Published in final edited form as:

Insect Biochem Mol Biol. 2018 December ; 103: 53–69. doi:10.1016/j.ibmb.2018.10.006.

Building a platform for predicting functions of serine protease-related proteins in *Drosophila melanogaster* and other insects

Xiaolong Cao and **Haobo Jiang**[*]

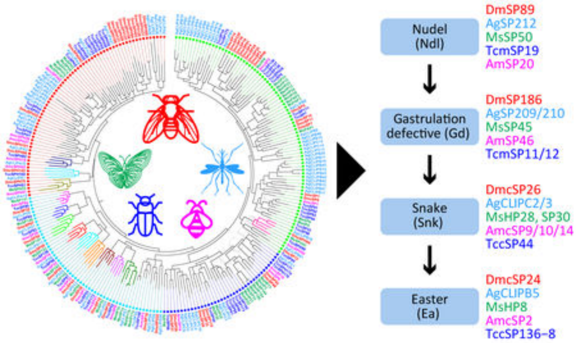
Department of Entomology and Plant Pathology, Oklahoma State University, Stillwater, OK 74078, USA

Abstract

Serine proteases (SPs) and serine protease homologs (SPHs) play essential roles in insect physiological processes including digestion, defense and development. Studies of insect genomes, transcriptomes and proteomes have generated a vast amount of information on these proteins, dwarfing the biological data acquired from a few model species. The large number and high diversity of homologous sequences makes it a challenge to use the limited functional information for making predictions across a broad taxonomic group of insects. In this work, we have extensively updated the framework of knowledge on the SP-related proteins in *Drosophila melanogaster* by identifying 52 new SPs/SPHs, classifying the 257 proteins into four groups (CLIP, gut, single- and multi-domain SPs/SPHs), and detecting inherent connections among phylogenetic relationships, genomic locations and expression profiles for 99 of the genes. Information on the existence of specific proteins in eggs, larvae, pupae and adults is presented to facilitate future research. More importantly, we have developed an approach to reveal close homologous or orthologous relationships among SPs/SPHs from *D. melanogaster*, *Anopheles gambiae*, *Apis mellifera*, *Manduca sexta*, and *Tribolium castaneum* thus inspiring functional studies in these and other holometabolous insects. This approach is useful for tackling similar problems on large and diverse protein families in other groups of organisms.

Graphical Abstract

[*]Corresponding author: Haobo Jiang, 127 Noble Research Center, Department of Entomology and Plant Pathology, Oklahoma State University, Stillwater, OK 74078, Tel: (405)-744-9400, Fax: (405)-744-6039, haobo.jiang@okstate.edu.



Keywords

phylogenetic analysis; gene duplication; chromosomal location; insect immunity; expression profiling; hemolymph protein; clip domain; serine protease cascade

1. Introduction

Serine proteases (SPs) of the S1A subfamily play important roles, ranging from digestion of dietary proteins to blood coagulation, complement activation, fertilization, and tumor metastasis in mammals (Foley and Conway, 2016; Wojtukiewicz et al., 2016). In arthropods, some SPs also form cascade pathways to establish dorsoventral polarity of embryos, mediate immune responses in various life stages, and cause hemolymph coagulation in horseshoe crabs (Kanost and Jiang, 2015; Kawabata and Muta, 2010). SPs digest foods and process protoxins in the midgut of most insects, including agricultural pests and human disease vectors (Rukmini et al., 2000; Zhu-Salzman and Zeng, 2015). Due to their crucial roles in various physiological processes, SPs and their non-catalytic homologs (SPHs) have been actively investigated in several species that are research models and/or agricultural pests (*e.g. Drosophila melanogaster*, *Tribolium castaneum*, *Manduca sexta*, *Helicoverpa armigera*, *Tenebrio molitor*), and insects of medical relevance (*e.g. Anopheles gambiae*, *Aedes aegypti*). SPHs are enzymatically inactive due to the lack of catalytic residue(s) but some of them (*e.g.* Masquerade) have known regulatory functions, including somatic muscle attachment, axonal guidance, post-mating responses, and prophenoloxidase (proPO) activation by proPO activating proteases (PAPs) (Kwon et al., 2000; Murugasu-Oei et al., 1996; Yu et al., 2003). POs catalyze the formation of reactive compounds and melanin that kill and sequester pathogens and parasites (Zhao et al., 2011). Upon infection, SPs also generate active cytokines such as Spatzle and stress responsive peptides that bind receptors (*e.g.* Tolls) to trigger intracellular signal transduction (Cao et al., 2015a; Schrag et al., 2017). Many members of the SP cascades in insects contain a disulfide-stabilized structure named clip domain and these SPs and SPHs are abbreviated as CLIPs and classified in subgroups A–E (Kanost and Jiang, 2015; Veillard et al., 2016).

In the past two decades, technological advances have greatly facilitated the acquisition of massive amounts of sequence data from living organisms, which led to the discovery of many genes homologous to those with known functions. For instance, there are 176 SP-related genes in the human genome (Puente et al., 2003) and, despite all the efforts to understand their physiological roles, only a small portion of them have been characterized at the functional level. Similarly, the holometabolous insects *D. melanogaster* (Ross et al., 2003), *A. gambiae* (Cao et al., 2017; Christophides et al., 2004), *Apis mellifera* (Zou et al., 2006), *T. castaneum* (Zou et al., 2007), and *M. sexta* (Cao et al., 2015b) possess 204, 337, 58, 168, and 193 SP-related genes in their genomes, respectively. Investigators routinely construct neighbor-joining trees to study evolutionary relationships among all members of the protein family in a single insect, since it takes a short time for the computation. Analyses of sequences from multiple species based on neighbor joining are not that trustworthy even if comparisons are limited to a subgroup of these proteins, such as CLIPs (Waterhouse et al., 2007). Thus, systematic classification and comparison of SP-like proteins across different orders of insects, which demand hundreds of highly reliable sequences and serial phylogenetic analyses, have not yet been reported. Another difficulty in such phylogenetic studies stems from the deeply rooted evolution of SP and SPH genes. Branching near roots of the trees is particularly problematic, as often indicated by low bootstrap values. The large numbers of SP-like genes and their high degrees of sequence variations in insects suggest that multiple series of gene duplication and sequence divergence occurred at different rates in some lineages of insects or subgroups of SP-related proteins. These complications interfere with the correct assignment of orthologous relationships among homologs in insects, the largest group of animals. Orthology is crucial for reliable extension of functional information from model species (*e.g. D. melanogaster*, *M. sexta*) to other insects of practical importance. This problem also exists for discovering such relationships in other large, diverse protein families (*e.g.* zinc finger proteins, Leu-rich repeat proteins, serine esterases) from various groups of living organisms.

To address these issues, we selected five species of holometabolous insects with broad biological significance, extensively updated the information on SP-related proteins in three of the five, and classified the resulting protein sequences into three to four large groups on the basis of domain structures, expression patterns and initial phylogenetic analyses. After that, the protein sequences from the five species were subjected in individual groups to sequential Bayesian analyses to reveal phylogenetic relationships. To better understand the functional data from *Drosophila*, we uncovered close correlations among chromosomal locations, phylogenetic tree positions, and expression profiles of the SPs and SPHs. Their mRNA levels and protein abundance in different tissues or life stages were extracted from published datasets to further assist the selection of genes for targeted reverse genetic analyses in the future. Functional data from *M. sexta* and the other species were integrated through hypothetical orthology to identify candidate genes for research in not only *Drosophila* but also mosquitoes (*e.g. A. gambiae*, *A. aegypti*), bees (*e.g. A. mellifera*), wasps (*e.g. Nasonia vitripennis*), moths (*e.g. H. armigera*), beetles (*e.g. T. castaneum*), and other insects. In summary, these explorations represent our newest efforts to organize and utilize a sizable body of information to expedite the functional elucidation of SP-related proteins based on their evolutionary relationships.

2. Materials and methods

2.1. *Identification and annotation improvement of the SP-related proteins in* D. melanogaster, T. castaneum, *and* A. mellifera

OPS Dmel-r6.09 was downloaded from FlyBase (https://flybase.org/). Domains in the Dmel-r6.09 sequences were identified on a local supercomputer using InterProScan5 v5.17 (Jones et al., 2014). Proteins containing a chymotrypsin-like (*i.e.* S1A subfamily) SP-related domain were extracted and combined. 376 RNA-seq files from the modENCODE project (Brown et al., 2014), 1.7 billion reads representing 30 different cDNA samples of whole insects at different life stages (Table S4), were downloaded from NCBI Sequence Read Archive (SRA). The reads were trimmed by Trimmomatic (Bolger et al., 2014) with the setting “SLIDINGWINDOW:4:20 LEADING:20 TRAILING:20 MINLEN:50” to remove low-quality bases and assembled by Trinity-2.3.2 (Haas et al., 2013), with the k-mer length of 25. SP-like sequences in Dmel-r6.09 were crosschecked and improved with the Trinity models. After removal of alternatively spliced and severely incomplete genes, the sequences were manually examined according to characteristic features of the S1A SPs, such as signal peptide and conserved regions. The SPs were compared with the newer OPS, Dmel-r6.14, to ensure all SPs were included. Similarly, *T. castaneum* and *A. mellifera* protein sequences were downloaded from NCBI RefSeq using accession numbers GCF_000002335 (Tcas5.2) and GCF_000002195 (Amel4.5). Domain(s) in the beetle and bee SP-like proteins were found using InterProScan5. To improve the SP/SPH gene models in Tcas5.2 and Amel4.5, all the RNA-seq data deposited before 6-1-2017 (6 projects of *T. castaneum*: PRJNA247821, PRJNA376868, PRJNA315762, PRJNA305354, PRJNA275195 and PRJNA266839; 5 projects of *A. mellifera*: PRJNA268450, PRJNA386067, PRJNA325930, PRJNA322249 and PRJNA275154) (Dippel et al., 2014; Greenwood et al., 2017; Kim et al., 2018; Manfredini et al., 2015; Mao et al., 2017; Stappert et al., 2016) were downloaded from NCBI SRA for read trimming and assembling with Trinity. The final outputs contained limited numbers of transcripts, possibly due to SNPs in the cDNA samples. As such, the intermediate files (data not shown) generated by Inchworm, a part of Trinity, were compared with the corresponding official gene models prior to manual improvement of the protein sequences. BLASTP searches against NR database of NCBI were also performed to improve the sequences, as some newly verified proteins were not updated in the RefSeq models.

2.2. *Properties of the SPs and SPHs in* D. melanogaster, T. castaneum, *and* A. mellifera

The SP-related sequences in each species were divided into SPs or SPHs based on the presence or absence of the His-Asp-Ser catalytic triad as described before (Cao et al., 2015b). Signal peptide was predicted using SignalP 4.1 (http://www.cbs.dtu.dk/services/SignalP/) (Petersen et al., 2011) and Signal-3L (http://www.csbio.sjtu.edu.cn/bioinf/Signal-3L/) (Shen and Chou, 2007). Some clip domains (as well as other structural modules) were identified by InterProScan5 while other clip domains were found through manual inspection of the sequences for a Cys doublet in the region close to the protease or protease-like domain (PD or PLD). SPs and SPHs with four additional Cys residues upstream of the doublet were designated CLIPs to indicate the presence of a clip domain (Kanost and Jiang, 2015). Residues 190, 216 and 226 (chymotrypsin numbering) (Perona and Craik, 1995) that form the primary substrate-binding pocket of each PD were identified in the aligned sequences for protease specificity prediction (Cao et al., 2015b).

