceous matrix, with water, catecholamines, and some lipids (Moussian, 2013). Despite what may seem like a limited pallet of materials, insects are able to synthesize cuticles that differ widely with respect to physical properties. Measurements of hardness have differed by more than 30 fold, while the stiffness of cuticle has been shown to vary by more than seven orders of magnitude (Klocke and Schmitz (2011) and references within; Vincent and Wegst, 2004). Investigations of the reasons behind these differences have focused on the role of dehydration and chemical cross-linking (Andersen, 2010; Sugumaran, 2010; Vincent 2009). However, it has been apparent for many years that differences exist between proteins extracted from different types of cuticle (hard versus soft), different developmental stages (i.e. larva, pupa, adult), as well as different time points (pre-molt versus post-molt) (Andersen *et al*., 1986, 1987, 1995a; Cox and Willis, 1985; Dittmer *et al*., 2012; Jensen *et al*., 1997; Kiely and Riddiford, 1985; Missios *et al*., 2000). Obtaining sequence information of these proteins was limited as it required either solubilization, purification, digestion, and sequencing of peptide fragments, or screening of cDNA expression libraries with antibodies raised against cuticle extracts in order to identify the proteins involved. Now, the emergence of large scale genomic, proteomic, and transcriptomic analyses has allowed for a renewed look at the importance of the protein content.

The first insect genome sequenced was that of *Drosophila melanogaster* (Adams *et al*., 2000); there are now 112 insect genome sequences available at NCBI. An analysis of 12 genomes by Ioannidou *et al*. (2014) suggested that genes encoding structural cuticular proteins (CP) represent on average 1% of the total protein-coding genes in insects. The total number of genes varied from 63 for *Pediculus humanus* to 301 for *Aedes aegypti*. These numbers, as well as those given for other insects, likely represent a minimum as most genomes have not gone through a rigorous manual annotation and rely on homology to other known CPs. Combining proteomics with genomics has been used to identify cuticular proteins for several insect species (Bae *et al*., 2011; Carrasco *et al*., 2011; Dittmer *et al*., 2012; Fu *et al*., 2011; He *et al*., 2007). Similarly, genomics and transcriptomics have contributed valuable information on CP gene expression, discerning variations in the timing (pre- or post-molt), developmental stage (larval, pupal, adult), and relative expression levels

of these genes (Cornman and Willis, 2009; Dittmer *et al*., 2012; Futahashi *et al*., 2008; Gallot *et al*., 2010; Liang *et al*., 2010; Okamoto *et al*., 2008; Suetsugu *et al*., 2013; Togawa *et al*., 2007, 2008).

Many CPs can be classified by the conserved sequence motifs they contain (Ioannidou *et al*., 2014; Willis *et al*., 2012). The largest group is known as CPs with the Rebers and Riddiford motif (CPR) that contain a core 28 amino acid sequence that is now recognized as part of a larger 63 amino acid consensus sequence (pfam00379) (Rebers and Riddiford, 1988; Willis *et al*., 2012). Variations in the extended consensus have been recognized, and CPs are often classified into one of three sub types: RR-1, RR-2, and RR-3. The number of CPR genes can vary greatly among species but many insects have more than 100 (Ioannidou *et al*., 2014). Additional groups having conserved sequence motifs include CPs analogous to peritrophins (CPAP) (Jasrapuria *et al*., 2010), CPs with a 44 amino acid motif (CPF) (Andersen *et al*., 1997; Togawa *et al*., 2007), CPF-like (CPFL) (Togawa et al., 2007), CPs with a Tweedle motif (CPT) (Guan *et al*., 2006), CPs with two or three repeats of C-X_5-C motif (CPCFC) (Jensen *et al*., 1997; Willis *et al*., 2012), and CPs with an 18 amino acid motif (Andersen, 2000; Nakato *et al*., 1990). Additionally, low-complexity proteins can be found in the cuticle that are rich in glycine or contain repeats of AAP(A/V), P(V/Y), GYGL, or GLLG (Willis *et al*., 2012). However, since these proteins lack any other distinctive sequences, presence of these repeats alone is not proof enough of their location in the cuticle. Excellent reviews on CPs can be found in Willis (2010) and Willis *et al*, (2012).

The goal of this research was to annotate the CP genes in the genome of the tobacco hornworm, *Manduca sexta*, with respect to the groups described above. These seven groups (CPR, CPAP, CPF, CPFL, CPT, CPCFC, and 18 aa motif) were chosen as they all contain conserved sequences previously shown to be present in known cuticular structural proteins and, therefore, can serve as a diagnostic feature. We compared the CP genes in *M. sexta* with those of *Bombyx mori* (Futahashi *et al*., 2008) in order to identify orthologs. Finally, we utilized 52 RNAseq libraries prepared from various tissues and developmental stages to look for patterns of coordinated expression among the CP genes. This analysis offers new insights into the CPs present in cuticle synthesized at different times and developmental stages.

2. Materials and Methods

2.1 CP gene annotation

To identify putative CP genes, the *M. sexta* genome and first official gene set (OGS1), available from the Agricultural Pest Genomics Resource Database (www.agripestbase.org), were searched by tBLASTn (Altschul *et al*., 1997) with the following sequences: for the CPR family, the consensus sequence GxFxYxxPDGxxxxVxYxADENGYQPxGAHLP was used to identify the RR-1 subtype, and EYDAxPxYxFxYxVxDxHTGDxKSQxExRDGDVVxGxYSLxExDGxxRTVxYTADxxNG FNAVVxxE was used to identify the RR-2 subtype (Figure 3 in Willis *et al*., 2005); classification into the appropriate subfamily was confirmed by the use of a profile hidden Markov model that discriminates between the two subtypes, available at the cuticleDB website (biophysics.biol.uoa.gr/cuticleDB/, Karouzou et al., 2007). For the CPAP family,

the *Tribolium castaneum* CPAP1-D (GenBank ACY95469) and CPAP3-A1 (GenBank ACY95475) sequences were used to identify CPAP family members. CPF genes were identified using the most highly conserved portion of the 44 amino acid motif, VSxYSKAVDTPFSSVRKxDxRIVNxA, (derived from Figure 1B in Togawa *et al*., 2007). CPFL genes were identified with the sequence LxYSAAPAVSHVAYxGxGxxYGW (derived from Figure 3 in Togawa *et al*., 2007). For Tweedle genes, the 100 amino acid weblogo sequence from Figure 3A in Willis (2010) was used to identify homologs. The sequence YPAGVNPAACPNYPYCD was used to identify members of the CPCFC family, and PVDTPEVAAAKAAHFAAH was used to identify genes encoding CPs with the 18 amino acid motif (Figure 4C and 4B in Willis, 2010).

For all genes except those of the CPAP family, names were assigned based on putative orthology to *B. mori* homologs. Reciprocal BLAST was performed to confirm that the *B. mori* protein identified as the top hit to a *M. sexta* query identified the same *M. sexta* protein as the top hit when it was used as the query sequence. Orthology was further established through the use of microsynteny. When the genes flanking a CP gene were homologous between *B. mori* and *M. sexta*, the CP genes were considered to be orthologous; identification of genes surrounding *B. mori* CP genes was inferred from the Gene Report page at NCBI associated with that particular CP gene. Because the CPAP genes in *B. mori* have not been annotated yet, naming of the *M. sexta* CPAP genes was based on a phylogenetic analysis of homologous proteins from several insect species; a detailed description of this analysis can be found in Tetreau *et al*. (2014, this issue).

2.2 Phylogenetic analysis

Phylogenetic analysis was performed using the corresponding protein sequence of selected genes from the CPR, CPAP, and CPFL groups. Details of the CPAP analysis can be found in a companion paper (Tetreau *et al*., 2014, this issue). Analysis was performed using MEGA software (v5.2.1; Tamura *et al*., 2011). Sequences were aligned globally using the ClustalW program in MEGA and then adjusted manually by eye. For the CPR group, the extended RR domain (pfam00379) was used; RR-1 and RR-2 subgroups were treated separately. For the CPFL group, the entire sequence of the *M. sexta* and *B. mori* proteins were used. Trees were constructed by the neighbor-joining method with a Poisson correction model. Gaps were treated by the pairwise deletion method and statistical analysis was performed by the bootstrap method using 1000 repetitions.

2.3 CP gene expression

Fifty-two RNAseq libraries had been prepared from various tissues and developmental stages as part of the Manduca Genome Sequencing Project (GenBank assembly accession number GCA_000262585.1). The libraries were sequenced by Illumina technology either at the Weill Cornel Medical College or the Baylor College of Medicine Human Genome Sequencing Center. The RNAseq reads were trimmed to 50 base pair and mapped onto the predicted open reading frame of the CP genes. Transcript abundance was determined by the FPKM method (fragments per kilobase of transcript per million fragments mapped). Read mapping and determination of FPKM values were performed using the RSEM software package (v1.2.12) (Li and Dewey, 2011). (The expression data for all genes in the *M. sexta*

OGS2 can be downloaded at ftp://ftp.bioinformatics.ksu.edu/pub/Manduca/OGS2/OSU_files/, file 20140508RSEM_OGS2_Gene_FPKM.xlsx.) The FPKM values ranged from 0 to greater than 10,000 and gene expression was grouped as follows: less than 1, no expression; 1 – 10, very low expression; 10 – 100, low expression; 100 – 1,000, moderate expression; 1,000 – 10,000, high expression; greater than 10,000, very high expression. Hierarchical clustering of CP gene expression was performed using MultiExperiment Viewer (v4.9) (Saeed *et al*., 2003) with the Pearson correlation-based metric and average linkage clustering method.

3. Results and Discussion

3.1.1 CP gene annotation—Our analysis of OGS1 identified 206 gene models that belong to seven different CP groups (CPR, CPAP, CPF, CPFL, CPT, CPCFC, 18 aa motif). However, upon annotation of these models it became apparent that 28 of them likely represented 44 genes that had been incorrectly spliced. An additional 26 genes, for which no OGS1 gene model was predicted, were identified by searching the *M. sexta* genome sequence. These findings were incorporated during the development of OGS2 and increased the total number of gene models to 276 (Table 1). Some of these models are likely allelic variants of the same gene. For example, *Msex2.14790* encodes a partial *CPR* gene that is 98% identical at the nucleotide level, and 100% identical at the amino acid level, to *Msex2.11885* (*MsCPR38*). Furthermore, *Msex2.14790* is on a small scaffold of ∼2000 nucleotides which is 94% identical at the nucleotide level to position 14,274 - 16,027 of scaffold00529, where *Msex2.11885* resides. Thus it was judged to be an allelic variant of *Msex2.11885*.