2.3. *Classification, multiple sequence alignment, and phylogenetic analyses of the SP-like proteins in* D. melanogaster, A. mellifera, *and* T. castaneum

The SP-related proteins in the three species were first classified into C (for CLIP), M (for multi-domain), and G-S (for gut and other single domain) groups based on the presence/absence of clip domain(s) and then two or more non-clip domains. While the G-S groups remained intact in *A. mellifera* and *T. castaneum*, single domain SPs/SPHs in *D. melanogaster* were further separated into G and S groups based on their expression profiles and positions in the phylogenetic tree of G-S. Members of the G group tend to associate with feeding, high-level expression, a typical size of about 270 residues, some gene clusters on chromosomes, and certain positions in the phylogenetic tree (Cao et al., 2017; Cao et al., 2015b). Multiple sequence alignments of the entire proteins in the C, G and S-M groups from *D. melanogaster* were performed using MUSCLE (Edgar, 2004), a module of MEGA 7.0 (Kumar et al., 2016), under the default settings with maximum iterations changed to 1,000. Alignments of the individual groups were converted to NEXUS format by MEGA, and phylogenetic analyses were conducted using MrBayes v3.2.6 (Ronquist et al., 2012) under the default model with the setting “nchains=12”. MCMC (Markov chain Monte Carlo) analyses were terminated after the standard deviations of two independent analyses was <0.01 or after 10 million generations (SD < 0.05). FigTree 1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/) was used to display the trees. To identify orthologous or close homologous relationships among CLIPs, S-group SP(H)s, Gd-domain SPs/SPH, and CUB-domain SPs from the five species, the sequences in each group were analyzed using MUSCLE and MrBayes. Based on the initial analysis of the 247 CLIPs, sequences of CLIPAs, CLIPBs, CLIPCs, and CLIPDs from the five species were separately examined by the Bayesian method.

2.4. *Chromosomal locations and expression profiling of the SP and SPH genes in* D. melanogaster

For most of the SP-related genes, their genomic locations were available in the information lists of the Dmel-r6.09 models. Retrieved position data were plotted using ArkMAP 2.0 (http://www.bioinformatics.roslin.ed.ac.uk/arkmap/) and improved using Adobe Illustrator. After excluding data from cell lines or challenged insects (Brown et al., 2014), the remaining 30 and 26 additional datasets (Table S4) were downloaded from NCBI SRA, representing whole bodies or tissues of healthy *D. melanogaster* under 52 conditions. Reads were trimmed with Trimmomatic to remove adaptors and low quality bases with the setting “SLIDINGWINDOW:4:30 LEADING:20 TRAILING:20 MINLEN:50”. Transcript sequences of the SP-related proteins in Dmel-r6.09 were replaced with the improved ones (Section 2.1). FPKM (f̲ragments p̲er k̲ilobase of transcript per m̲illion mapped reads) values for genes in different libraries were calculated using Bowtie2 2.2.3 (Langmead and Salzberg, 2012) and RSEM 1.2.15 (Li and Dewey, 2011). FPKM values in libraries from biological replicates were averaged to represent gene expression in those types of samples. Hierarchically clustered gradient heat maps of $\log_2$(FPKM+1) values were plotted using the clustermap function of Seaborn (https://seaborn.pydata.org/), a Python data visualization library, with the average linkage method and Euclidean matrix.

2.5. *Abundances of the SP-related proteins in* D. melanogaster *eggs*, *larvae*, *pupae*, *and adults*

Label-free quantification (LFQ) values of the SPs/SPHs in whole insects at 14 life stages, each with four biological replicates (Table S5), were extracted from the published data (Casas-Vila et al., 2017). Proteins identified in at least 3 out of the 68 samples were kept. After normalization with 5×10^6, which is close to 5,103,000 or the lowest non-zero LFQ, relative protein abundances, as represented by $\log_2(\text{LFQ}/5\times10^6 + 1)$ values, were subjected to hierarchical cluster analysis and plotted as a gradient heat map.

3. Results and discussion

3.1. Current status of the research area and outlines of our approach

Genome-wide investigations of insect SPs and SPHs have facilitated functional studies of this family of proteins after the genome sequence of *D. melanogaster* was first published in 2000. We described and analyzed sequences of about 200 SP-related proteins in the fruit fly (Ross et al., 2003), and these were further annotated by functional residue clustering (Shah et al., 2008). Similar studies yielded overviews of the S1A protease subfamily in the yellow fever and African malaria mosquitoes (Cao et al., 2017; Waterhouse et al., 2007), honeybee (Zou et al., 2006), red flour beetle (Zou et al., 2007), silkworm (Zhao et al., 2010), diamond back moth (Lin et al., 2015), and tobacco hornworm (Cao et al., 2015b). However, due to the large family sizes, most of the investigations focused on CLIPs (clip-domain SPs and SPHs) but it has been quite difficult to define orthologous relationships, for example, among the *A. gambiae* and *D. melanogaster* CLIPBs. Incorrect gene models and incomplete genome coverage also compromised the quality of those studies, leading to weak support for functional predictions. When the role of an SP is known in species A but no clear ortholog stands out in species B, the amount of experimental work required to validate the assumed function increases proportionally with the number of candidates. To deal with these problems, we first improved the gene models in *D. melanogaster*, *T. castaneum* and *A. mellifera*, separated the family into 3–4 groups with similarities in sequence and domain structure, performed serial phylogenetic analyses at the group and subgroup levels along with the homologs in *A. gambiae* and *M. sexta*, and predicted functions mainly based on the literature and closest possible homolog.

3.2. Overview of the SP-related protein family in the five insect species with known genomes

To provide a summary of the family, we have selected five species from four different orders of holometabolous insects: *D. melanogaster* (Diptera) for its wide-ranging genetic data (Veillard et al., 2016), *A. gambiae* (Diptera) for its medical implications and available results on CLIPs (Cao et al., 2017), *A. mellifera* (Hymenoptera) for its agricultural importance and a small set of the core SPs and SPHs (Zou et al., 2006), *T. castaneum* (Coleoptera) for its unique phylogenetic position, amenability to RNAi assays, and representation of the largest order of animals (Zou et al., 2007), and *M. sexta* (Lepidoptera) for its extensive biochemical data on SP-related proteins and closeness to agricultural pests (Kanost and Jiang, 2015). Since the *M. sexta* and *A. gambiae* SPs and SPHs were recently examined (Cao et al., 2017; Cao et al., 2015b), we only updated the information on SP-like proteins in the fly, bee, and beetle.

3.2.1. *The fruit fly* D. melanogaster—In contrast to AgamP4.5, in which 117 of the 337 gene models were improved using the MCOT method (Cao and Jiang, 2015), the same approach only resulted in six corrections in the *Drosophila* sequences, reflecting the high quality of the genome assembly and gene annotation. Compared with results of the initial analysis (Ross et al., 2003) of *D. melanogaster* SP/SPH genes, major improvements are made in four areas (Table S1): 1) we identified 52 new SP or SPH genes in the genome, making 257 the total count of SP-like genes, 2) all CLIPs but four have a signal peptide, an entire clip domain, and a full-length protease or protease-like domain (PD/PLD). cSP44 and cSP56 have a predicted transmembrane region rather than a signal peptide, whereas cSP54 and cSP229 lack both, 3) we compiled a complete set of information including the group-chromosome-tree-expression (G-C-T-E) characterization for easily locating the genes on the chromosomes, phylogenetic trees, and expression profiles, and 4) we separated SPs and SPHs into four groups related to their domain structure and putative functions: C (for **C**LIP), M (for other **m**ulti-domain), G (for **g**ut) and S (for other **s**ingle domain) groups of SP-related proteins. Among the 190 SP and 67 SPH genes (Table 1) in the fly genome, 52 encode CLIPs (34 SPs and 18 SPHs with at least one clip domain), 21 are in the "M" group of non-CLIP, multi-domain SPs (17) and SPHs (4), 65 gut and 119 other single-domain SPs/SPHs.

Of the 257 proteins, 190 are predicted to be activated by cleavage after Arg(161)/Lys(29) by SPs with trypsin-like specificity, 26 after Phe(8)/Tyr(3)/Leu(15) by chymotrypsin-like SPs, and 11 after Ala(1)/Gly(2)/Val(1)/Ile(2)/Met(2)/Ser(3) by elastase-like SPs, and 5 after His by SPs with special features. Regarding catalytic specificity, 114, 61 and 15 of the 190 SPs have trypsin-, chymotrypsin-, and elastase-like properties, respectively. The dominance of trypsin-like SPs is also clear in the other insects (Cao et al., 2015b and 2017; Tables S2 and S3).

3.2.2. *The honeybee* A. mellifera—Compared with the initial analysis (Zou et al., 2006), there are major improvements in the honeybee SP/SPH sequences: 44 of the 57 SP-related gene models had been updated by BeeBase personnel; 31 of the updated and 2 newly identified genes were fully confirmed using the assembled RNA-seq data; 13 of the previously updated gene annotations were further improved in this work (Table S2). The genome contains 46 SP and 13 SPH genes encoding 14 SPs and 7 SPHs with at least one clip domain (C), 10 SPs and 3 SPHs with other domains (M), and 22 SPs and 3 SPHs with only a PD or PLD (G-S). In the third group, the shorter ones are more likely related to digestion whereas the longer ones may participate in processes that need regulatory region(s) for proper functioning.

3.2.3. *The red flour beetle* T. castaneum—The repertoire of 98 SP and 70 SPH genes in the beetle (Zou et al., 2007) is larger than that in the bee. After we improved the gene models using the RNA-seq data, the genome encodes 55 CLIPs (36 SPs and 19 SPHs), 20 SP-related proteins with other domains (M, 18 SPs and 2 SPHs), and 102 with one

PD/PLD (G and S unseparated, 28 SPs and 39 SPHs with 231–270 residues; 24 SPs and 11 SPHs with 271–387 residues). With nine newly identified SP/SPH genes, a total of 177 SP-related genes (Table S3) are further examined below.

3.3. *General structural features and classification of the SP-related proteins in* D. melanogaster

Among the 52 CLIPs most have a single clip domain. The exceptions are cSP14, cSP18, cSPH58, cSPH121, cSPH142, and cSPH79, which contain 4, 2, 2, 2, 3, and 5 clip domains, respectively (Table 1). InterProScan5 only recognized 25 of the 64 clip domains in these CLIPs but it successfully identified 21 non-CLIP multi-domain SP-like proteins, including SP89/Ndl, SP186/Gd, SP53/ModSP, SP72/Tequila, and SP75/Corin (Fig. 1). There are 12 different types of regulatory domains (*e.g.* LDL, Sushi) in these proteins. SP55, SP60, SPH144, SP186/Gd, and SP212 have a similar domain structure of 1–2 Gd-PD/PLD; SP80a, SP80b, and SP120 have a CUB domain followed by a PD; SP49, SP77, SP222, SP248, and SP251 contain a PD and a PLD; SPH241 and SPH247 have two tandem PLDs. Functions of the CLIPs and other multi-domain SPs/SPHs are worth exploring in the future. The classification of gut (G) and single-domain (S) SPs/SPHs is not as clear-cut as CLIPs or multi-domain SPs/SPHs. Therefore, we took a combined approach (Section 2.3) and separated them into 57 gut SPs, 8 gut SPHs, 82 single-domain SPs, and 37 single-domain SPHs (Table 1).

3.4. *Phylogenetic relationships, chromosomal locations, and expression patterns of the* D. melanogaster *SP-related genes*

The phylogenetic trees exhibit complex relationships among the CLIPs, gut, and single- and multi-domain SPs/SPHs (Fig. 2). In panel A, all 18 CLIPBs (cSP38 to cSP12, cSP1/Grass, cSP8) reside in branch A; all 7 CLIPCs (cSP42, 48, 115, 33, 26, 31, 28) constitute the entire B branch; 11 CLIPDs (cSP36, 232-18, 67-54, 34-19, and 44-32-59-56) are located in branches a, A, b, c and C, respectively; 16 CLIPAs (cSPH142/Scaf to cSPH125) are in branches d, D–F, and e. Note that the CLIP A–D classification system was developed on the basis of sequence alignments (Fig. S1) (Kanost and Jiang, 2015) and amended to include CLIPEs (Cao et al., 2017). In panel B, 26 gut SPs and 1 gut SPH are analyzed: 16 Jonahs (Jon25B$_{\text{i–iii}}$, Jon44E, Jon65A$_{\text{i–iv}}$, Jon66C$_{\text{i, ii}}$, Jon74E, 99C$_{\text{i–iii}}$, Jon99F$_{\text{i, ii}}$) on branch G and 11 trypsins (α–ι, λ, and κ/SPH109) on branch I. In panel C, two groups of multi-domain SP-like proteins are discovered: branch R has three CUB-domain SPs (80a, 80b and 120); branch V consists of Gd-domain SP55, SP60, SP186/Gd, SP212 and SPH144 (Fig. 1). Other close relationships occur in groups G, S and M, with large branches (*e.g.* G, I, L, N) formed as a result of extensive putative gene duplications.

Locations of SP-like genes on chromosomes also reflect the history of family expansion (Fig. 3). There are 32 clusters (A–Z, 1–6) containing 3 to 13 genes in each. For example, all thirteen genes (trypsins α, β, γ, δ, ε, ζ, η, θ, ι, κ/SPH109, λ, SP6, and SP215) in cluster I belong to the G group and are responsible for digestion of dietary proteins (Davis et al., 1985). Likewise, clusters C, P and #4 contain 3, 4, and 3 Jonah genes that also participate in food digestion (Akam and Carlson, 1985; Carlson and Hogness, 1985). Except for SP186/Gd, all the Gd-domain SPs/SPH are in cluster V. Other gene clusters include H and Z for CLIPs, F for multi-domain, W and #6 for gut, and B, J, K and N for single-domain SP(H)s (Fig. 3). Interestingly, SPs/SPHs involved in development or immunity (*e.g.* Ndl, Gd, Snk, Ea, Mas, Sb, Tequila, ModSP, Grass, Psh, SPE, MP1) are often encoded by genes that are not members of clusters. Maintenance of the solitary state and, hence, proper dosage could be crucial for the conserved functions. A simpler explanation is that mutation in one member of a gene cluster with overlapping function may not produce a phenotype and, in other words, finding these mutants may be the result of a bias toward genes with unique function (*e.g.* defense).

Gene expression patterns associate with functions. Thus, we extracted and analyzed the publicly available RNA-seq data on *D. melanogaster* SPs and SPHs. The highest transcript levels occur in the gut group (Fig. 4B). Except for SP50, SP107, SP141 and SPH196a, all the SP-related transcripts in the G group are detected at high levels [$\log_2$(FPKM+1): 8–16] in larval and adult gut tissues. The mRNA levels are so high in most cases that they are easily detected in the whole body samples at the feeding stages. The low expression in late embryo (16–24 h) may allow the 1st instar to digest food right after hatching. In contrast, the production of digestive proteases ceases completely in the pupal stage until right before adult emergence (Carlson and Hogness, 1985). In comparison, SPs and SPHs in the S and M groups are expressed at low levels [$\log_2$(FPKM+1): 1–7] (Fig. 4C). A group of genes with similar expression profiles is defined as an assembly here. Single-domain SP63/Send1 and SP190/Send2 are in assembly I; SPH255/Aqrs, SPH194a/Spx1, SPH194b/Spx2, SPH202/Intr and SP168/Sems are in assembly M. Sphinx1/2 are required for Toll activation (Kambris et al., 2006). Send1, Send2, Aquarius, Intrepid, Seminase, and others form a protein network required for sperm storage in mated females (Findlay et al., 2014; Schnakenberg et al., 2011). Multi-domain SP75/Corin and SP89/Ndl belong to assembly L; SP72/Tequila and SP53/ModSP to N. Detailed expression patterns of these and other related genes in the 52 datasets are shown (Fig. 4C). Among the CLIPs, cSP3, cSP61, cSPH93, and cSPH156 have unique expression patterns, whereas expression profiles of the other genes fall into assemblies A–D (Fig. 4A).

Gene duplication is a major mechanism for forming gene clusters. In this process, regulatory elements are often duplicated with the coding region, leading to similar profiles of expression of the gene copies unless mutations occur in these elements. We thus predicted and later observed a considerable correlation among phylogenetic relationships (T), chromosomal locations (C), and expression patterns (E) (Fig. 5). Of the 257 genes, 216, 147, and 249 form branches, clusters, and assemblies in the T, C and E groups, respectively, and 99 genes in 33 groups have T-C-E three-way consistency (Fig. 5, B and C). Such relationships are vital for future research on these genes by taking into account possible functional redundancy and coordinated regulation of gene expression. For instance, we may not observe any phenotype if we only knock out 1 or 2 of cSP32, cSP44 and cSP59.

3.5. Orthologous relationships and function prediction of SP-related proteins from insects

Orthology disclosed by multiple sequence comparison and phylogenetic analysis is critical for inferring equivalence of function. To identify such relationships, we performed an initial phylogenetic analysis of all CLIPs from the fly, mosquito, bee, moth and beetle, and found that *A. gambiae* subgroup E differs drastically from subgroups A–D (Cao et al., 2017). After removing the CLIPEs, we generated a MrBayes tree of 247 CLIPs (Fig. S1), which exhibits a clear separation of clades A–D (probabilities or p: 0.95 to 0.99). We then aligned the sequences in each subgroup and constructed five trees using MrBayes v3.2.6 (Fig. 6). Three CLIPA branches and one CLIPB branch display 1:1:1:0:1 or 1:1:1:1:1 (Dm:Ag:Am:Ms:Tc, same species order below) orthologous relationships (p: 1.00). Lineage-specific expansions are common in subgroups A–C and E, suggesting that, in each cohort of descendants, at least one member is functional in that species to play their ancestor's role. In comparison, seven 1:1:1:1:1, one 1:1:1:2:1, one 1:1:0:1:1, and one 2:3:1:2:2 orthologous sets (p: 1.00) constitute the entire subgroup D. We assume ancestors of this subgroup existed long before the radiation of holometabolous insects but do not understand why gene duplication seldom occurred. In spite of the complexity in parts of the trees (Fig. 6, A–C), our stepwise analysis of CLIPs yielded insights into their evolutionary relationships useful for guiding focused functional studies.

Comparison of protein domain structures is another way to uncover hypothetical orthologous relationships among non-CLIP multi-domain SPs/SPHs, since their regulatory modules are likely associated with unique functions of the entire proteins. As shown in Table 2, nine orthologous groups (*e.g.* Nudel, Tequila, Corin, ModSP, Masquerade, Gd) are identified in these five insects with 1:1:1:1:1, 1:1:1:0:1 or a:b:c:d:e, where a–e are any integers from 1 to 6. While no functional data is available for any member of the SEA, TSP and CUB sets, future research should cast light on roles of these proteins with the conserved domain structures.

We have used the known SP-SPH pathways in *D. melanogaster* and *M. sexta* as templates to overlay their putative orthologs or orthologous groups and hypothesize the presence of similar systems in the other species, which likely stemmed from their last common ancestors existing before the radiation of holometabolous insects (Fig. 7). By inspiring research to test the overarching hypothesis, the current knowledge on these systems will be expanded and enriched. Panel A surmises that the *Drosophila* embryonic SP pathway for establishing dorsoventral polarity (Moussian and Roth, 2005) may be conserved in the five insects of complete metamorphosis. The detection of *M. sexta* SP50, SP45, HP28/SP30 and HP8 transcripts in adult ovaries and early embryos (Cao et al., 2015b) provided initial experimental support for the proposed SP cascade in the moth. Besides, if silencing of *T. castaneum* mSP19, mSP11/12, cSP44, or cSP136–138 gene expression in ovaries impairs embryo ventralization, the pathway may be functional in the beetle as well.

The immune SP-SPH systems for proPO and proSpätzle activation in hemolymph have been understood to different degrees in several insects (Fig. 7B) (Kanost and Jiang, 2015; Park et al., 2010; Veillard et al., 2016). PO produces reactive compounds to kill and sequester pathogens; Spatzle elicits the Toll pathway to induce antimicrobial peptide synthesis. We propose that the putative protease orthologs may play similar roles: *M. sexta* HP14a/b corresponds to DmModSP, AgSP217, TcmSP3/13 and AmSP49; MsHP1a/b is closely related to DmcSP34, AgCLIPD1, TccSP52 and AmcSP8. While these initiation steps often involve one ortholog per species, subsequent steps of the pathways contain 1 to 9 SP/SPH paralogs, depending on the extent of lineage-specific gene expansion. For instance, *M. sexta* HP2, HP13, HP18a/b. HP21, HP22, SP33, and SP144 form one orthologous set. MsHP14a activates proHP2 in wandering larvae and pupae, and HP2 activates proPAP2 (He et al., 2018). HP21, activated by MsHP14a and activating both proPAP2 and 3 (Gorman et al., 2007; Wang and Jiang, 2007), has a functional overlap with HP2, but it acts mainly in hemolymph of the feeding larvae. In another example, *Drosophila* cSP24/Easter and cSP4/SPE activate Spätzle in embryos and adults (Jang et al., 2006; Mulinari et al., 2006), respectively, and so does their ortholog MsHP8 in larval hemolymph (An et al., 2010). It is possible that DmcSP3 and the other six CLIPBs in the same orthologous set (Fig. 7B) activate DmSpätzle1–6 in the same or different tissues or life stages. Likewise, functional redundancy may be extensive in the group of *M. sexta* SPH1 and the eleven *A. gambiae* CLIPA/E's. While MsPAP1–3 all need the presence of a high M_r complex of SPH1 and SPH2 to generate highly active PO (Wang and Jiang, 2004), it is unclear if AgCLIPB9 also needs one or more of CLIPA4–7, 12–14, 26, 31, 32, E6, and E7 to generate active POs. MsPAP1 or AgCLIPB9 alone cleaved *M. sexta* proPOs at the correct peptide bond but yielded POs with a low specific activity (An et al., 2011; Gupta et al., 2005) and so did DmMP2 to DmPPO1 (An et al., 2013).

Genetic research on *Drosophila* immunity provided support for the hypothetical immune SP-SPH system (Fig. 7B). ModSP integrates signals from pattern recognition receptors and indirectly activates Grass and then cSP4/SPE (Buchon et al., 2009). Microbial proteases can activate cSP28/Psh that induces the Toll pathway via SPE and Spätzle (Issa et al., 2018). A wound in the integument induces cSP31/Hayan activation and melanization (Nam et al., 2012). The functions of Psh and Hayan are consistent with the central role of *M. sexta* HP6 in proSpatzle-1 and proPO activation (An et al., 2009 and 2010). The roles of cSP25/MP1 and cSP7/MP2 appear controversial: *in vitro* test indicates that MP2 is a PAP (An et al., 2013) but *in vivo* data suggest that MP2 activates proMP1 and MP1 activates proPO (Tang et al., 2006). This situation may reflect the limitation of our approach or, in contrast, the limitation of a genetic study in discerning the position or order of pathway members.