Additionally, some CP genes were also split between multiple gene models due to gaps in the genome assembly (supplementary file 1). This was evident for *MsCPAP3-B*. Both *Msex2.08808* and *15015* show high sequence identity at the amino acid level (80% and 96% respectively) with obstructor-B (i.e. CPAP3-B) from *Papilio xuthus* (GenBank BAM17933). Interestingly, Msex2.08808's shared identity with PxObst-B covers amino acids 1-110 and 245-291, while Msex2.15015 aligns with residues 113-242. An examination of the genomic sequence for *Msex2.08808* reveals a gap between exons 3 and 4 which likely contains the missing exons encoding residues 111-244 present in Msex2.15015. Thus, the correct sequence for *MsCPAP3-B* is a composite of *Msex2.08808* and *Msex2.15015*. This conclusion is supported by the RNAseq data assembled by the Trinity software package which predicts a 1.24 kb transcript encoding a 296 amino acid protein that is the same as the composite sequence and 87% identical to PxObst-B (data not shown). (*Msex2.14185* encodes for a protein that is 99% identical at both the amino acid and nucleotide levels with Msex2.15015 and therefore was judged to be an allele.) After accounting for multiple gene models representing the same gene, we identified 248 distinct *CP* genes in *M. sexta*.

3.1.2 CPR genes—Two-hundred and seven of the annotated genes belong to the CPR family: 79 RR-1, 124 RR-2, and 4 RR-3 (Table 1). This represents the largest number of CPR genes in any insect genome annotated to date and is 36% greater than the number of CPR genes in the silkworm *B. mori*: 56 RR-1, 93 RR-2, and 4 RR-3; 149 were originally annotated by Futahashi et al. (2008) including the misclassified CPH5, three more were

added by Liang et al. (2010), and one additional gene, LOC101743422, was discovered in our analysis. Similar to observations for other insects (Cornman *et al*., 2008; Dittmer *et al*., 2012; Gallot *et al*., 2010; Liang *et al*., 2010), many of the genes occurred in clusters: 55 of the RR-1 genes were found in clusters of 11-17 genes on four scaffolds (00081, 00136, 00224, 00529), while 86 of the RR-2 genes were found in clusters of 8-26 genes on six scaffolds (00006, 00227, 00685, 00717, 01004, 01013) (Table 1).

As automated annotation is being carried out on several sequenced lepidopteran genomes, *B. mori* genes are frequently the top hit resulting in some continuity in gene naming among the lepidoptera. We endeavored to continue this practice by naming the *M. sexta CP* genes based on their deduced orthology to *B. mori* as determined by sequence identity, phylogenetic analysis, and microsynteny. For some of the *M. sexta* genes no clear orthology with *B. mori* could be established. For others, gene duplication resulted in multiple *M. sexta* genes being orthologous to a single *B. mori* gene. In these cases, new CPR numbers were assigned that begin with *CPR152*. Therefore, the CPR gene numbering went as high as *CPR250* even though there are only 207 genes. Nevertheless, we believe that those numbered *CPR1 - 151* are orthologous to the *B. mori* gene with the same name.

We were able to assign orthology to 52 of the 56 *B. mori* RR-1 genes; only *BmCPR29*, *31*, *50*, and *53* did not have clear orthologs in *M. sexta*. (*BmCPR50* and *53* may be orthologs of *Msex2.04420* and *06326* but this could not be confirmed by synteny, therefore, the *M. sexta* genes were given the names *CPR159* and *162*). Sixteen of the additional RR-1 genes in *M. sexta* likely arose through duplication of just four genes: *MsCPR2*, *13*, *41*, and *46* (boxed in grey in Table 1). For example, in *B. mori* the five genes *CPR41-46* are flanked by *CPG21* on one end and *ORC3* and *T-related protein-like* on the other (Figure 1). These same three genes in *M. sexta* flank a cluster of 17 *CPR* genes on scaffold00136. BLAST results and phylogenetic analysis indicate that *Msex2.06320* and *06321* are co-orthologs of *BmCPR41*, *Msex2.06322 – 06325* are orthologous to *BmCPR42-45*, and *Msex2.06327 – 06336* are all most closely related to *BmCPR46*. As no ortholog could be identified by sequence similarity or phylogenetic analysis, *Msex2.06336* was given the name *MsCPR46* because of its location and orientation with respect to *ORC3* and *T-related protein-like*. Similar analysis indicates that *Msex2.00115* and *00116* are paralogs of *Msex2.00117* (*MsCPR2*), and *Msex2.08403 – 08406* are paralogs of *Msex2.08402* (*MsCPR13*).

Determining orthology among the RR-2 genes was more difficult as only 51 of the 93 RR-2 genes in *B. mori* had clear one- to-one orthologs in *M. sexta*. Interestingly, orthology could be established to most of the *B. mori* genes from *CPR57 – 84* and *121 – 145*; it was only for *CPR85 – 120* (with the exception of *90* and *91*) that direct orthology could not be conclusively resolved, although groups of relatedness were apparent. Phylogenetic analysis showed that 11 genes clustered on scaffold00685 (*Msex2.12553 – 12561*, *15538*, *15539*, which we have named *MsCPR183 – 193*) were most similar to *BmCPR85 – 89* (Figure S1). The common node for these two groups suggest that they arose from a common ancestral gene and that duplication after speciation gave rise to additional genes. This conclusion is further supported by the location of *MsCPR183 – 193* between genes *MsCPR84* (*Msex2.12552*) and *MsCPR90* (*Msex2.12562*) on scaffold00685, thus making it likely that

they would be related to *BmCPR85 – 89*. Similar homologous groupings are shown in Table 2 and Figure S2.

Four RR-3 genes were identified, *Msex2.06360*, *11251*, *07106*, and *04414*, that all have orthologs in *B.mori* (*CPR146 – 149*). *BmCPR149* was originally annotated as encoding an RR-2 protein (GenBank ACY06906), however, we have classified the M.sexta gene as an RR-3 based on greater sequence identity with the RR-3 consensus sequence (Andersen, 2000; data not shown). MsCPR149 is unique in that it encodes a protein with two RR domains, one each at the N- and C-termini of the protein; this model is supported by the RNAseq data suggesting that it is not an error in the gene assembly resulting from the fusion of two adjacent genes. It is similar to CPR58 of Nansonia vitripennis (GenBank XP_001601627) which has a comparable domain architecture (Willis, 2010).

3.1.3 CPAP genes—The second largest group of *CP* genes annotated was the *CPAPs*. *CPAPs* are classified according to the number of peritrophin A-type chitin binding domains they contain, with CPAP1 proteins containing one domain and CPAP3 proteins containing three domains (Jasrapuria *et al*., 2010). The *CPAP3* genes were first recognized in *D. melanogaster* and alternately named *gasp* (gene analogous to small peritrophins) and *obstructor* (Barry et al., 1999; Behr and Hoch, 2005). This family of genes was further divided into two groups consisting of *obstructor-A – E*, and *F – J*. Genes orthologous to *obst-A – E* have been identified in other insects while *obst-F – J* have only been described for *D. melanogaster*. The related *CPAP1* genes were identified and characterized by Jasrapuria *et al*. (2010) and contained ten family members named *CPAP1-A – J*.

We identified 15 putative *CPAP1* genes and 10 *CPAP3* genes (Table 1). This is greater than the 10 *CPAP1* and 7 *CPAP3* genes described from *T. castaneum* for which the most extensive annotation of this gene family has been carried out (Jasrapuria *et al*., 2010). Since similar work had not yet been performed for *B. mori*, we named the *M. sexta* genes based on orthology to *T. castaneum* and homologous genes in other insects. This analysis is described in a companion paper detailing the different groups of proteins containing peritrophin A-type chitin-binding domains in *M. sexta* (Tetreau *et al*., 2014, this issue) but will be summarized here. *M. sexta* has orthologs to 9 of the 10 *CPAP1* genes in *T. castaneum* with *1-E* being the exception. The six additional *CPAP1* genes in *M. sexta* all have orthologs in other insect species and we have named these *CPAP1-B2* and *1-K – O*. *M. sexta* has orthologs to all seven of the *CPAP3* genes in *T. castaneum*, and phylogenetic analysis shows that the three additional genes cluster with *CPAP3-E* and, therefore, we have named them *E2 – E4*; it should be noted that *MsCPAP3-E2* (*Msex2.03294*) is not orthologous to *Obst-E2* of *D. melanogaster* as the latter is a spliced isoform of the *Obst-E* gene while the former is a separate gene distinct from *MsCPAP3-E1* (*Msex2.03293*). Similar to what Jasrapuria *et al*. (2010) reported for *T. castaneum*, the *CPAP3-C* ortholog of *M. sexta* (*Msex2.08810*) has the potential for alternative splicing of exon 5, giving rise to two different proteins. Eight of the ten *CPAP3* genes are present in tandem arrays on two scaffolds: *A1*, *A2*, *D1*, *B*, and *C* on scaffold00255 (with one gene between *B* and *C*), and *E1 – E3* on scaffold00068. The CPAP1 genes don't exhibit the same clustering characteristic although *B1* and *B2* are

adjacent on scaffold00007, while *H*, *F*, and *N* are grouped on scaffold00032 (with one gene between *F* and *N*) (Table 1).

In general, very little annotation of these genes has been carried out for insect genomes. Recently, Ioannidou and co-workers (2014) developed profile hidden Markov models to identify CPAP and other cuticle structural proteins from sequenced genomes and applied this analysis to 14 arthropod genomes (12 insect and 2 crustacean). The number of putative *CPAP* genes identified ranged from 10 to 20 for *CPAP1* and from 5 to 12 for *CPAP3* (supplementary file 2 in Ioannidou *et al*., 2014). Combining our analysis of *M. sexta* with their analysis of several insect genomes allowed us to assign orthology to the *CPAP* genes from eight insect species (Table 3). The *CPAP3* genes had the highest degree of conservation, with *Acyrthosiphon pisum*, *Apis mellifera*, *B. mori*, *M. sexta*, and *T. castaneum* all having orthologs to the seven described genes. *Anopheles gambiae* and *P. humanus* had orthologs to six of these genes with both lacking an ortholog to *CPAP3-A2*, although *A. gambiae* appears to have two duplications of *CPAP3-A1* and *P. humanus* a duplication of *CPAP3-D1* (Table 3). Interestingly, we observed that the gene model *PHUM434540* appears to be a fusion of two genes with the 5-prime half orthologous to *CPAP3-B* and the 3-prime half orthologs to *CPAP3-A1*. As detailed by Behr and Hoch (2005), *D. melanogaster* has five *CPAP3* (a.k.a. *obstructor*) genes, lacking orthologs to *CPAP3-A2* and *CPAP3-D2*.