3.6. *Levels of SP-related proteins in* Drosophila *whole body samples at different life stages*

Protein abundances are useful information for functional studies, simply because the concerned proteins must exist at certain levels to act properly in a specific tissue or stage. This is particularly important if a large set of orthologs (*e.g.* DmcSP3, 4/SPE, 14, 16, 24/Easter, 25/MP1, 38, 61, 229) is in question. mRNA levels are informative and easier to get, but may not be well correlated with protein abundances. For instance, levels of 571 transcripts in *M. sexta* fat body corresponded with concentrations of their proteins in hemolymph with correlation coefficients of 0.41 (naive larvae) and 0.43 (immune challenged) (He et al., 2016). Therefore, we extracted LFQ values of SPs/SPHs from the published work (Casas-Vila et al., 2017) in whole insects at various life stages (Table S5) and displayed their relative abundances in the labeled heat map (Fig. 8). Among the 110 proteins identified, 44 are GPs/GPHs, 29 are other single domain SPs/SPHs, 30 are CLIPs, and 7 are non-CLIP, multi-domain SPs/SPHs. They account for 68%, 24%, 58%, and 41% of the G (65), S (119), C (52), and M (17) groups, respectively. Overrepresentation of the G and C groups indicates gut SPs/SPHs and CLIPs are present at higher levels than the average to be detected in whole insects by mass spectrometry. In contrast, lower abundances of the single-domain SP(H)s lead to the S group underrepresentation, which is in general consistent with the transcriptome data (Fig. 4). When we compared the mRNA and protein levels of specific genes, discrepancies were noticed. For instance, SP89/Ndl mRNA levels [$\log_2$(FPKM+1): 1–4] are low and scattered (Fig. 4C) but its protein levels [$\log_2(\text{LFQ}/5\times10^6 + 1)$: 6–9] in adult females and eggs are among the highest (Fig. 8). Np/cSP59 mRNA (level: 1–5, Fig. 4A) is present in embryo, larva, pupa, but its protein (level: 6–8) is mainly found in pupa to affect bristle and wing development (Bridges et al., 1936). While the expression profiles of Np/cSP59, Sb/cSP56, and Masquerade/cSPH79 mRNA (levels: 1–6, Fig. 4A) closely resemble, the levels of Sb and Mas are below the detection limit in the whole insect. Fine dissection of tissue samples at various life stages and more sensitive proteomics analyses in the future should provide data on relative abundances of the SP-related proteins in this important model species.

3.7. Conclusions

Serine proteases (SPs) and their catalytically inactive homologs participate in food digestion, reproduction, embryo development, innate immunity, and other insect physiological processes. We have verified a total of 191 SP and 66 SPH genes in the *Drosophila* genome, 52 more than previously reported. These include 52 CLIPs, 21 multi-domain, 65 gut, and 119 other single-domain SPs/SPHs. Revelation of the links among phylogenetic relationships, genomic locations, and expression patterns for 99 SP-related genes in 33 groups accounts for a substantial portion of the evolutionary history of this gene family in the model organism. mRNA and protein levels in different tissues or stages are presented as background information to help with predictions of potential function. In addition, we have elucidated hypothetical orthologous relationships of SPs/SPHs from *D. melanogaster*, *A. gambiae*, *A. mellifera*, *M. sexta*, and *T. castaneum* to inspire functional studies and information exchange among researchers studying these holometabolous insects and beyond.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We would like to thank Drs. Michael Kanost and Ulrich Melcher for their critical comments, which help us greatly in the manuscript improvement. This work was supported by NIH grants AI112662 and GM58634. This work was approved for publication by the Director of Oklahoma Agricultural Experimental Station and supported in part under project OKLO2450. Computation for this project was performed at OSU High Performance Computing Center at Oklahoma State University supported in part through the National Science Foundation grant OCI–1126330.

Abbreviations:

SP	serine protease
SPH	(non-catalytic) serine protease homolog
PD	SP catalytic domain

PLD	protease-like domain in SPH
CLIP	clip-domain SP or SPH
GP and GPH	gut serine protease and gut serine protease homolog
CUB	a domain in **c**omplement 1r/s, **u**egf and **b**mp1
FPKM	fragments per kilobase of transcript per million mapped reads
G-C-T-E	group-chromosome-tree-expression
Gd	a domain in *Drosophila* **g**astrulation **d**efective
LDL	low-density lipoprotein receptor class A repeat
LFQ	label-free quantification
PO and proPO	phenoloxidase and its proenzyme
PAP	proPO activating protease
SEA	a domain in **s**perm protein, **e**nterokinase and **a**grin
SRA	sequence read archive
TSP	thrombospondin

References

Akam ME, Carlson JR, 1985 The detection of Jonah gene transcripts in *Drosophila* by in situ hybridization. EMBO J. 4, 155–161. [PubMed: 2410252]

An C, Budd A, Kanost MR, Michel K, 2011 Characterization of a regulatory unit that controls melanization and affects longevity of mosquitoes. Cell. Mol. Life Sci 68, 1929–1939. [PubMed: 20953892]

An C, Ishibashi J, Ragan E, Jiang H, Kanost MR, 2009 Functions of *Manduca* sexta hemolymph proteinases HP6 and HP8 in two innate immune pathways. J. Biol. Chem 284, 19716–19726. [PubMed: 19487692]

An C, Jiang H, Kanost MR, 2010 Proteolytic activation and function of the cytokine Spätzle in innate immune response of a lepidopteran insect, *Manduca sexta*. FEBS J. 277, 148–162. [PubMed: 19968713]

An C, Zhang M, Chu Y, Zhao Z, 2013 Serine protease MP2 activates prophenoloxidase in the melanization immune response of *Drosophila melanogaster*. PLoS One 8, e79533. [PubMed: 24260243]

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

Bridges CB, Skoog EN, Li J, 1936 Genetical and cytological studies of a deficiency (notopleural) in the second chromosome of *Drosophila melanogaster*. Genetics 21, 788–795. [PubMed: 17246819]

Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, Booth BW, Wen J, Park S, Suzuki A, Wan KH, Yu C, Zhang D, Carlson JW, Cherbas L, Eads BD, Miller D, Mockaitis K, Roberts J, Davis CA, Frise E, Hammonds AS, Olson S, Shenker S, Sturgill D, Samsonova AA, Weiszmann R, Robinson G, Hernandez J, Andrews J, Bickel PJ, Carninci P, Cherbas P, Gingeras TR, Hoskins RA, Kaufman TC, Lai EC, Oliver B, Perrimon N, Graveley BR, Celniker SE, 2014 Diversity and dynamics of the *Drosophila* transcriptome. Nature 512, 393–399. [PubMed: 24670639]

Buchon N, Poidevin M, Kwon HM, Guillou A, Sottas V, Lee BL, Lemaitre B, 2009 A single modular serine protease integrates signals from pattern-recognition receptors upstream of the *Drosophila* Toll pathway. Proc. Natl. Acad. Sci. USA 106, 12442–12447. [PubMed: 19590012]

Casas-Vila N, Bluhm A, Sayols S, Dinges N, Dejung M, Altenhein T, Kappei D, Altenhein B, Roignant JY, Butter F, 2017 The developmental proteome of *Drosophila melanogaster*. Genome Res. 27, 1273–1285. [PubMed: 28381612]

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. [PubMed: 25612938]

Cao X, Gulati M, Jiang H, 2017 Serine protease-related proteins in the malaria mosquito, *Anopheles gambiae*. Insect Biochem. Mol. Biol 88, 48–62. [PubMed: 28780069]

Cao X, He Y, Hu Y, Wang Y, Chen Y-RR, Bryant B, Clem RJ, Schwartz LM, Blissard G, Jiang H, 2015a The immune signaling pathways of *Manduca sexta*. Insect Biochem. Mol. Biol 62, 64–74. [PubMed: 25858029]

Cao X, He Y, Hu Y, Zhang X, Wang Y, Zou Z, Chen Y, Blissard GW, Kanost MR, 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. [PubMed: 25530503]

Carlson JR, Hogness DS, 1985 Developmental and functional analysis of Jonah gene expression. Dev. Biol 108, 355–368. [PubMed: 2416611]

Christophides GK, Zdobnov E, Barillas-Mury C, Birney E, Blandin S, Blass C, Brey PT, Collins FH, Danielli A, Dimopoulos G, Hetru C, Hoa NT, Hoffmann JA, Kanzok SM, Letunic I, Levashina EA, Loukeris TG, Lycett G, Meister S, Michel K, Moita LF, Müller HM, Osta MA, Paskewitz SM, Reichhart JM, Rzhetsky A, Troxler L, Vernick KD, Vlachou D, Volz J, von Mering C, Xu J, Zheng L, Bork P, Kafatos FC, 2002 Immunity-related genes and gene families in *Anopheles gambiae*. Science 298, 159–165. [PubMed: 12364793]

Davis CA, Riddell DC, Higgins MJ, Holden JJ, White BN, 1985 A gene family in *Drosophila melanogaster* coding for trypsin-like enzymes. Nucleic Acids Res. 13, 6605–6619. [PubMed: 2414727]

Dippel S, Oberhofer G, Kahnt J, Gerischer L, Opitz L, Schachtner J, Stanke M, Schütz S, Wimmer EA, Angeli S, 2014 Tissue-specific transcriptomics, chromosomal localization, and phylogeny of chemosensory and odorant binding proteins from the red flour beetle *Tribolium castaneum* reveal subgroup specificities for olfaction or more general functions. BMC Genomics 15, 1141. [PubMed: 25523483]

Edgar RC, 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797. [PubMed: 15034147]

Findlay GD, Sitnik JL, Wang W, Aquadro CF, Clark NL, Wolfner MF, 2014 Evolutionary rate covariation identifies new members of a protein network required for *Drosophila melanogaster* female post-mating responses. PLoS Genet. 10, e1004108. [PubMed: 24453993]

Foley JH, Conway EM, 2016 Cross Talk Pathways Between Coagulation and Inflammation. Circ Res 118, 1392–1408. [PubMed: 27126649]

Gorman MJ, Wang Y, Jiang H, Kanost MR, 2007 *Manduca sexta* hemolymph proteinase 21 activates prophenoloxidase-activating proteinase 3 in an insect innate immune response proteinase cascade. J. Biol. Chem 282, 11742–11749. [PubMed: 17317663]

Greenwood JM, Milutinović B, Peuß R, Behrens S, Esser D, Rosenstiel P, Schulenburg H, Kurtz J, 2017 Oral immune priming with *Bacillus thuringiensis* induces a shift in the gene expression of *Tribolium castaneum* larvae. BMC Genomics 18, 329. [PubMed: 28446171]

Gupta S, Wang Y, Jiang H, 2005 *Manduca sexta* prophenoloxidase (proPO) activation requires proPO-activating proteinase (PAP) and serine proteinase homologs (SPHs) simultaneously. Insect Biochem. Mol. Biol 35, 241–248. [PubMed: 15705503]

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, Macmanes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, Leduc RD, Friedman N, Regev A, 2013 De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature protocols 8, 1494–1512. [PubMed: 23845962]

He Y, Cao X, Zhang S, Rogers J, Hartson S, Jiang H, 2016 Changes in the plasma proteome of *Manduca sexta* larvae in relation to the transcriptome variations after an immune challenge: evidence for high molecular weight immune complex formation. Mol. Cell. Proteomics 15, 1176–1187. [PubMed: 26811355]

He Y, Wang Y, Hu Y, Jiang H, 2018 *Manduca sexta* hemolymph protease-2 (HP2) activated by HP14 generates prophenoloxidase-activating protease-2 (PAP2) in wandering larvae and pupae. Insect Biochem Mol Biol. 101, 57–65. [PubMed: 30098411]

He Y, Wang Y, Yang F, Jiang H, 2017 *Manduca sexta* hemolymph protease-1, activated by an unconventional non-proteolytic mechanism, mediates immune responses. Insect Biochem. Mol. Biol 84, 23–31. [PubMed: 28366787]

Issa N, Guillaumot N, Lauret E, Matt N, Schaeffer-Reiss C, Van Dorsselaer A, Reichhart JM, Veillard F, 2018 The circulating protease Persephone is an immune sensor for microbial proteolytic activities upstream of the *Drosophila* Toll pathway. Mol. Cell 69, 539–550. [PubMed: 29452635]

Jang IH, Chosa N, Kim SH, Nam HJ, Lemaitre B, Ochiai M, Kambris Z, Brun S, Hashimoto C, Ashida M, Brey PT, Lee WJ, 2006 A Spätzle-processing enzyme required for toll signaling activation in *Drosophila* innate immunity. Dev. Cell 10, 45–55. [PubMed: 16399077]

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S, 2014 InterProScan 5: genome-scale protein function classification. Bioinformatics 30, 1236–1240. [PubMed: 24451626]

Kambris Z, Brun S, Jang IH, Nam HJ, Romeo Y, Takahashi K, Lee WJ, Ueda R, Lemaitre B, 2006 *Drosophila* immunity: a large-scale in vivo RNAi screen identifies five serine proteases required for Toll activation. Curr. Biol 16, 808–813. [PubMed: 16631589]

Kanost MR, Jiang H, 2015 Clip-domain serine proteases as immune factors in insect hemolymph. Curr. Opin. Insect Sci 11, 47–55. [PubMed: 26688791]

Kawabata S, Muta T, 2010 Sadaaki Iwanaga: Discovery of the lipopolysaccharide- and β-1,3-D-glucan-mediated proteolytic cascade and unique proteins in invertebrate immunity. J. Biochem 147, 611–618. [PubMed: 20406733]

Kim K, Kim JH, Kim YH, Hong SE, Lee SH, 2018 Pathway profiles based on gene-set enrichment analysis in the honey bee *Apis mellifera* under brood rearing-suppressed conditions. Genomics 110, 43–49. [PubMed: 28803879]

Kumar S, Stecher G, Tamura K, 2016 MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol. Biol. Evol 33, 1870–1874. [PubMed: 27004904]

Kwon TH, Kim MS, Choi HW, Joo CH, Cho MY, Lee BL, 2000 A masquerade-like serine proteinase homologue is necessary for phenoloxidase activity in the coleopteran insect, *Holotrichia diomphalia* larvae. Eur. J. Biochem 267, 6188–6196. [PubMed: 11012672]

Langmead B, Salzberg SL, 2012 Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. [PubMed: 22388286]

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

Lin H, Xia X, Yu L, Vasseur L, Gurr GM, Yao F, Yang G, You M, 2015 Genome-wide identification and expression profiling of serine proteases and homologs in the diamondback moth, *Plutella xylostella* (L.). BMC Genomics 16, 1054. [PubMed: 26653876]

Manfredini F, Brown MJ, Vergoz V, Oldroyd BP, 2015 RNA-sequencing elucidates the regulation of behavioural transitions associated with the mating process in honey bee queens. BMC Genomics 16, 563. [PubMed: 26227994]

Mao W, Schuler MA, Berenbaum MR, 2017 Disruption of quercetin metabolism by fungicide affects energy production in honey bees (*Apis mellifera*). Proc. Natl. Acad. Sci. USA 114, 2538–2543. [PubMed: 28193870]

Moussian B, Roth S, 2005 Dorsoventral axis formation in the *Drosophila* embryo--shaping and transducing a morphogen gradient. Curr. Biol 15, R887–899. [PubMed: 16271864]

Mulinari S, Häcker U, Castillejo-López C, 2006 Expression and regulation of Spätzle-processing enzyme in *Drosophila*. FEBS Lett. 580, 5406–5410. [PubMed: 16996061]

Murugasu-Oei B, Balakrishnan R, Yang X, Chia W, Rodrigues V, 1996 Mutations in masquerade, a novel serine-protease-like molecule, affect axonal guidance and taste behavior in *Drosophila*. Mech. Dev 57, 91–101. [PubMed: 8817456]

Nam HJ, Jang IH, You H, Lee KA, Lee WJ, 2012 Genetic evidence of a redox-dependent systemic wound response via Hayan protease-phenoloxidase system in *Drosophila*. EMBO J. 31, 1253–1265. [PubMed: 22227521]

Park JW, Kim CH, Rui J, Park KH, Ryu KH, Chai JH, Hwang HO, Kurokawa K, Ha NC, Söderhäll I, Söderhäll K, Lee BL, 2010 Beetle immunity. In “Invertebrate Immunity” (Söderhäll K ed), Adv. Exp. Med. Biol 708, 163–180. [PubMed: 21528698]

Perona JJ, Craik CS 1995 Structural basis of substrate specificity in the serine proteases. Protein Sci. 4, 337–360. [PubMed: 7795518]

Petersen TN, Brunak S, von Heijne G, Nielsen H, 2011 SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat. Methods 8, 785–786. [PubMed: 21959131]

Puente XS, Sanchez LM, Overall CM, Lopez-Otin C, 2003 Human and mouse proteases: a comparative genomic approach. Nat. Rev. Genet 4, 544–558. [PubMed: 12838346]

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP, 2012 MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol 61, 539–542. [PubMed: 22357727]

Ross J, Jiang H, Kanost MR, Wang Y, 2003 Serine proteases and their homologs in the *Drosophila melanogaster* genome: an initial analysis of sequence conservation and phylogenetic relationship. Gene 304, 117–131. [PubMed: 12568721]

Rukmini V, Reddy CY, Venkateswerlu G, 2000 *Bacillus thuringiensis* crystal delta-endotoxin: role of proteases in the conversion of protoxin to toxin. Biochimie 82, 109–116. [PubMed: 10727765]

Schnakenberg SL, Matias WR, Siegal ML, 2011 Sperm-storage defects and live birth in *Drosophila females* lacking spermathecal secretory cells. PLoS Biol. 9, e1001192. [PubMed: 22087073]

Schrag LG, Herrera AI, Cao X, Prakash O, Jiang H, 2017 Structure and function of stress responsive peptides in insects In: Srivastava VP (Ed.), Peptide-based Drug Discovery: Challenges and New Therapeutics. Royal Society of Chemistry, London, UK, pp. 438–451.

Shah PK, Tripathi LP, Jensen LJ, Gahnim M, Mason C, Furlong EE, Rodrigues V, White KP, Bork P, Sowdhamini R, 2008 Enhanced function annotations for *Drosophila* serine proteases: a case study for systematic annotation of multi-member gene families. Gene 407, 199–215. [PubMed: 17996400]

Shen HB, Chou KC, 2007 Signal-3L: A 3-layer approach for predicting signal peptides. Biochem. Biophys. Res. Commun 363, 297–303. [PubMed: 17880924]

Stappert D, Frey N, von Levetzow C, Roth S, 2016 Genome-wide identification of *Tribolium* dorsoventral patterning genes. Development 143, 2443–2454. [PubMed: 27287803]

Tang H, Kambris Z, Lemaitre B, Hashimoto C, 2006 Two proteases defining a melanization cascade in the immune system of *Drosophila*. J. Biol. Chem 281, 28097–28104. [PubMed: 16861233]

Veillard F, Troxler L, Reichhart JM, 2016 *Drosophila melanogaster* clip-domain serine proteases: structure, function and regulation. Biochimie 122, 255–269. [PubMed: 26453810]

Wang Y, Jiang H, 2004 Prophenoloxidase (proPO) activation in *Manduca sexta*: an analysis of molecular interactions among proPO, proPO-activating proteinase-3, and a cofactor. Insect Biochem. Mol. Biol 34, 731–742. [PubMed: 15262278]

Wang Y, Jiang H, 2007 Reconstitution of a branch of the *Manduca sexta* prophenoloxidase activation cascade in vitro: Snake-like hemolymph proteinase 21 (HP21) cleaved by HP14 activates prophenoloxidase-activating proteinase-2 precursor. Insect Biochem. Mol. Biol 37, 1015–1025. [PubMed: 17785189]

Wang Y, Jiang H, 2008 A positive feedback mechanism in the *Manduca sexta* prophenoloxidase activation. Insect Biochem. Mol. Biol 38, 763–769. [PubMed: 18625399]

Wang Y, Lu Z, Jiang H, 2014 *Manduca sexta* proprophenoloxidase activating proteinase-3 (PAP3) stimulates melanization by activating proPAP3, proSPHs, and pro-POs. Insect Biochem. Mol. Biol 50, 82–91. [PubMed: 24768974]

Waterhouse RM, Kriventseva EV, Meister S, Xi Z, Alvarez KS, Bartholomay LC, Barillas-Mury C, Bian G, Blandin S, Christensen BM, Dong Y, Jiang H, Kanost MR, Koutsos AC, Levashina EA, Li J, Ligoxygakis P, Maccallum RM, Mayhew GF, Mendes A, Michel K, Osta MA, Paskewitz S, Shin SW, Vlachou D, Wang L, Wei W, Zheng L, Zou Z, Severson DW, Raikhel AS, Kafatos FC, Dimopoulos G, Zdobnov EM, Christophides GK, 2007 Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. Science 316, 1738–1743. [PubMed: 17588928]

Wojtukiewicz MZ, Hempel D, Sierko E, Tucker SC, Honn KV, 2016 Thrombin-unique coagulation system protein with multifaceted impacts on cancer and metastasis. Cancer Metastasis Rev. 35, 213–233. [PubMed: 27189210]

Yu XQ, Jiang H, Wang Y, Kanost MR, 2003 Nonproteolytic serine proteinase homologs are involved in prophenoloxidase activation in the tobacco hornworm, *Manduca sexta*. Insect Biochem. Mol. Biol 33, 197–208. [PubMed: 12535678]

Zhao P, Lu Z, Strand MR, Jiang H, 2011 Antiviral, anti-parasitic, and cytotoxic effects of 5,6-dihydroxyindole (DHI), a reactive compound generated by phenoloxidase during insect immune response. Insect Biochem Mol Biol. 41, 645–652. [PubMed: 21554953]

Zhao P, Wang GH, Dong ZM, Duan J, Xu PZ, Cheng TC, Xiang ZH, Xia QY, 2010 Genome-wide identification and expression analysis of serine proteases and homologs in the silkworm *Bombyx mori*. BMC Genomics 11, 405. [PubMed: 20576138]

Zhu-Salzman K, Zeng R, 2015 Insect response to plant defensive protease inhibitors. Ann. Rev. Entomol 60, 233–252. [PubMed: 25341101]

Zou Z, Lopez DL, Kanost MR, Evans JD, Jiang H, 2006 Comparative analysis of serine protease-related genes in the honeybee genome: possible involvement in embryonic development and innate immunity. Insect Mol. Biol 15, 603–614. [PubMed: 17069636]

Zou Z, Evans J, Lu Z, Zhao P, Williams M, Sumathipala N, Hetru C, Hultmark D, Jiang H, 2007 Comparative genome analysis of the *Tribolium* immune system. Genome Biol. 8, R177. [PubMed: 17727709]

Highlights:

- Extensively updated knowledge on serine protease-related proteins in the fruit fly, honey bee, and flour beetle;
- Inner links detected among phylogenetic relationships, genomic locations and expression profiles for 99 of the *Drosophila* genes;
- Putative orthologous relationships revealed among S1A members from the five holometabolous insects;
- Physiological roles predicted on the basis of orthologous relationships and functional data.

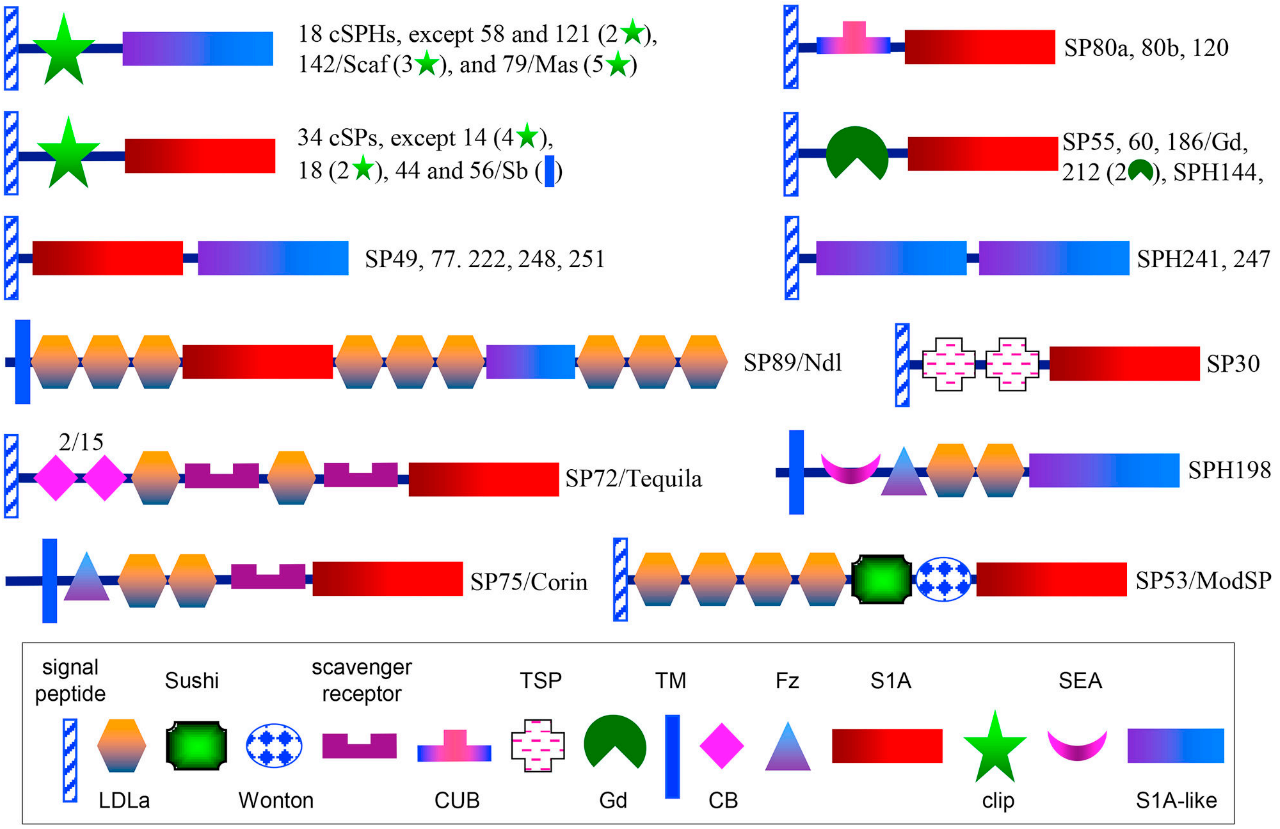


Fig. 1.