More interspecies variability was observed with the *CPAP1* genes. Of the original ten genes described by Jasrapuria *et al*. (2010), the number of genes present in the eight species analyzed varied from five (*A. pisum*) to ten (*T. castaneum*), with most having eight or nine (Table 3). All eight species had orthologs to the newly described *CPAP1-B2* branch, and at least three orthologs to *CPAP1-K – O* groups. A few genes from *A. pisum*, *A. mellifera*, and *P. humanus* could not be clearly grouped within any of the 16 CPAP1 branches and were left unclassified. It should be noted that although both our analysis (this report and Tetreau *et al*., 2014, this issue) and that of Ioannidou *et al*. (2014) independently verified the close relationship of the new members of the CPAP1 group to the original ten genes described previously for *T. castaneum*, the presence of the encoded proteins in the cuticle must still be established (and hence are they truly cuticle proteins). The *D. melanogaster* ortholog to *CPAP1-K* (*CG7549*), alternately known as *mind-the-gap (mtg)*, is expressed in neuronal cells and encodes a protein critical for the organization of the synaptic extracelluar matrix of neuromuscular junctions (Rohrbough *et al*., 2007; Rushton *et al*., 2012). Interestingly, *mtg* null mutants, which die during the embryonic stage, exhibit a weakened cuticle and herniated hindgut phenotype (Rohrbough *et al*., 2007). This phenotype can be rescued by ubiquitous expression of a wildtype transgene (Rushton *et al*., 2009). According to FlyBase, moderate levels of *mtg* expression was also detected in the larval hindgut, tracheae, and carcass, all cuticle expressing tissues. Thus, it seems that *mtg*, and possibly other CPAPs, may have multiple roles.

3.1.4 Other CP genes—All of the remaining CP genes we annotated had identifiable orthologs in *B. mori* (Table 1). *M. sexta* and *B. mori* each have four *CPT* genes, one *CPCFC* gene, and one *CPF* gene. *B. mori* has five genes with the 18 amino acid motif (*BmCPH14*, *15*, *16*, *30*, and *31*) whereas *M. sexta* has only four. An examination of the protein sequences

reveals that BmCPH14 and 15 are 92% identical and likely resulted from a gene duplication event that did not occur in *M. sexta*. The *M. sexta* ortholog (Msex2.01072) had a slightly higher sequence identity to BmCPH15 and, therefore, was given that name. Whereas six *CPFL* genes were identified in *M. sexta*, only four were in *B. mori*. The corresponding proteins of genes *Msex2.04035*, *04036*, and *04037* were all similar to BmCPFL4. Msex2.04036 had the highest sequence similarity and was designated the ortholog; the remaining two genes were named *MsCPFL5* and *MsCPFL6*.

3.2 CP gene expression

Fifty-two RNAseq libraries were created to provide transcript information to aid in the annotation of the genome. We sought to take advantage of this information to examine the expression profiles of the CP genes across multiple tissues and developmental stages. One caveat is that a majority of the libraries were created from tissues collected during the post-molt period. Thus, expression data of genes expressed primarily or exclusively during pre-molt (pharate) stages is limited. A second complication concerns the potential (if not likely) contamination of non-cuticle synthesizing samples with cuticle expressing tissues such as trachea or epidermis during the collection process. For example, we would not expect CP genes to be expressed in fat body, midgut, Malpighian tubules, or muscle. However, these samples may also contain pieces of epidermis or trachea and, therefore, CP gene expression was observed in some of these libraries. Finally, most libraries were sequenced only once, thus caution is advised when interpreting expression levels. (Four libraries were sequenced by both paired-end reads and single-end reads: midgut 5th instar wandering, and abdominal muscle (proleg) 5th instar at 12 h, pre-wandering, and wandering stages.)

The average number of CP genes expressed across all libraries was 88, with a minimum of 6, a maximum of 216, and a median of 72 (Figure 2A). The libraries with the highest number of CP genes expressed were the whole head and abdominal muscle (proleg) libraries of 4th instar larva after head capsule slippage (HCS, i.e. pharate 5th instar), each with 216 genes, while the library with the fewest CP genes expressed was prepared from adult midguts 3 -5 days post eclosion (6 genes). In total, 21 libraries had “above average” (> 88 genes) expression. This includes the 11 whole head libraries, the 5 libraries prepared form eggs or whole larvae (1st through 3rd instar), 3 of the abdominal muscle (proleg) libraries (4th instar HCS and 5th instar 12 h), midgut from 3rd instar HCS, and fat body from 4th instar HCS and pupa days 15-18; the latter three libraries would not be predicted to have abundant CP gene expression but this may be an indication of contamination from trachea or epidermis.

The libraries having the highest levels of CP transcripts (as judged by total FPKM values) are predictably from tissues enriched for epidermal cells; namely whole heads, whole larvae, or muscle scraped from the abdominal body wall either just prior to or just after molting (Figure 2B). CPR RR-1 transcripts dominated most libraries, having the highest total FPKM values in all but nine libraries (Table 4). Genes from RR-1, RR-2, CPH, and CPFL all had high levels of expression in the whole head libraries (supplementary file 1) with RR-2 having the highest transcript levels in the 4th instar HCS library and CPH the highest in adult day 1. On average, the top ten expressed genes in each library accounted for nearly 77% of

the CP transcripts, and we view them as representing the major CPs expressed in that library. We compared the top ten expressed genes in various libraries to identify differences and commonalities between them. Transcripts for RR-1 genes *CPR154* and *156*, the RR-2 gene *CPR63*, and genes *CPH1* and *CPH15* (CPCFC and 18 aa groups, respectively) are among the top ten in most post-molt larval head libraries but not adult. Conversely, *CPR19*, *23*, *37*, *CPH30*, and *CPH31* are more highly expressed in the adult head than in the larva (Table 5). In contrast, *CPFL1* and *Msex2.02570* were commonly found in both the larval and adult whole head libraries. (*Msex2.02570* encodes a low complexity protein that is not a member of the CP groups discussed here but it is annotated in GenBank as a pupal cuticle protein; accession number AY585211.) Only two of the whole head libraries were made from tissue collected pre-molt. For the 4th instar HCS (pharate 5th instar) library, six of the ten most highly expressed genes were RR-2, as compared to only two in the post-molt larval whole head libraries. A different pattern was observed when comparing the pre-molt adult head library (pupa days 21-22) with post-molt adult head libraries in which many of the top expressed genes were the same (*CPR19*, *23*, *37*, *67*, *CPH30*, *31*, and *CPFL1*) (Table 5).

Differences between pre-molt and post-molt CP expression were also observed with the 3 day old eggs (pharate 1st instar) and the whole larva libraries (1st through 3rd instars 1 day post-molt). The whole larva libraries shared seven genes in common among the top ten (RR-1 genes *CPR2*, *3*, *4*, *5*, *154*, *156*, and *170*) of which only two (*CPR3* and *4*) were also highly expressed in 3 day old eggs. Top CP expressing genes in 3 day old eggs include *CPT2*, *CPT3*, *CPAP3-A1* and *CPAP3-Ca* (supplementary file 1). Differences between pre-molt and post molt CP expression was not seen with the three libraries from abdominal muscle (proleg) having high CP transcript levels (4th instar HCS and 5th instar 12 h both single and paired end reads); CPR RR-1 transcripts dominated the list with six of the top ten genes the same between the three libraries (*CPR3*, *4*, *46*, *154*, *156*, and *163*; supplementary file 1). Common top expressed genes among the larval libraries (whole head, whole larvae, abdominal muscle) include the RR-1 genes *CPR3*, *4*, *5*, *154*, and *156* (supplementary file 1 and Fig S3). In a similar fashion, the corresponding genes in *B. mori* (*BmCPR2-5*) were among the top expressing genes during the pre- and post-molt 5th instar larval stage (Okamoto et al., 2008). Recently, Qiao and coworkers (2014) identified *BmCPR2* as the gene responsible for the *stony* mutant in *B. mori*. This mutation lead to significant defects in cuticle extension, tensile strength, larval mobility, and body shape. It was also observed for *B. mori* that *CPR46* was highly expressed post-molt in larval epidermis. As noted in section 3.1.2, the *M. sexta* ortholog to *CPR46* has undergone gene expansion that encompasses 10 genes, *CPR46* and *CPR163 – 171* (Table 1). Members of this co-orthologous group appear as top expressed genes in many of the post-molt larval libraries in *M. sexta* as well (supplementary file 1).

Twenty of the annotated genes showed no expression in any of the libraries; 19 of these belong to the CPR RR-2 group and one belonged to the CPR RR-1 group (see supplementary file 1). Furthermore, 10 of these 20 genes belonged to the same cluster of related genes (group 4 in Table 2). An additional 29 genes were not detected in at least 47 of the 52 libraries (90%), with 22 of these belonging to the CPR RR-2 group; 8 of these had established orthology with *B. mori* genes (with CPR numbers between 121 – 145) while 7

belonged to a cluster whose orthology could not be established. The lone CPF gene in *M. sexta*, *Msex2.05601*, was not detected in 48 of the libraries, but had high expression in the whole head library from pupa day 21 – 22 (pharate adult) and then decreased thereafter (low expression in adult day 1 whole head and very low expression in adult days 3 and 7 whole head libraries). Similarly, transcripts for *CPAP3-E4* (*Msex2.14886*) were detected at moderate levels in 3 h old eggs but no other libraries. These results suggest that genes showing very low or no expression in most (or all) libraries may have a restricted tissue or temporal expression, or that they may be pseudogenes.

Clustering analysis was performed to identify genes that had similar expression patterns across all 52 libraries. Figure 3 shows selected profiles from this analysis. (A heat map depicting the grouping and relative expression levels of all the CP genes is shown in supplementary Figure S3.) One of the most striking observation was of eight genes that had low to moderate levels of expression in nearly all of the libraries, producing an "expression band" that extended the width of the heat map (Figure 3A and supplementary Figure S3). Five of the eight genes in this cluster belonged to the CPAP group (*CPAP1-C*, *1-H*, *1-M*, and *CPAP3-D2* and *3-Cb*) and two were RR-3 (*CPR146* and *149*). The nearly ubiquitous expression pattern of these genes suggests that they may be important for general cuticle synthesis or synthesis of tracheal cuticle. Another six genes were nearly exclusively expressed only in pre-molt libraries during larval development (the exception being expression in fat body from pupa days 15 – 18) (Figure 3B). Remarkably, three of the genes belonged to the Tweedle family (of which only four were found in *M. sexta*) and the other three are CPR RR-2 genes that are clustered together in the genome. Similar observations were reported for *B. mori* in that *CPT2-4* shared the same expression pattern with the highest expression occurring at the larval stage during the molt (Liang et al., 2010).

Several genes showed expression in just a few libraries such as whole heads 4th instar HCS and 3 day old eggs (Figure 3C). What these two libraries have in common is that they were prepared from tissue synthesizing larval head capsule cuticle pre-molt. The fact that those genes were not expressed in other larval pre-molt libraries that may contain tracheal or epidermis contamination such as fat body 4th instar HCS, midgut 3rd and 4th instars HCS, or abdominal muscle (proleg) 4th instar HCS, suggests that those genes are not involved in synthesis of tracheal or body wall cuticle. Similarly, several genes were expressed almost exclusively in 4th instar whole heads and abdominal muscle (proleg) HCS libraries, indicating likely function in larval head capsule and body wall synthesis pre-molt (Figure 3D). Additional interesting patterns observed include restriction of expression to whole heads of 5th instar day 2 and abdominal muscle (proleg) 5th instar post-molt (Figure 3E), 3 day old eggs and midguts of 2nd and 3rd instar larvae (Figure 3F), and eggs 3 hours old (Figure 3G). Three of the four genes in the last example are the additional members of the CPAP3-E group found in *B. mori* and *M. sexta* but not the other species presented in Table 3. This time point is too early for embryonic or 1st instar cuticle synthesis (Konopova and Zrzavy, 2005; Ziese and Dorn, 2003) but the transcripts may be maternally loaded.

4. Conclusions

Combining reciprocal BLAST, phylogenetic analysis, and microsynteny proved an effective method for identifying orthologous CP genes between *M. sexta* and *B. mori*. The most challenging identification was for the CPR RR-2 group which appears to be undergoing duplication after speciation more often than for other CP genes. For the remaining groups examined, one to one orthology was more easily established. The 207 CPR genes identified are currently the most in any insect for which this type of annotation has been carried out (although recent analysis by Ioannidou et al. (2014) indicates that the yellow fever mosquito, *Aedes aegypti*, may have as many as 240), and is 37% more than the number of CPR genes identified in *B. mori*; it remains to be determined which number is more typical for the Lepidoptera. Our analysis has also expanded the number of genes in the CPAP1 group from 10 to 16 but the veracity of the new genes as true CPs must be confirmed.

The RNAseq libraries provided valuable data on gene expression. In most libraries, CPR RR-1 genes accounted for most of the CP gene transcripts but this may have been influenced by the number of libraries prepared from larval tissues (32 of 52) or post-molt tissues (40 of 52). RR-1 genes have been proposed to be more prevalent in soft cuticle than hard cuticle (Andersen, 2000). As larva are mostly soft bodied and cuticle synthesized post-molt is thought to undergo less (or no) sclerotization than cuticle synthesized pre-molt (Andersen *et al*., 1995b), this may account for the abundance of RR-1 transcripts detected. Additionally, trachea is a likely contamination of the libraries prepared from midguts, Malpighian tubules, and fat body. The flexibility needed may necessitate a less sclerotized tracheal cuticle and, therefore, it is reasonable to anticipate a greater number of RR-1 proteins. In *B. mori*, Okamoto and co-workers (2008) also found that the majority of CP transcripts from two larval EST libraries (a pharate 5^{th} instar and 5^{th} instar day 3) were from RR-1 genes. Nevertheless, several distinct patterns of CP gene expression were observed, with differences noted between larval and adult cuticle or pre-molt and post-molt cuticle. As the purpose of the RNAseq libraries was to aid in the genome annotation, it was desirable to sample multiple tissues at several time points and all developmental stages. As a consequence, careful selection of a single tissue (i.e. epidermis, trachea, or hard cuticle or soft cuticle) or more refined time points was not a necessity. A more meticulous collection on both accounts in future experiments may yield a better characterization of CP gene expression.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by grants from the National Institutes of Health (GM41247) and the National Science Foundation (IOS1257961) to M. Kanost, from the NIH (GM58634) to H. Jiang, and by the Cornell University Agricultural Experiment Station federal formula funds received from the USDA Cooperative State Research, Education, and Extension Service to P. Wang. This is contribution number 15-142-J from the Kansas Agricultural Experiment Station.

References

Adams DA, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al. The genome sequence of *Drosophila melanogaster*. Science. 2000; 287:2185–2195. [PubMed: 10731132]

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–3402. [PubMed: 9254694]

Andersen SO. Studies on proteins in post-ecdysial nymphal cuticle of locust, *Locusta migratoria*, and cockroach, *Blaberus craniifer*. Insect Biochem Mol Biol. 2000; 30:569–577. [PubMed: 10844249]

Andersen SO. Insect cuticular sclerotization: a review. Insect Biochem Mol Biol. 2010; 40:166–178. [PubMed: 19932179]

Andersen SO, Højrup P. Extractable proteins from abdominal cuticle of sexually mature locusts, *Locusta migratoria*. Insect Biochem. 1987; 17:45–51.

Andersen SO, Højrup P, Roepstorff P. Characterization of cuticular proteins from the migratory locust, *Locusta migratoria*. Insect Biochem. 1986; 16:441–447.

Andersen SO, Rafin K, Krogh TN, Højrup P, Roepstorff P. Comparison of larval and pupal cuticular proteins in *Tenebrio molitor*. Insect Biochem Molec Biol. 1995a; 2:177–187. [PubMed: 7711749]

Andersen SO, Højrup P, Roepstorff P. Insect cuticular proteins. Insect Biochem Mol Biol. 1995b; 25:153–176. [PubMed: 7711748]

Andersen SO, Rafn K, Roepstorff P. Sequence studies of proteins from larval and pupal cuticle of the yellow meal worm, *Tenebrio molitor*. Insect Biochem Mol Biol. 1997; 27:121–131. [PubMed: 9066122]

Bae N, Lödl M, Pollak A, Lubec G. Mass spectrometrical analysis of cuticular proteins from the wing of *Hebemoia glaucippe* (Linnaeus, 1758) (Lepidoptera: Pieridae). J Proteomics. 2011; 75:517–531. [PubMed: 21903182]

Barry MK, Triplett AA, Christensen AC. A peritrophin-like protein expressed in the embryonic tracheae of *Drosophila melanogaster*. Insect Biochem Mol Biol. 1999; 29:319–327. [PubMed: 10333571]

Behr M, Hoch M. Identification of the novel evolutionary conserved obstructor multigene family in invertebrates. FEBS Letters. 2005; 579:6827–6833. [PubMed: 16325182]

Carrasco MA, Buechler SA, Arnold RJ, Sformo T, Barnes BM, Duman JG. Elucidating the biochemical overwintering adaptations of larval *Cucujus clavipes puniceus*, a nonmodel organism, via high throughput proteomics. J Proteome Res. 2011; 10:4634–4646. [PubMed: 21923194]

Cornman RS, Willis JH. Annotation and analysis of low-complexity protein families of *Anopheles gambiae* that are associated with cuticle. Insect Mol Biol. 2009; 18:607–622. [PubMed: 19754739]

Cornman RS, Togawa T, Dunn WA, He N, Emmons AC, Willis JH. Annotation and analysis of a large cuticular protein family with the R&R consensus in *Anopheles gambiae*. BMC Genomics. 2008; 9:22. [PubMed: 18205929]

Cox DL, Willis JH. The cuticular proteins of *Hyalophora cecropia* from different anatomical regions and metamorphic stages. Insect Biochem. 1985; 15:349–362.

Dittmer NT, Hiromasa Y, Tomich JM, Lu N, Beeman RW, Kramer KJ, Kanost MR. Proteomic and transcriptomic analyses of rigid and membranous cuticles and epidermis from the elytra and hindwings of the red flour beetle, *Tribolium castaneum*. J Proteome Res. 2012; 11:269–78. [PubMed: 22087475]

Fu Q, Li P, Xu YM, Zhang S, Jia L, Zha XF, Xiang ZH, He NJ. Proteomic analysis of larval integument, trachea and adult scale from the silkworm, *Bombyx mori*. Proteomics. 2011; 11:3761–3767. [PubMed: 21761556]

Futahashi R, Okamoto S, Kawasaki H, Zhong YS, Iwanaga M, Mita K, Fujiwara H. Genome-wide identification of cuticular genes in the silkworm, *Bombyx mori*. Insect Biochem Mol Biol. 2008; 38:1138–1146. [PubMed: 19280704]

Gallot A, Rispe C, Leterme N, Gauthier JP, Jaubert-Possamai S, Tagu D. Cuticular proteins and seasonal photoperiodism in aphids. Insect Biochem Mol Biol. 2010; 40:235–240. [PubMed: 20018241]

Guan X, Middlebrooks BW, Alexander S, Wasserman SA. Mutation of TweedleD, a member of an unconventional cuticle protein family, alters body shape in *Drosophila*. Proc Natl Acad Sci USA. 2006; 103:16794–16799. [PubMed: 17075064]

He N, Botelho JMC, McNall RJ, Belozerov V, Dunn WA, Mize T, Orlando R, Willis JH. Proteomic analysis of cast cuticles from *Anopheles gambiae* by tandem mass spectrometry. Insect Biochem Mol Biol. 2007; 37:135–146. [PubMed: 17244542]

Ioannidou ZS, Theodoropoulou MC, Papandreou NC, Willis JH, Hamodrakas SJ. CutProtFam-Pred: detection and classification of putative structural cuticular proteins from sequence alone, based on profile hidden Markov models. Insect Biochem Mol Biol. 2014; 52:51–59. [PubMed: 24978609]

Jasrapuria S, Arakane Y, Osman G, Kramer KJ, Beeman RW, Muthukrishnan S. Genes encoding proteins with peritrophin A-type chitin-binding domains in *Tribolium castaneum* are grouped into three distinct families based on phylogeny, expression and function. Insect Biochem Mol Biol. 2010; 40:214–227. [PubMed: 20144715]

Jensen UG, Rothmann A, Skou L, Andersen SO, Roepstorff P, Højrup P. Cuticular proteins from the giant cockroach, *Blaberus craniifer*. Insect Biochem Mol Biol. 1997; 27:109–120. [PubMed: 9066121]

Karouzou MV, Spyropoulos Y, Iconomidou VA, Cornman RS, Hamodrakas SJ, Willis JH. *Drosophila* cuticular proteins with the R&R consensus: annotation and classification with a new tool for discriminating RR-1 and RR-2 sequences. Insect Biochem Mol Biol. 2007; 37:754–760. [PubMed: 17628275]

Kiely ML, Riddiford LM. Temporal programming of epidermal cell protein synthesis during the larval-pupal transformation of *Manduca sexta*. Roux's Arch Dev Biol. 1985; 194:325–335.