Domain organization of 66 multi-domain SPs and SPHs in *D. melanogaster*. Signal peptide and other structural elements (see symbols in inset) were predicted. The schematic diagrams are not drawn to scale.

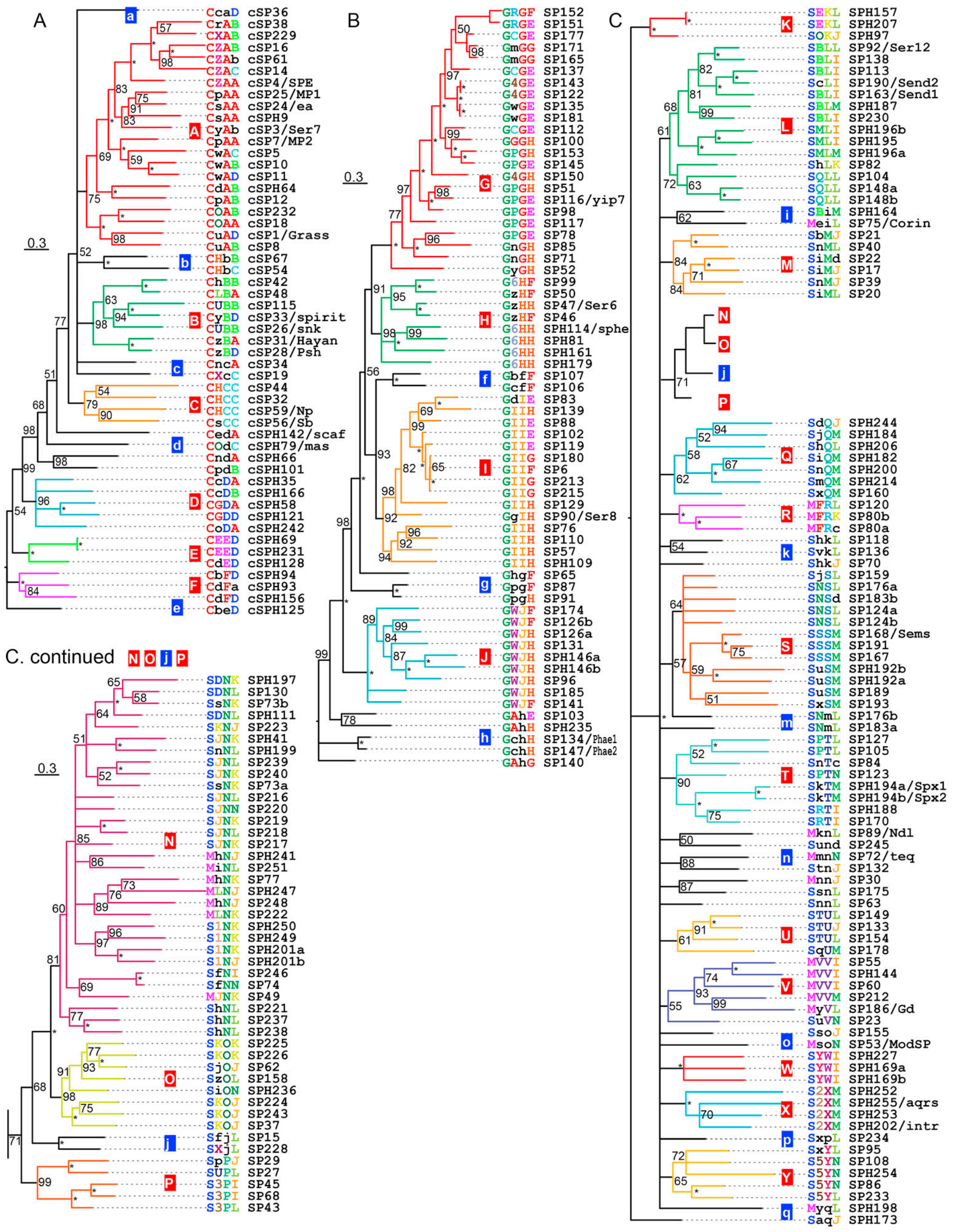


Fig. 2.

Phylogenetic relationships of the 52 CLIPs (**A**), 65 GP(H)s (**B**), and other 140 SP(H)s (including 21 multi-domain proteins lacking clip domain) (**C**) in *D. melanogaster*. Complete sequences of the proteins in each group were aligned and a phylogenetic tree was constructed using MrBayes v3.2.6. Probability values for branches are indicated near the branching points, with "*" representing 100%. Those on red background with the capital letters (A–Y) represent branches (probability ≥55%) on the tree, and other branches with lower probabilities are assigned with lower case letters (a–q) on blue background. **G**roup name (C for clip, G for gut, M for other multi-domain, S for other single domain SP(H)), **c**hromosomal location (A–Z, 1–6, and a–z in Fig. 3), **t**ree position (A–Y and a–q in this figure), and **e**xpression profile (A–N and a–D in Fig. 4) are used to generate G-C-T-E identification code for each SP-related gene. These IDs, in various colors depending on their categories, are listed before the corresponding systemic names.

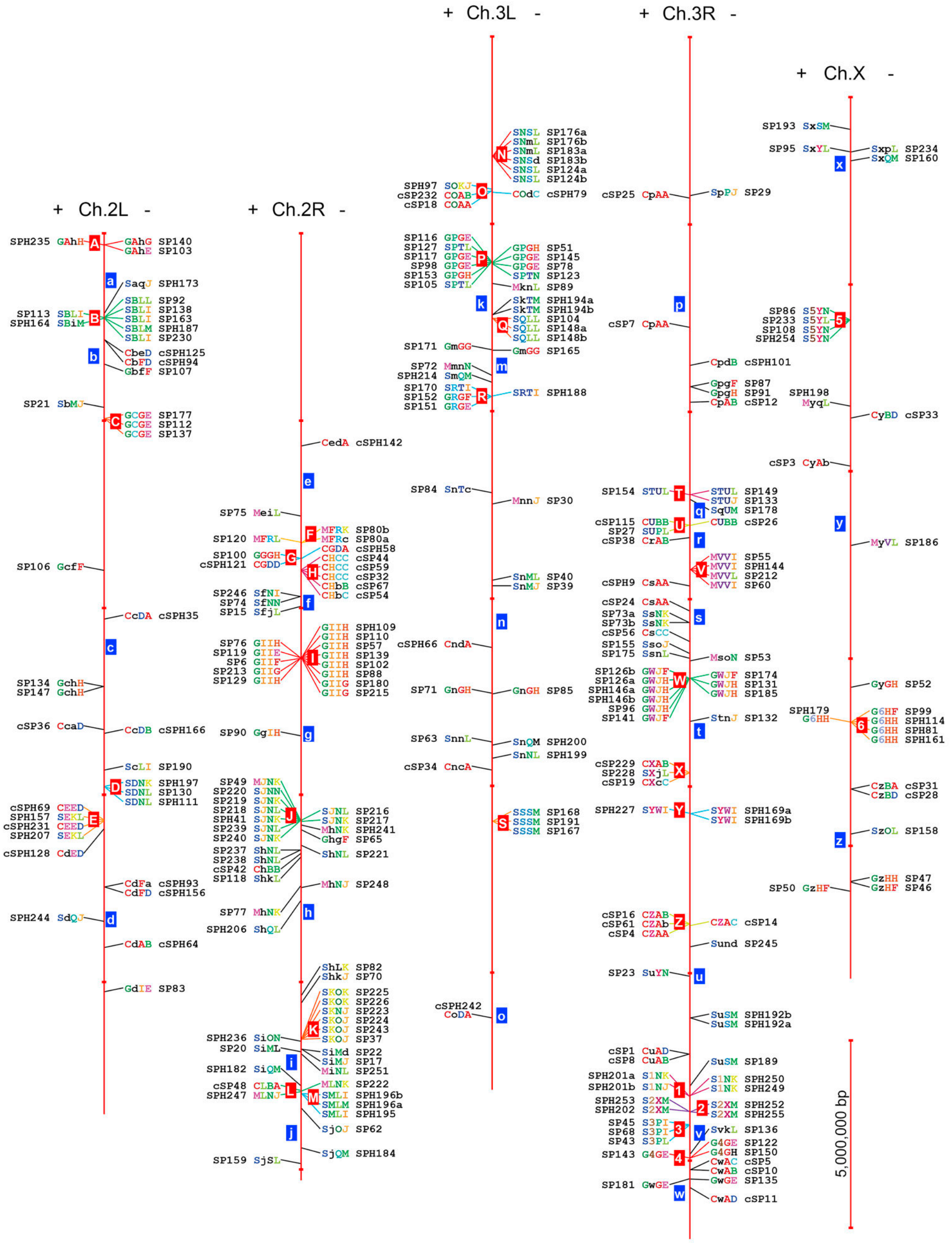
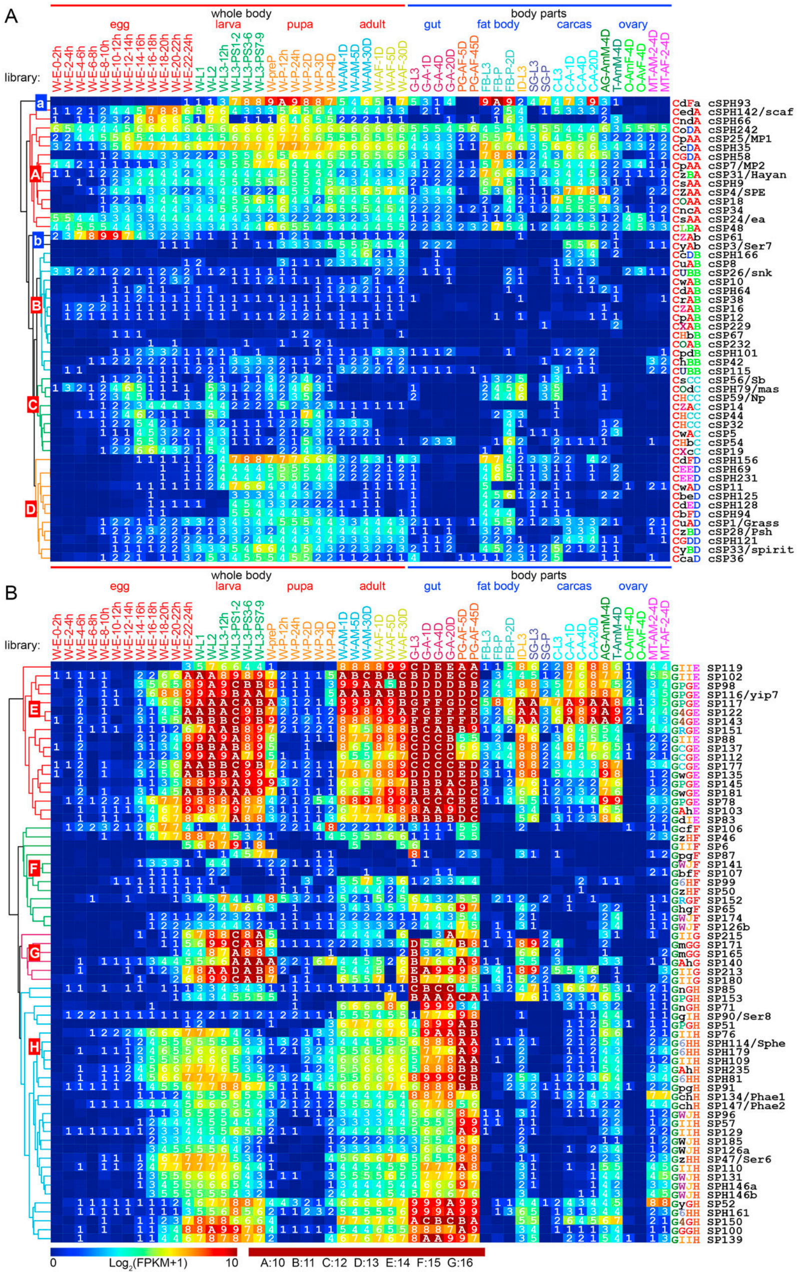


Fig. 3.