Klocke D, Schmitz H. Water as a major modulator of the mechanical properties of insect cuticle. Acta Biomater. 2011; 7:2935–2942. [PubMed: 21515418]

Konopová B, Zrzavý J. Ultrastructure, development, and homology of insect embryonic cuticle. J Morphol. 2005; 264:339–362. [PubMed: 15838850]

Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011; 12:323. [PubMed: 21816040]

Liang J, Zhang L, Xiang Z, He N. Expression profile of cuticular genes of silkworm, *Bombyx mori*. BMC Genomics. 2010; 11:173. [PubMed: 20226095]

Missios S, Davidson HC, Linder D, Mortimer L, Okobi AO, Doctor JS. Characterization of cuticular proteins in the red flour beetle, *Tribolium castaneum*. Insect Biochem Mol Biol. 2000; 30:47–65. [PubMed: 10646970]

Moussian, B. The arthropod cuticle. In: Minelli, A.; Boxshall, G.; Fusco, G., editors. Arthropod Biology and Evolution. Springer-Verlag; Berlin Heidelberg: 2013. p. 171-196.

Nakato H, Toriyama M, Izumi S, Tomino S. Structure and expression of a mRNA for a pupal cuticle protein of the silkworm, *Bombyx mori*. Insect Biochem. 1990; 20:667–678.

Okamoto S, Futahashi R, Kojima T, Mita K, Fujiwara H. Catalogue of epidermal genes: genes expressed in the epidermis during larval molt of the silkworm *Bombyx mori*. BMC Genomics. 2008; 9:396. [PubMed: 18721459]

Qiao L, Xiong G, Wang RX, He SZ, Chen J, Tong XL, Hu H, Li CL, Gai TT, Xin YQ, et al. Mutation of a cuticular protein, *BmorCPR2*, alters larval body shape and adaptability in silkworm, *Bombyx mori*. Genetics. 2014; 196:1103–1115. [PubMed: 24514903]

Rebers JF, Riddiford LM. Structure and expression of a *Manduca sexta* larval cuticle gene homologous to *Drosophila* cuticle genes. J Mol Biol. 1988; 203:411–423. [PubMed: 2462055]

Rohrbough J, Rushton E, Woodruff E III, Fergestad T, Vigneswaran K, Broadie K. Presynaptic establishment of the synaptic cleft extracellular matrix is required for post-synaptic differentiation. Genes Dev. 2007; 21:2607–2628. [PubMed: 17901219]

Rushton E, Rohrbough J, Broadie K. Presynaptic secretion of mind-the-gap organizes the synaptic extracellular matrix-integrin interface and postsynaptic environments. Dev Dyn. 2009; 238:554–71. [PubMed: 19235718]

Rushton E, Rohrbough J, Deutsch K, Broadie K. Structure-function analysis of endogenous lectin mind-the-gap in synaptogenesis. Dev Neurobiol. 2012; 72:1161–1179. [PubMed: 22234957]

Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, et al. TM4: a free, open-source system for microarray data management and analysis. Biotechniques. 2003; 34:374–378. [PubMed: 12613259]

Suetsugu Y, Futahashi R, Kanamori H, Kadono-Okuda K, Sasanuma SI, Narukawa J, Ajimura M, Jouraku A, Namiki N, Shimomura M, et al. Large scale full-length cDNA sequencing reveals a unique genomic landscape in a lepidopteran model insect, Bombyx mori. G3. 2013; 3:1481–1492. [PubMed: 23821615]

Sugumaran M. Chemistry of cuticular sclerotization. Adv Insect Physiol. 2010; 39:151–209.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution. 2011; 28:2731–2739. [PubMed: 21546353]

Tetreau G, Dittmer N, Jasrapuria S, Cao X, Chen YR, Muthukrishnan S, Jiang H, Blissard GW, Kanost MR, Wang P. Analysis of chitin-binding proteins from *Manduca sexta* provides new insights into evolution of peritrophin A-type chitin-binding domains in insects. Insect Biochem Mol Biol. 2014 this issue. Submitted.

Togawa T, Dunn WA, Emmons AC, Willis JH. CPF and CPFL, two related gene families encoding cuticular proteins of *Anopheles gambiae* and other insects. Insect Biochem Mol Biol. 2007; 37:675–688. [PubMed: 17550824]

Togawa T, Dunn WA, Emmons AC, Nagao J, Willis JH. Developmental expression patterns of cuticular protein genes with the R&R Consensus from *Anopheles gambiae*. Insect Biochem Mol Biol. 2008; 38:508–519. [PubMed: 18405829]

Vincent JFV. If it's tanned it must be dry: a critique. J Adhes. 2009; 85:755–769.

Vincent JFV, Wegst UGK. Design and mechanical properties of insect cuticle. Arthropod Struct Dev. 2004; 33:187–199. [PubMed: 18089034]

Willis JH. Structural cuticular proteins from arthropods: annotation, nomenclature, and sequence characteristics in the genomics era. Insect Biochem Mol Biol. 2010; 40:189–204. [PubMed: 20171281]

Willis, JH.; Iconomidou, VA.; Smith, RF.; Hamodrakas, SJ. Cuticular Proteins. In: Gilbert, L.; Iatrou, K.; Gill, SS., editors. Comprehensive Molecular Insect Science. Vol. 4. Elsevier Pergamon; Oxford: 2005. p. 79-109.

Willis, JH.; Papandreou, NC.; Iconomidou, VA.; Hamodrakas, SJ. Cuticular Proteins. In: Gilbert, L., editor. Insect Molecular Biology and Biochemistry. Elsevier B.V.; 2012. p. 134-166.

Ziese S, Dorn A. Embryonic integument and “molts” in *Manduca sexta* (Insecta, Lepidoptera). J Morphol. 2003; 255:146–161. [PubMed: 12474263]

- 248 cuticular protein (CP) genes were annotated in *Manduca sexta*
- Orthology was established between *M. sexta* and *Bombyx mori* CP genes
- Gene expression was analyzed via 52 RNAseq libraries
- Diverse expression was observed across various tissues and developmental stages

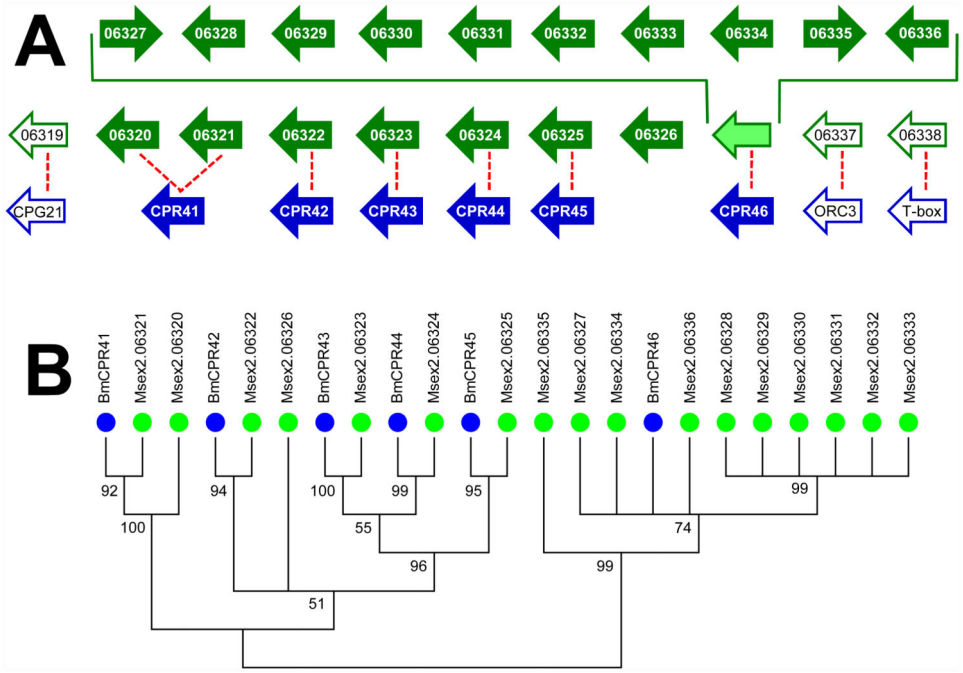

Figure 1. Evidence of gene duplication in *M. sexta*

(A) Alignment of *M. sexta* genes (green arrows) on scaffold00136 with the corresponding region in *B. mori* chromosome 22 (blue arrows). Filled arrows represent CPR genes while open arrows represent non-CPR genes; the arrows are only meant to show gene order and orientation and are not representative of gene size or distances between genes. The dashed red line denotes putative orthologs. (B) Phylogenetic analysis of the CPR genes shown in A. All branches with a bootstrap value less than 50 were collapsed.

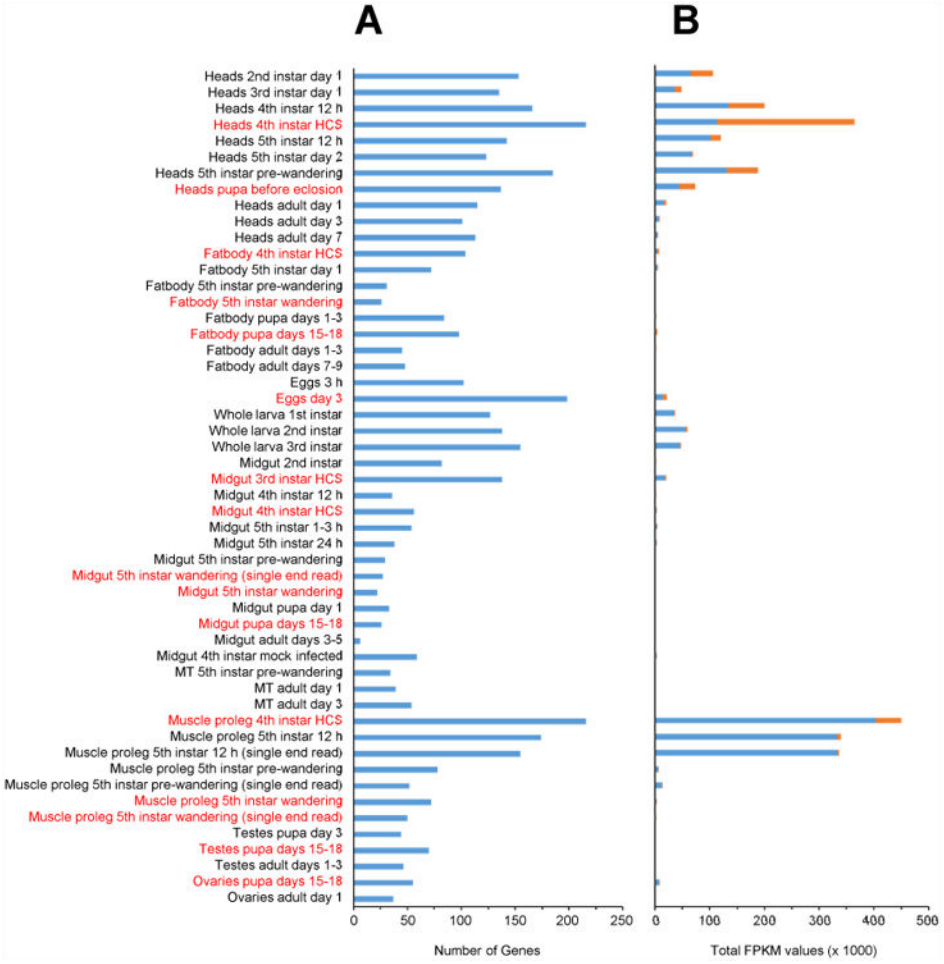

Figure 2. CP gene expression in the RNAseq libraries

(A) Total number of CP genes expressed in each library. (B) Total FPKM values for CPR RR-1 and RR-2 genes in each library; the pattern looks the same when the total FPKM value for all CP genes is displayed. Red type indicates libraries prepared from pre-molt (pharate) stages.

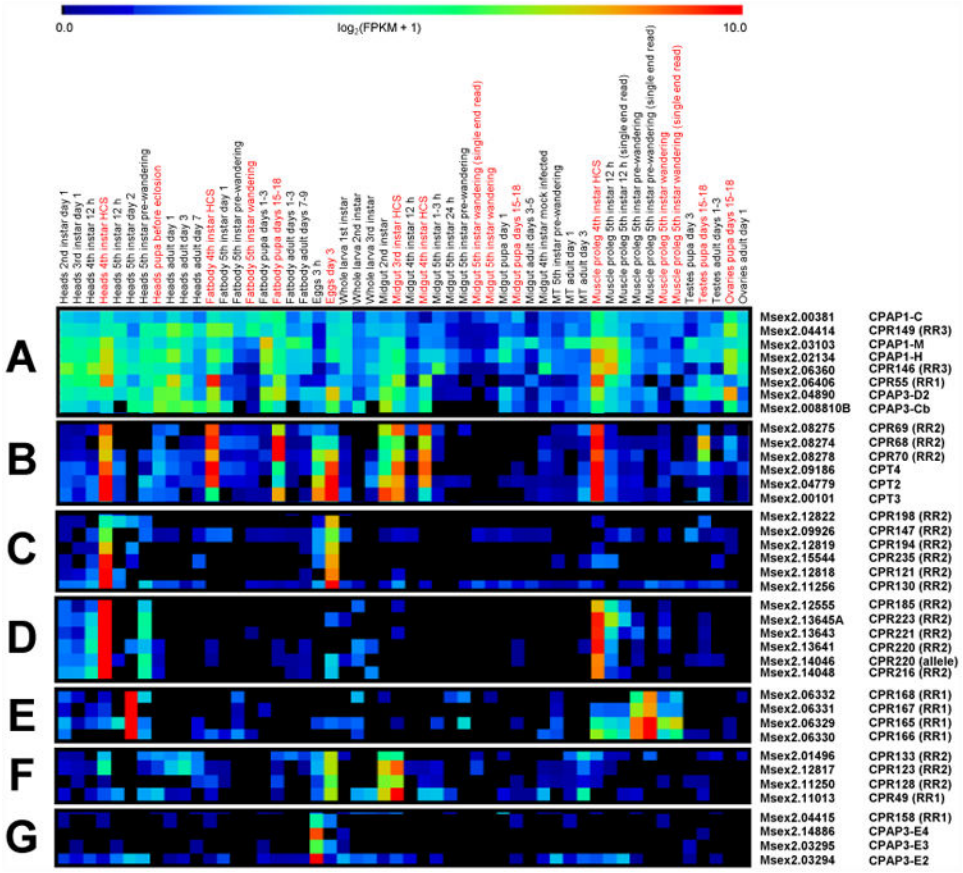

Figure 3. Selected expression patterns from cluster analysis

The expression pattern of each CP gene across all 52 RNAseq libraries were grouped by cluster analysis as described in section 2.3. A – G highlight some of the unique expression patterns observed; a full heat map of all of the genes is shown in supplemental Figure S3. Red type indicates libraries prepared from pre-molt (pharate) stages.

Table 1

***Manduca* structural cuticle protein genes**

OGS2 Gene ID	Gene Name [a]	Family	Scaffold No. [b]
Msex2.07923	MsCPR1	CPR-RR1	scaffold00208
Msex2.00114	MsCPR154 (LCP14)	CPR-RR1	scaffold00003
Msex2.00115	MsCPR155	CPR-RR1	scaffold00003
Msex2.00116	MsCPR156	CPR-RR1	scaffold00003
Msex2.00117	MsCPR2	CPR-RR1	scaffold00003
Msex2.00118	MsCPR3	CPR-RR1	scaffold00003
Msex2.09274	MsCPR4 (CP14.6)	CPR-RR1	scaffold00267
Msex2.09273	MsCPR5	CPR-RR1	scaffold00267
Msex2.11639	MsCPR6	CPR-RR1	scaffold00482
Msex2.05318	MsCPR7	CPR-RR1	scaffold00107
Msex2.05319	MsCPR8	CPR-RR1	scaffold00107
Msex2.05320	MsCPR9 (CP20)	CPR-RR1	scaffold00107
Msex2.05189	MsCPR10	CPR-RR1	scaffold00106
Msex2.08400	MsCPR11	CPR-RR1	scaffold00224
Msex2.08401	MsCPR12	CPR-RR1	scaffold00224
Msex2.08402	MsCPR13	CPR-RR1	scaffold00224
Msex2.08403	MsCPR174	CPR-RR1	scaffold00224
Msex2.08404	MsCPR175	CPR-RR1	scaffold00224
Msex2.08405	MsCPR176	CPR-RR1	scaffold00224
Msex2.08406	MsCPR177	CPR-RR1	scaffold00224
Msex2.08407	MsCPR14	CPR-RR1	scaffold00224
Msex2.08408	MsCPR15 (CP36)	CPR-RR1	scaffold00224
Msex2.08409	MsCPR16 (CP27)	CPR-RR1	scaffold00224
Msex2.08410	MsCPR17	CPR-RR1	scaffold00224
Msex2.08411	MsCPR18	CPR-RR1	scaffold00224
Msex2.08412	MsCPR19	CPR-RR1	scaffold00224
Msex2.08413	MsCPR52	CPR-RR1	scaffold00224
Msex2.00741	MsCPR20	CPR-RR1	scaffold00009
Msex2.00742	MsCPR21	CPR-RR1	scaffold00009
Msex2.04086	MsCPR157	CPR-RR1	scaffold00073
Msex2.04415	MsCPR158	CPR-RR1	scaffold00081
Msex2.04416	MsCPR22	CPR-RR1	scaffold00081
Msex2.04417	MsCPR23	CPR-RR1	scaffold00081
Msex2.04418	MsCPR24	CPR-RR1	scaffold00081
Msex2.04419	MsCPR25	CPR-RR1	scaffold00081
Msex2.04420	MsCPR159	CPR-RR1	scaffold00081

OGS2 Gene ID	Gene Name [a]	Family	Scaffold No. [b]
Msex2.04424	MsCPR26	CPR-RR1	scaffold00081
Msex2.04425	MsCPR30	CPR-RR1	scaffold00081
Msex2.04427	MsCPR160	CPR-RR1	scaffold00081
Msex2.04428	MsCPR27	CPR-RR1	scaffold00081
Msex2.04429, Msex2.14415	MsCPR28	CPR-RR1	scaffold00081, scaffold01804
Msex2.11897	MsCPR182	CPR-RR1	scaffold00529
Msex2.11896	MsCPR181	CPR-RR1	scaffold00529
Msex2.11895	MsCPR180	CPR-RR1	scaffold00529
Msex2.11894	MsCPR179	CPR-RR1	scaffold00529
Msex2.11892	MsCPR178	CPR-RR1	scaffold00529
Msex2.11891	MsCPR32	CPR-RR1	scaffold00529
Msex2.11890	MsCPR33	CPR-RR1	scaffold00529
Msex2.11889	MsCPR34	CPR-RR1	scaffold00529
Msex2.11888	MsCPR35	CPR-RR1	scaffold00529
Msex2.11887	MsCPR36	CPR-RR1	scaffold00529
Msex2.11886	MsCPR37	CPR-RR1	scaffold00529
Msex2.11885, Msex2.14790	MsCPR38	CPR-RR1	scaffold00529, scaffold03863
Msex2.11884, Msex2.14482	MsCPR39	CPR-RR1	scaffold00529, scaffold02011
Msex2.14550, Msex2.11883	MsCPR40	CPR-RR1	scaffold02427, scaffold0529
Msex2.06320	MsCPR161	CPR-RR1	scaffold00136
Msex2.06321	MsCPR41	CPR-RR1	scaffold00136
Msex2.06322	MsCPR42	CPR-RR1	scaffold00136
Msex2.06323	MsCPR43	CPR-RR1	scaffold00136
Msex2.06324	MsCPR44	CPR-RR1	scaffold00136
Msex2.06325	MsCPR45	CPR-RR1	scaffold00136
Msex2.06326	MsCPR162	CPR-RR1	scaffold00136
Msex2.06327	MsCPR163	CPR-RR1	scaffold00136
Msex2.06328	MsCPR164	CPR-RR1	scaffold00136
Msex2.06329	MsCPR165	CPR-RR1	scaffold00136
Msex2.06330	MsCPR166	CPR-RR1	scaffold00136
Msex2.06331	MsCPR167 (LCP16/17)	CPR-RR1	scaffold00136
Msex2.06332	MsCPR168	CPR-RR1	scaffold00136
Msex2.06333	MsCPR169	CPR-RR1	scaffold00136
Msex2.06334	MsCPR170	CPR-RR1	scaffold00136
Msex2.06335	MsCPR171	CPR-RR1	scaffold00136
Msex2.06336	MsCPR46	CPR-RR1	scaffold00136
Msex2.13066	MsCPR47	CPR-RR1	scaffold00793
Msex2.08425	MsCPR48	CPR-RR1	scaffold00221
Msex2.11013	MsCPR49	CPR-RR1	scaffold00396