Chromosomal locations of the 257 genes coding for *D. melanogaster* SP-related proteins. As indicated by the scale bar, positions of the genes are plotted in proportion on chromosomes, with “+” and “−” indicating positive and negative strands on the left and right, respectively. The G-C-T-E identification code for each SP-related gene is defined in the legend to Fig. 2, indicated along with the systemic name, and linked to its location by a straight line. Adjacent genes with high sequence similarities are grouped by lines in the same color and marked by a series of location IDs (A–Z and 1–6) on red background for different chromosomal segments. Regions in between are labeled with IDs on blue background (a–z).


A
whole body
body parts
egg
larva
pupa
adult
gut
fat body
carcas
ovary
library:
W-E-0-2h
W-E-2-4h
W-E-4-6h
W-E-6-8h
W-E-8-10h
W-E-10-12h
W-E-12-14h
W-E-14-16h
W-E-16-18h
W-E-18-20h
W-E-20-22h
W-E-22-24h
W-L1
W-L2
W-L3-12h
W-L3-PS1-2
W-L3-PS3-6
W-L3-PS7-9
W-preP
W-P-12h
W-P-24h
W-P-2D
W-P-3D
W-P-4D
W-AM-1D
W-AM-5D
W-AM-30D
W-AF-1D
W-AF-5D
W-AF-30D
G-L3
G-A-1D
G-A-4D
G-A-20D
PG-AF-5D
PG-AF-45D
FB-L3
FB-P
FB-P-2D
ID-L3
SG-L3
SG-P
C-L3
C-A-1D
C-A-4D
C-A-20D
AG-AmM-4D
T-AmM-4D
O-AmF-4D
O-AvF-4D
MT-AM-2-4D
MT-AF-2-4D
a
A
b
B
C
D
CdFa cSPH93
CedA cSPH142/scaf
CndA cSPH66
CoDA cSPH242
CpAA cSP25/MP1
CcDA cSPH35
CGDA cSPH58
CpAA cSP7/MP2
CzBA cSP31/Hayan
CsAA cSPH9
CZAA cSP4/SPE
COAA cSP18
CncA cSP34
CsAA cSP24/ea
CLBA cSP48
CZAb cSP61
CyAb cSP3/Ser7
CcDB cSPH166
CuAB cSP8
CUBB cSP26/snk
CwAB cSP10
CdAB cSPH64
CrAB cSP38
CZAB cSP16
CpAB cSP12
CXAB cSP229
CHbB cSP67
COAB cSP232
CpdB cSPH101
ChBB cSP42
CUBB cSP115
CsCC cSP56/Sb
COdC cSPH79/mas
CHCC cSP59/Np
CZAC cSP14
CHCC cSP44
CHCC cSP32
CwAC cSP5
CHbC cSP54
CXcC cSP19
CdFD cSPH156
CEED cSPH69
CEED cSPH231
CwAD cSP11
CbeD cSPH125
CdED cSPH128
CbFD cSPH94
CuAD cSP1/Grass
CzBD cSP28/Psh
CGDD cSPH121
CyBD cSP33/spirit
CcaD cSP36
B
E
F
G
H
GIIE SP119
GIIE SP102
GPGE SP98
GPGE SP116/yip7
GPGE SP117
G4GE SP122
G4GE SP143
GRGE SP151
GIIE SP88
GCGE SP137
GCGE SP112
GCGE SP177
GwGE SP135
GPGE SP145
GwGE SP181
GPGE SP78
GAhE SP103
GdIE SP83
GcfF SP106
GzHF SP46
GIIF SP6
GpgF SP87
GWJF SP141
GbfF SP107
G6HF SP99
GzHF SP50
GRGF SP152
GhgF SP65
GWJF SP174
GWJF SP126b
GIIG SP215
GmGG SP171
GmGG SP165
GAhG SP140
GIIG SP213
GIIG SP180
GnGH SP85
GPGH SP153
GnGH SP71
GgIH SP90/Ser8
GPGH SP51
GIIH SP76
G6HH SPH114/Sphe
G6HH SPH179
GIIH SPH109
GAhH SPH235
G6HH SPH81
GpgH SP91
GchH SP134/Phae1
GchH SP147/Phae2
GWJH SP96
GIIH SP57
GIIH SP129
GWJH SP185
GWJH SP126a
GzHH SP47/Ser6
GIIH SP110
GWJH SP131
GWJH SPH146a
GWJH SPH146b
GyGH SP52
G6HH SPH161
G4GH SP150
GGGH SP100
GIIH SP139
0
Log2(FPKM+1)
10
A:10 B:11 C:12 D:13 E:14 F:15 G:16

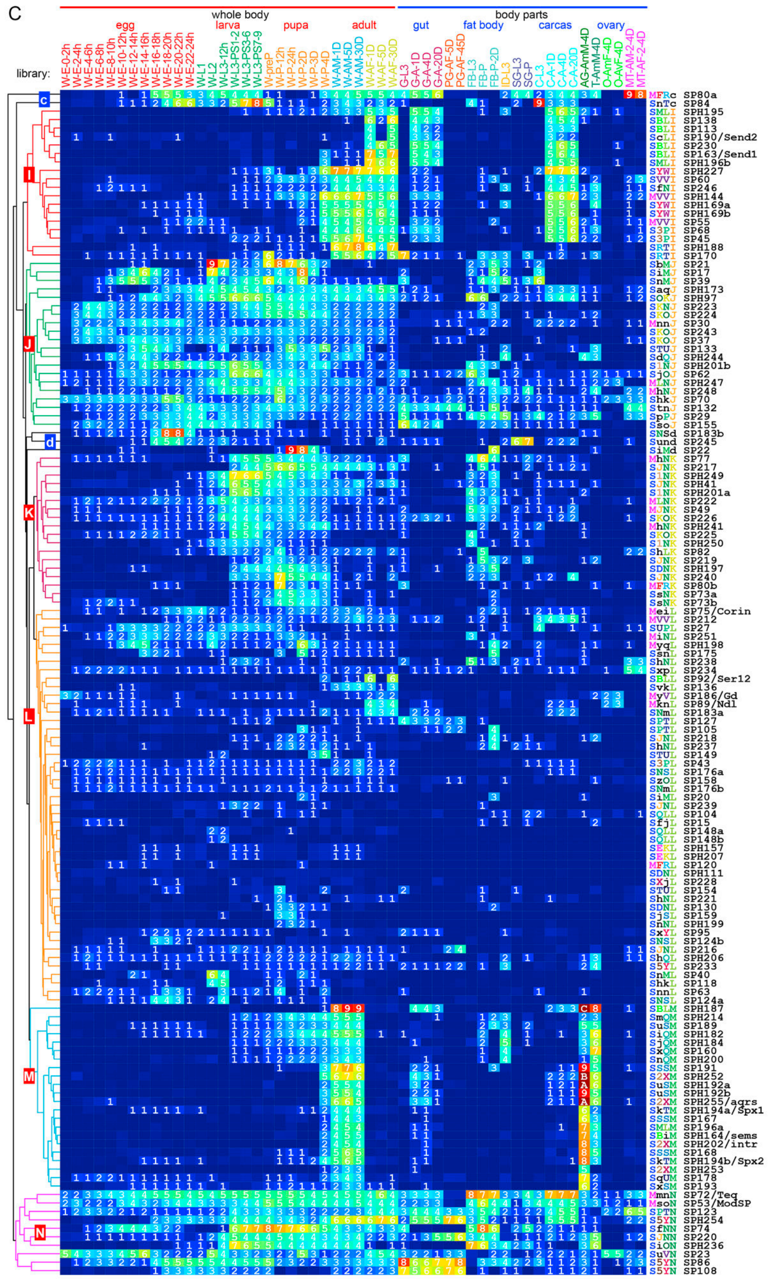
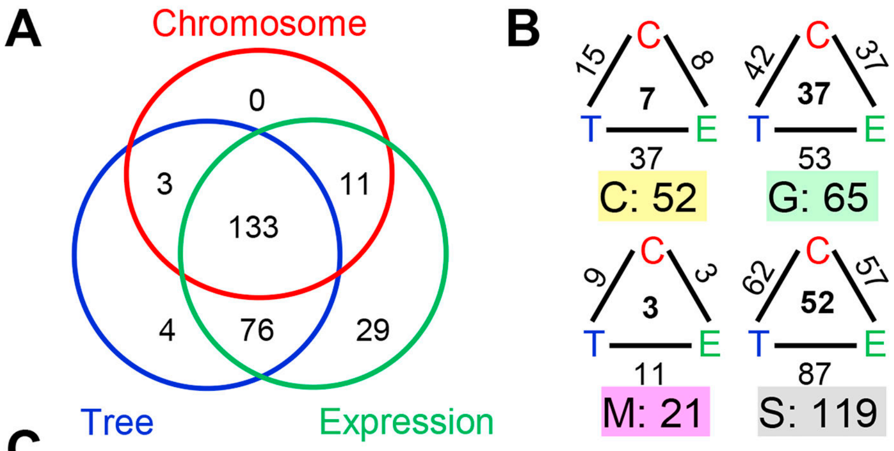


Fig. 4.

Transcript profiles of the 52 CLIPs (**A**), 65 GPs/GPHs (**B**) and 140 SPs/SPHs (including multi-domain SP-related ones) (**C**) in *D. melanogaster*. The mRNA levels in 52 kinds of tissue samples, as represented by $\log_2$(FPKM+1) values, are shown in the hierarchically clustered gradient heatmap from blue (0) to maroon (≥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, … 15.50–16.49 are labeled in the color blocks as 0, 1, 2 … 9, A, B, … G, respectively. Groups of genes with similar expression patterns are shown in different colors. The branches on red background with the capital letters A–N represent reliable assemblies, while other branches are assigned with lower case letters a–D on blue background. Systemic names are listed on the right along with the G-C-

T-E IDs (see definition in the legend to Fig. 2). Briefly, in the first position, C (in red) for CLIPs, G (in green) for gut SPs/SPHs, M (in pink) for other multi-domain SP-like proteins, and S (in blue) for other single domain SP(H)s. In the 2nd to 4th positions, chromosomal location (Fig. 3), tree position (Fig. 2), and expression assembly (Fig. 4) are shown in various colors as defined in the corresponding figures. The libraries names are abbreviated using: W, whole insect; E, embryo; L, larval; preP, prepupal; P, pupal; A, adult; M, male; F, female; h, hour; D, day; G, gut; FB, fat body; ID, imaginal discs, SG, salivary glands; C, carcass; AG-AmM, accessary glands of adult mated male; T, testes; O, ovaries of adult mated female; AvF, adult virgin female; MT, Malpighian tubules.