OGS2 Gene ID	Gene Name [a]	Family	Scaffold No. [b]
Msex2.13464	MsCPR51	CPR-RR1	scaffold01003
Msex2.06407	MsCPR54	CPR-RR1	scaffold00137
Msex2.06406	MsCPR55	CPR-RR1	scaffold00137
Msex2.04351	MsCPR56	CPR-RR1	scaffold00079
Msex2.00132	MsCPH5	CPR-RR2	scaffold00003
Msex2.02008	MsCPR57	CPR-RR2	scaffold00029
Msex2.00295	MsCPR58	CPR-RR2	scaffold00006
Msex2.00296	MsCPR59	CPR-RR2	scaffold00006
Msex2.00297	MsCPR61	CPR-RR2	scaffold00006
Msex2.00298	MsCPR62	CPR-RR2	scaffold00006
Msex2.00299	MsCPR63	CPR-RR2	scaffold00006
Msex2.00301	MsCPR64	CPR-RR2	scaffold00006
Msex2.00302	MsCPR65	CPR-RR2	scaffold00006
Msex2.00303	MsCPR134	CPR-RR2	scaffold00006
Msex2.09232	MsCPR66	CPR-RR2	scaffold00266
Msex2.08273	MsCPR67	CPR-RR2	scaffold00227
Msex2.08276	MsCPR172	CPR-RR2	scaffold00227
Msex2.08279	MsCPR173	CPR-RR2	scaffold00227
Msex2.08274	MsCPR68	CPR-RR2	scaffold00227
Msex2.08275	MsCPR69	CPR-RR2	scaffold00227
Msex2.08278	MsCPR70	CPR-RR2	scaffold00227
Msex2.08280	MsCPR71	CPR-RR2	scaffold00227
Msex2.08282	MsCPR72	CPR-RR2	scaffold00227
Msex2.08283	MsCPR73	CPR-RR2	scaffold00227
Msex2.05356	MsCPR74	CPR-RR2	scaffold00108
Msex2.04619	MsCPR76	CPR-RR2	scaffold00086
Msex2.04620	MsCPR77	CPR-RR2	scaffold00086
Msex2.03062	MsCPR79	CPR-RR2	scaffold00054
Msex2.03061	MsCPR80	CPR-RR2	scaffold00054
Msex2.15511	MsCPR153 [c]	CPR-RR2	scaffold00054
Msex2.03060	MsCPR152 [c]	CPR-RR2	scaffold00054
Msex2.12550	MsCPR82	CPR-RR2	scaffold00685
Msex2.12551	MsCPR83	CPR-RR2	scaffold00685
Msex2.12552	MsCPR84	CPR-RR2	scaffold00685
Msex2.12553	MsCPR183	CPR-RR2	scaffold00685
Msex2.12554	MsCPR184	CPR-RR2	scaffold00685
Msex2.12555, Msex2.14741	MsCPR185	CPR-RR2	scaffold00685, scaffold03458
Msex2.15538	MsCPR186	CPR-RR2	scaffold00685
Msex2.15539	MsCPR187	CPR-RR2	scaffold00685

OGS2 Gene ID	Gene Name [a]	Family	Scaffold No. [b]
Msex2.12556	MsCPR188	CPR-RR2	scaffold00685
Msex2.12557, Msex2.14630	MsCPR189	CPR-RR2	scaffold00685, scaffold02706
Msex2.12558	MsCPR190	CPR-RR2	scaffold00685
Msex2.12559	MsCPR191	CPR-RR2	scaffold00685
Msex2.12560	MsCPR192	CPR-RR2	scaffold00685
Msex2.12561	MsCPR193	CPR-RR2	scaffold00685
Msex2.12562	MsCPR90	CPR-RR2	scaffold00685
Msex2.12563	MsCPR91	CPR-RR2	scaffold00685
Msex2.12837	MsCPR214	CPR-RR2	scaffold00717
Msex2.15541	MsCPR213	CPR-RR2	scaffold00717
Msex2.12836	MsCPR212	CPR-RR2	scaffold00717
Msex2.12835	MsCPR211	CPR-RR2	scaffold00717
Msex2.12834	MsCPR210	CPR-RR2	scaffold00717
Msex2.12833	MsCPR209	CPR-RR2	scaffold00717
Msex2.12832, Msex2.14053	MsCPR208	CPR-RR2	scaffold00717, scaffold01268
Msex2.12831, Msex2.14054	MsCPR207	CPR-RR2	scaffold00717, scaffold01268
Msex2.12830, Msex2.14055	MsCPR206	CPR-RR2	scaffold00717, scaffold01268
Msex2.12829, Msex2.14056	MsCPR205	CPR-RR2	scaffold00717, scaffold01268
Msex2.12828, Msex2.14057	MsCPR204	CPR-RR2	scaffold00717, scaffold01268
Msex2.12827, Msex2.14058, Msex2.14210	MsCPR203	CPR-RR2	scaffold00717, scaffold01268, scaffold04887
Msex2.12826	MsCPR202	CPR-RR2	scaffold00717
Msex2.12825	MsCPR201	CPR-RR2	scaffold00717
Msex2.12824	MsCPR200	CPR-RR2	scaffold00717
Msex2.12823	MsCPR199	CPR-RR2	scaffold00717
Msex2.12822	MsCPR198	CPR-RR2	scaffold00717
Msex2.12821	MsCPR197	CPR-RR2	scaffold00717
Msex2.15540	MsCPR196	CPR-RR2	scaffold00717
Msex2.12820	MsCPR195	CPR-RR2	scaffold00717
Msex2.12819	MsCPR194	CPR-RR2	scaffold00717
Msex2.12818	MsCPR121	CPR-RR2	scaffold00717
Msex2.12816	MsCPR122	CPR-RR2	scaffold00717
Msex2.12817	MsCPR123	CPR-RR2	scaffold00717
Msex2.12815	MsCPR124	CPR-RR2	scaffold00717
Msex2.12814	MsCPR150	CPR-RR2	scaffold00717
Msex2.12194	MsCPR125	CPR-RR2	scaffold00562
Msex2.12193	MsCPR126	CPR-RR2	scaffold00562
Msex2.11250	MsCPR128	CPR-RR2	scaffold00476
Msex2.11252	MsCPR129	CPR-RR2	scaffold00476
Msex2.11256	MsCPR130	CPR-RR2	scaffold00476

OGS2 Gene ID	Gene Name [a]	Family	Scaffold No. [b]
Msex2.11351	MsCPR131	CPR-RR2	scaffold00447
Msex2.13191, Msex2.14515	MsCPR132	CPR-RR2	scaffold00893, scaffold02588
Msex2.01496	MsCPR133	CPR-RR2	scaffold00024
Msex2.08525	MsCPR135	CPR-RR2	scaffold00228
Msex2.08526	MsCPR136	CPR-RR2	scaffold00228
Msex2.08527	MsCPR137	CPR-RR2	scaffold00228
Msex2.08354	MsCPR140 (Pro-resilin)	CPR-RR2	scaffold00240
Msex2.04385	MsCPR141	CPR-RR2	scaffold00081
Msex2.09926	MsCPR142	CPR-RR2	scaffold00325
Msex2.11692	MsCPR143	CPR-RR2	scaffold00565
Msex2.12993	MsCPR144	CPR-RR2	scaffold00779
Msex2.12994	MsCPR145	CPR-RR2	scaffold00779
Msex2.09922	MsCPR151	CPR-RR2	scaffold00325
Msex2.13636, Msex2.14049	MsCPR215	CPR-RR2	scaffold01004, scaffold01304
Msex2.13637, Msex2.14048	MsCPR216	CPR-RR2	scaffold01004, scaffold01304
Msex2.13638	MsCPR217	CPR-RR2	scaffold01004
Msex2.13639, Msex2.15551	MsCPR218	CPR-RR2	scaffold01004, scaffold01304
Msex2.13640, Msex2.14047	MsCPR219	CPR-RR2	scaffold01004, scaffold01304
Msex2.13641, Msex2.14046	MsCPR220	CPR-RR2	scaffold01004, scaffold01304
Msex2.13643	MsCPR221	CPR-RR2	scaffold01004
Msex2.13644	MsCPR222	CPR-RR2	scaffold01004
Msex2.13645A	MsCPR223	CPR-RR2	scaffold01004
Msex2.13645C	MsCPR224	CPR-RR2	scaffold01004
Msex2.13656, Msex2.14579	MsCPR225	CPR-RR2	scaffold01013, scaffold02441
Msex2.13657, Msex2.15550	MsCPR226	CPR-RR2	scaffold01013, scaffold02441
Msex2.13659	MsCPR227	CPR-RR2	scaffold01013
Msex2.13660	MsCPR228	CPR-RR2	scaffold01013
Msex2.13661	MsCPR229	CPR-RR2	scaffold01013
Msex2.13662	MsCPR230	CPR-RR2	scaffold01013
Msex2.13663	MsCPR231	CPR-RR2	scaffold01013
Msex2.13664	MsCPR232	CPR-RR2	scaffold01013
Msex2.13665	MsCPR233	CPR-RR2	scaffold01013
Msex2.13666	MsCPR234	CPR-RR2	scaffold01013
Msex2.15544	MsCPR235	CPR-RR2	scaffold01013
Msex2.13667	MsCPR236	CPR-RR2	scaffold01013
Msex2.13668	MsCPR237	CPR-RR2	scaffold01013
Msex2.13669	MsCPR238	CPR-RR2	scaffold01013
Msex2.13670	MsCPR239	CPR-RR2	scaffold01013
Msex2.13671	MsCPR240	CPR-RR2	scaffold01013

OGS2 Gene ID	Gene Name [a]	Family	Scaffold No. [b]
Msex2.13672	MsCPR241	CPR-RR2	scaffold01013
Msex2.13814	MsCPR242	CPR-RR2	scaffold01084
Msex2.14031	MsCPR243	CPR-RR2	scaffold01333
Msex2.14032, Msex2.14432	MsCPR244	CPR-RR2	scaffold01333, scaffold01931
Msex2.14033, Msex2.15549	MsCPR245	CPR-RR2	scaffold01333, scaffold01931
Msex2.14385	MsCPR246	CPR-RR2	scaffold01713
Msex2.14386	MsCPR247	CPR-RR2	scaffold01713
Msex2.14960	MsCPR248	CPR-RR2	scaffold05463
Msex2.15340	MsCPR249	CPR-RR2	scaffold12390
Msex2.15463	MsCPR250	CPR-RR2	scaffold17868
Msex2.06360	MsCPR146	CPR-RR3	scaffold00136
Msex2.11251	MsCPR147	CPR-RR3	scaffold00476
Msex2.07106	MsCPR148	CPR-RR3	scaffold00167
Msex2.04414	MsCPR149	CPR-RR3	scaffold00081
Msex2.13160, Msex2.11877	MsCPAP1-A	CPAP1	scaffold00817, scaffold00519
Msex2.00613	MsCPAP1-B	CPAP1	scaffold00007
Msex2.00614	MsCPAP1-B2	CPAP1	scaffold00007
Msex2.00381	MsCPAP1-C	CPAP1	scaffold00004
Msex2.03859	MsCPAP1-D	CPAP1	scaffold00064
Msex2.02135	MsCPAP1-F	CPAP1	scaffold00032
Msex2.00269	MsCPAP1-G	CPAP1	scaffold00006
Msex2.02134	MsCPAP1-H	CPAP1	scaffold00032
Msex2.03076	MsCPAP1-I	CPAP1	scaffold00054
Msex2.03236	MsCPAP1-J	CPAP1	scaffold00052
Msex2.01703	MsCPAP1-K	CPAP1	scaffold00025
Msex2.09867	MsCPAP1-L	CPAP1	scaffold00311
Msex2.03103	MsCPAP1-M	CPAP1	scaffold00054
Msex2.02137	MsCPAP1-N	CPAP1	scaffold00032
Msex2.08722	MsCPAP1-O	CPAP1	scaffold00236
Msex2.08805	MsCPAP3-A1	CPAP3	scaffold00255
Msex2.08806	MsCPAP3-A2	CPAP3	scaffold00255
Msex2.08808, Msex2.15015, Msex2.14785	MsCPAP3-B	CPAP3	scaffold00255, scaffold06133, scaffold01361
Msex2.08810	MsCPAP3-C	CPAP3	scaffold00255
Msex2.08807, Msex2.14226	MsCPAP3-D1	CPAP3	scaffold00255, scaffold01383
Msex2.04890	MsCPAP3-D2	CPAP3	scaffold00101
Msex2.03293	MsCPAP3-E	CPAP3	scaffold00068
Msex2.03294, Msex2.15120	MsCPAP3-E2	CPAP3	scaffold00068, scaffold07708
Msex2.03295	MsCPAP3-E3	CPAP3	scaffold00068
Msex2.14886	MsCPAP3-E4	CPAP3	scaffold04686

OGS2 Gene ID	Gene Name [a]	Family	Scaffold No. [b]
Msex2.06560	MsCPT1	Tweedle	scaffold00143
Msex2.04779	MsCPT2	Tweedle	scaffold00091
Msex2.00101	MsCPT3	Tweedle	scaffold00003
Msex2.09186	MsCPT4	Tweedle	scaffold00268
Msex2.03833	MsCPH1	CPCFC	scaffold00064
Msex2.05601	MsCPF1	CPF	scaffold00116
Msex2.02569	MsCPFL1	CPFL	scaffold00039
Msex2.04027	MsCPFL2	CPFL	scaffold00070
Msex2.04028	MsCPFL3	CPFL	scaffold00070
Msex2.04036	MsCPFL4	CPFL	scaffold00070
Msex2.04037	MsCPFL5	CPFL	scaffold00070
Msex2.04035	MsCPFL6	CPFL	scaffold00070
Msex2.01072	MsCPH15	18 aa	scaffold00013
Msex2.01073	MsCPH16	18 aa	scaffold00013
Msex2.03612	MsCPH30	18 aa	scaffold00059
Msex2.03635	MsCPH31	18 aa	scaffold00059

[a] Gene names were assigned based on putative orthology to *B. mori*. Red type indicates genes in which one to one orthology could not be clearly established and, therefore, new gene numbers were assigned (MsCPR154-250). Blue type indicates genes previously deposited in GenBank (names given in parentheses). Genes boxed in gray identify expansion in *M. sexta* (putative paralogs) for which only a single ortholog occurs in *B. mori*.

[b] Scaffolds boxed in yellow or orange contain clusters of five or more CP genes.

[c] Genes MsCPR152 and 153 are orthologous to *B. mori* LOC101743422.

Table 2

Orthologous groups of CPR RR-2 genes between *B. mori* and *M. sexta* [a]

Group No.	*B. mori*	*M. sexta* [b]
1	CPR85, 86, 87, 88, 89	CPR183, 184, 185, 186, 187, 188, 189, 190, 191, 192, 193, 246, 247
2	CPR95, 96, 97, 98, 99, 100	CPR215, 216, 217, 218, 219, 220, 244, 245, 248
3	CPR101, 102, 103, 104, 105	CPR221, 222, 223, 224, 240, 241
4	CPR107, 109, 111, 120	CPR195, 196, 197, 199, 201, 207, 209, 225, 227, 229, 232
5	CPR106, 108, 110	CPR211, 213

[a] See supplementary figure S2 for phylogenetic analysis

[b] See Table 1 for OGS2 Gene ID

Table 3

***CPAP* gene orthologs from eight insect species**

Gene	*A. pisum*	*A. gambiae*	*A. melifera*	*B. mori*	*D. melanogaster*	*M. sexta*	*P. humanus*	*T. castaneum*
CPAP1-A	ACYPI45536	AGAP001203	GB51442	LOC101737637	CG32036	Msex2.13160	PHUM600940	TC004733
CPAP1-B	ACYPI001105	AGAP009480	GB44831	LOC101741926	CG14301	Msex2.00613		TC000587
CPAP1-B2	ACYPI007439	AGAP009479	GB44832	LOC101743233	CG14304	Msex2.00614	PHUM376120	TC000588
CPAP1-C		AGAP003751	GB42318	LOC101740411	CG14880	Msex2.00381		TC000316
CPAP1-D				LOC101744246		Msex2.03859		TC009263
CPAP1-E			GB41618		CG14959		PHUM034260	TC009887
CPAP1-F		AGAP006435	GB41624	LOC101746039	CG13675	Msex2.02135	PHUM034480	TC009893
CPAP1-G		AGAP007613	GB49021	LOC101737536	CG8192	Msex2.00269	PHUM071270	TC008877
CPAP1-H	ACYPI004632	AGAP005489	GB41625	LOC101740962	CG13676	Msex2.02134	PHUM263920	TC009894
CPAP1-I	ACYPI009675	AGAP002052	GB41792	LOC101742705	CG13643	Msex2.03076	PHUM575010	TC012766
CPAP1-J	ACYPI000845	AGAP010302	GB54921	LOC101747065	CG14608	Msex2.03236	PHUM355660	TC011101
CPAP1-K	ACYPI008601	AGAP001597	GB49734	LOC101736353	CG7549 (mtg)	Msex2.01703	PHUM601170	TC013568
CPAP1-L	ACYPI009002			LOC101746263	CG14607	Msex2.09867	PHUM355640	LOC657301
CPAP1-M	ACYPI002996	AGAP028105	GB48858	LOC101738097		Msex2.03103	PHUM135070	TC011724
CPAP1-N		AGAP005586	GB41623	LOC101740538	CG12009	Msex2.02137		TC009890
CPAP1-O		AGAP007089		LOC101735358	CG5756	Msex2.08722		
CPAP1 Unclassified	ACYPI30472		GB51439, GB53319				PHUM106940	
CPAP3-A1	ACYPI007911	AGAP000989	GB50636	LOC101739926	CG17052 (Obst-A)	Msex2.08805	PHUM434540	TC011140
CPAP3-A1-like		AGAP000987, AGAP000988						
CPAP3-A2	ACYPI006031		GB41945	LOC101740062		Msex2.08806		TC011141
CPAP3-B	ACYPI004093	AGAP009790	GB41227	LOC101740330	CG4778 (Obst-B)	Msex2.08808 + Msex2.15015	PHUM434540	TC011139
CPAP3-C	ACYPI007860	AGAP003308	GB41203	LOC733010	CG10287 (Obst-C; gasp)	Msex2.08810	PHUM179210	TC001169
CPAP3-D1	ACYPI009786	AGAP000986	GB41946	LOC101740197	CG17058 (Obst-D)	Msex2.08807	PHUM434550, PHUM434570	TC011142
CPAP3-D2	ACYPI001579	AGAP002909	GB41270	LOC101742643		Msex2.04890	PHUM434580	TC001350
CPAP3-E	ACYPI000583	AGAP009405	GB52854	LOC101743739	CG11142 (Obst-E1)	Msex2.03293	PHUM291220	TC011349
CPAP3-E2				LOC101743600	CG11142 (Obst-E2)	Msex2.03294		
CPAP3-E3						Msex2.03295		

Gene	*A. pisum*	*A. gambiae*	*A. mellifera*	*B. mori*	*D. melanogaster*	*M. sexta*	*P. humanus*	*T. castaneum*
CPAP3-E4				LOC101743460		Msex2.14886		

Gene IDs were derived from Jasrapuria (2011) and Ioannidou *et al.* (2014) except for *B. mori* and *D. melanogaster* which were from GenBank or Behr and Hoch (2005), and *M. sexta* from this report.

Table 4
Most abundant CP group expressed per library

CP Group	Library	FPKM[a]	%[b]
RR2	Head 4th instar HCS,	250,848	56,
	Fat body pupa days 15-18,	3,669	59,
	Testes pupa days 15-18	827	48
CPAP1	Fat body pupa days 1-3,	1,899	42,
	Midgut pupa day 1,	91	37,
	Ovaries adult day 1	135	40
CPAP3	Eggs 3 h	8,116	86
CPH	Head adult day 1	25,587	41
CPT	Testes pupa day 3	247	32
RR1	All others	32-402,401	30-90

[a] Total FPKM value for indicated CP group.

[b] Relative to total FPKM value for all CP genes.

Table 5

Top ten expressing CP genes in whole head libraries[a]

2nd instar day 1	3rd instar day 1	4th instar day 1	4th instar HCS	5th instar 12 h	5th instar day 2	5th instar pre-wandering	Pupa days 21-22	Adult day 1	Adult day 3	Adult day 7
CPR134	CPH15	CPR134	CPR216	CPR46	CPR169	CPR46	CPFL1	CPH31	CPR19	CPR19
CPH15	CPH1	CPR154	CPR241	CPR163	CPR164	CPR134	CPH31	CPH30	CPR56	Msex2.02570
CPH1	CPR154	CPH15	CPFL1	CPR169	CPR165	CPR63	CPR37	CPFL1	CPR37	CPR56
CPR170	CPR134	CPR156	CPR3	CPR63	CPR46	CPR163	CPFL3	CPFL5	CPH30	CPH30
CPR42	CPR156	CPR3	CPR43	CPR5	CPR166	CPR4	CPR73	CPR19	CPH31	CPR23
CPR154	CPR170	CPH1	CPR83	CPR4	CPR167	CPR169	CPR19	CPR37	Msex2.02570	CPR37
CPR156	CPR42	Msex2.02570	CPH31	CPR134	CPR168	Msex2.02570	CPR67	CPR63	CPR23	CPR67
CPR3	CPR63	CPFL1	CPR184	CPR154	CPR63	CPR3	CPH30	CPR10	CPFL1	CPR191
CPFL1	CPR3	CPH31	CPR215	CPR156	CPR42	CPR5	CPR32	CPR23	CPR67	CPR141
CPR63	Msex2.02570	CPR63	CPR248	CPH15	CPR6	CPR156	CPR74	CPAP3-Ca	CPR41	CPH31

[a] As determined by FPKM values; red type indicates RR-1 genes and blue type indicates RR-2 genes.