C

	C-T-E	Names
C	EED	cSPH69, 231
	HCC	cSP32, 44, 59
	UBB	cSP26, 115
G	CGE	SP112, 137, 177
	IIE	SP88, 102, 119
	IIG	SP180, 213, 215
	IIH	SP57, 76, 110, 129 139, SPH109
	PGE	SP78, 98, 116, 117, 145
	PGH	SP51, 153
	WJF	SP126b, 141, 174
	WJH	SP96, 126a, 131, 185, SPH146a, 146b
	4GE	SP122, 143
	6HH	SPH81, 114, 161, 179
M	VVI	SP55, 60, SPH144
S	BLI	SP113, 138, 163, 230
	DNL	SP130, SPH111
	EKL	SPH157, 207
	JNK	SP217, 219, 240, SPH41
	JNL	SP216, 218, 239
	KOJ	SP37, 224, 243,
	KOK	SP225, 226
	MLI	SPH195, 196b
	NSL	SP124a, 124b, 176a
	PTL	SP105, 127
	QLL	SP104, 148a, 148b
	RTI	SP170, SPH188
	SSM	SP167, 168, 191
	TUL	SP149, 154
	YWI	SPH169a, 169b, 227
	1NK	SPH201a, 249, 250
	2XM	SPH202, 252, 253, 255
	3PI	SP45, 68
	5YN	SP86, 108, SPH254

Fig. 5.

Correlations of the *D. melanogaster* SP-related genes in clusters on chromosome (C), branches of phylogenetic tree (T), and assemblies of expression (E). (**A**) Venn diagram of numbers the genes located in clusters, branches, and assemblies of the entire S1A SP/SPH family. (**B**) Two- and three-way matches of the genes in the C(LIP), G(ut), M(ulti-domain), and S(ingle domain) groups. For each CTE triangle, numbers of 2- (C-T, T-E, or E-C) and 3-(C-T-E) way matched genes are indicated on the sides and center. Each match has at least two genes. For instance, in the G group, 42 C-T, 53 T-E, 37 E-C, and 37 C-T-E matches are found, involving a total of 65 SPs/SPHs. (**C**) A list of SP/SPH names with C-T-E matches in the C, G, M and S groups. Among the 52 CLIPs, 7 belong to three 3-way matches:

cSPH69-231, cSP32-44-59, and cSP26-115. For the G group, there are ten C-T-E matches with 2–6 members in each and 37 proteins in total.

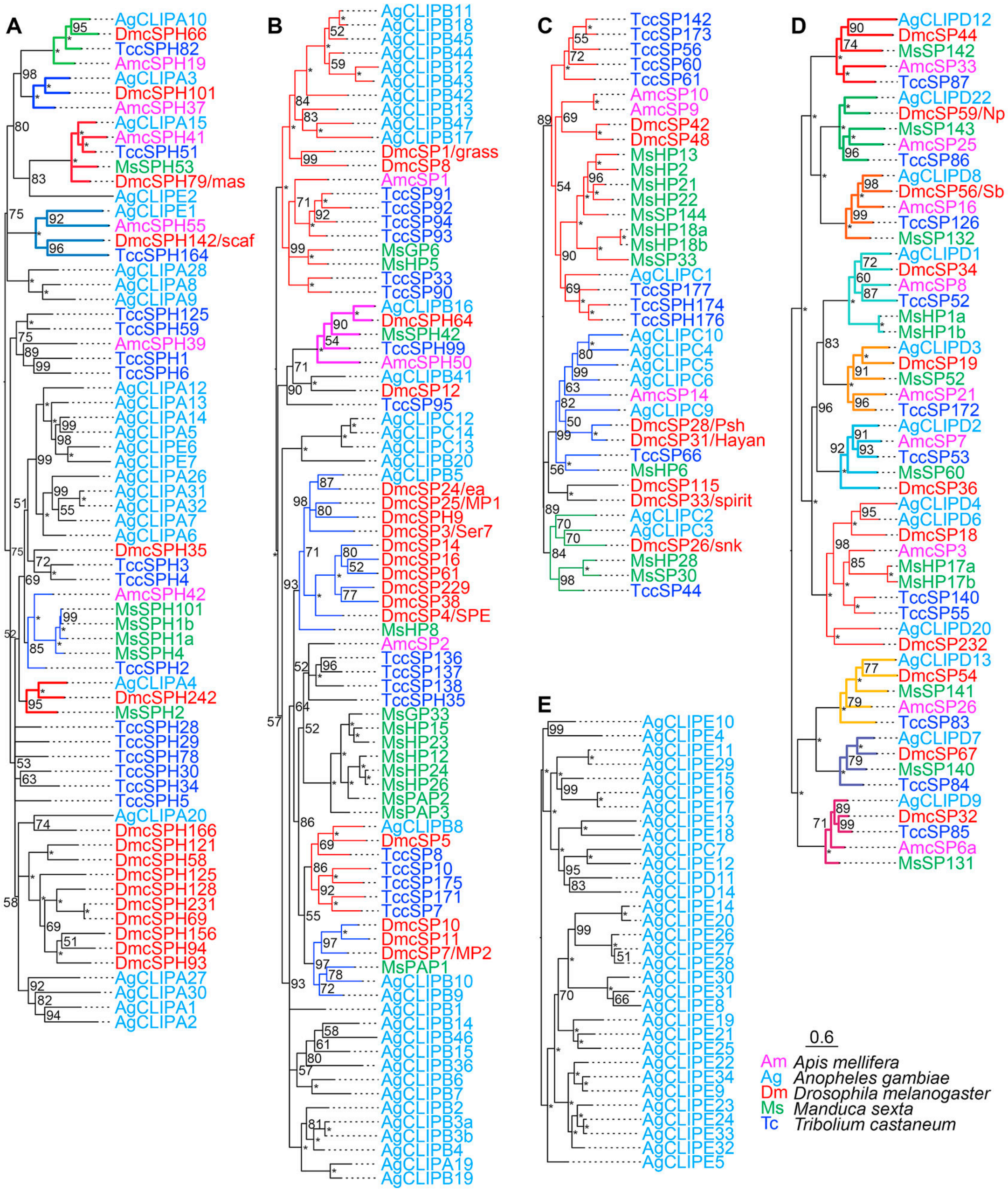


Fig. 6.

Phylogenetic trees of the CLIPAs (**A**), CLIPBs (**B**), CLIPCs (**C**), CLIPDs (**D**), and CLIPEs (**E**) in the five insects. Based on the initial analysis of 247 CLIPA–D's (Fig. S1), entire protein sequences in each subgroup were aligned for building a phylogenetic tree using MrBayes v3.2.6. Probability values are indicated near the branching points, with "*" representing 100 and colored branches representing various sets of potential orthologous genes (probability >80, 3–5 species). The bold branches represent 1:1:1, 1:1:1:1 or 1:1:1:1:1 orthology, except for the set containing closely linked MsHP1a and MsHP1b. As shown in the inset, the protein names are in different colors.

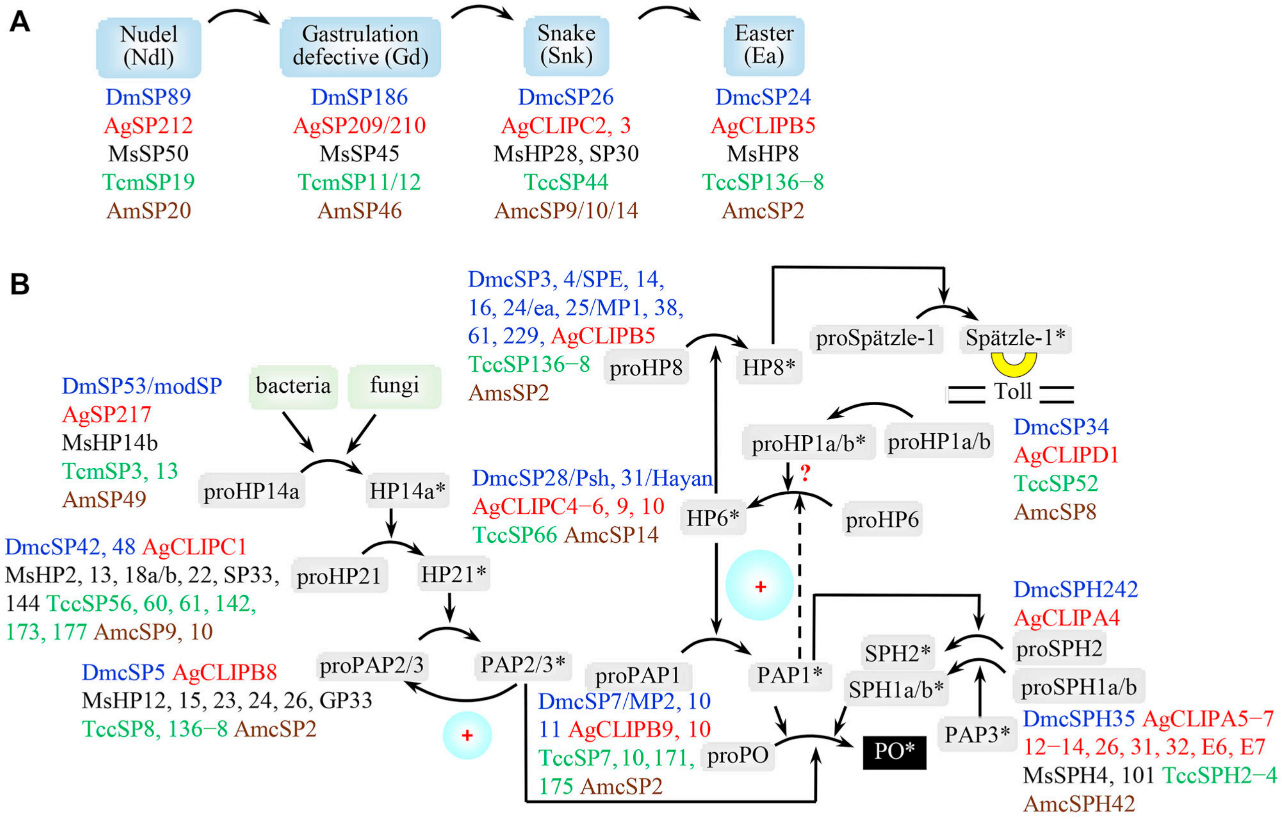


Fig. 7.

Predicted functions of the CLIPs and other multi-domain SPs and SPHs in the insect SP-SPH pathways. (**A**) The SP cascade that establishes the dorsal-ventral axis of *D. melanogaster* embryo is used as a template to identify ortholog sets for proposing similar pathways in the other insects. When an ortholog is not identified in the phylogenetic analysis, its closest homolog(s) are listed to assist functional exploration: for example, *A. mellifera* cSP9/10/14 under cSP26/Snk and *T. castanusm* cSP136–8 under cSP24/Ea. (**B**) Members of the SP-SPH network that mediates proPO and proSpätzle-1 activation in *M. sexta* are presented along with their orthologs or close homologs, as revealed by the serial phylogenetic analyses and domain structure comparison (Fig. 6, Table 2). The protein names are in blue (Dm), red (Ag), black (Ms), green (Tc), and brown (Am) fonts. Circles with a "+" sign represent positive feedback mechanisms, one being auto-activation of proPAP3 by PAP3 (Wang et al., 2014) and the other being indirect activation proHP6 by PAP1 (dashed arrow) and direct activation of proPAP1 by HP6 (Wang and Jiang, 2008). "?" indicates that the step (*i.e.* cleavage activation of proHP6 by uncut but active proHP1) is partially established (He et al., 2017).

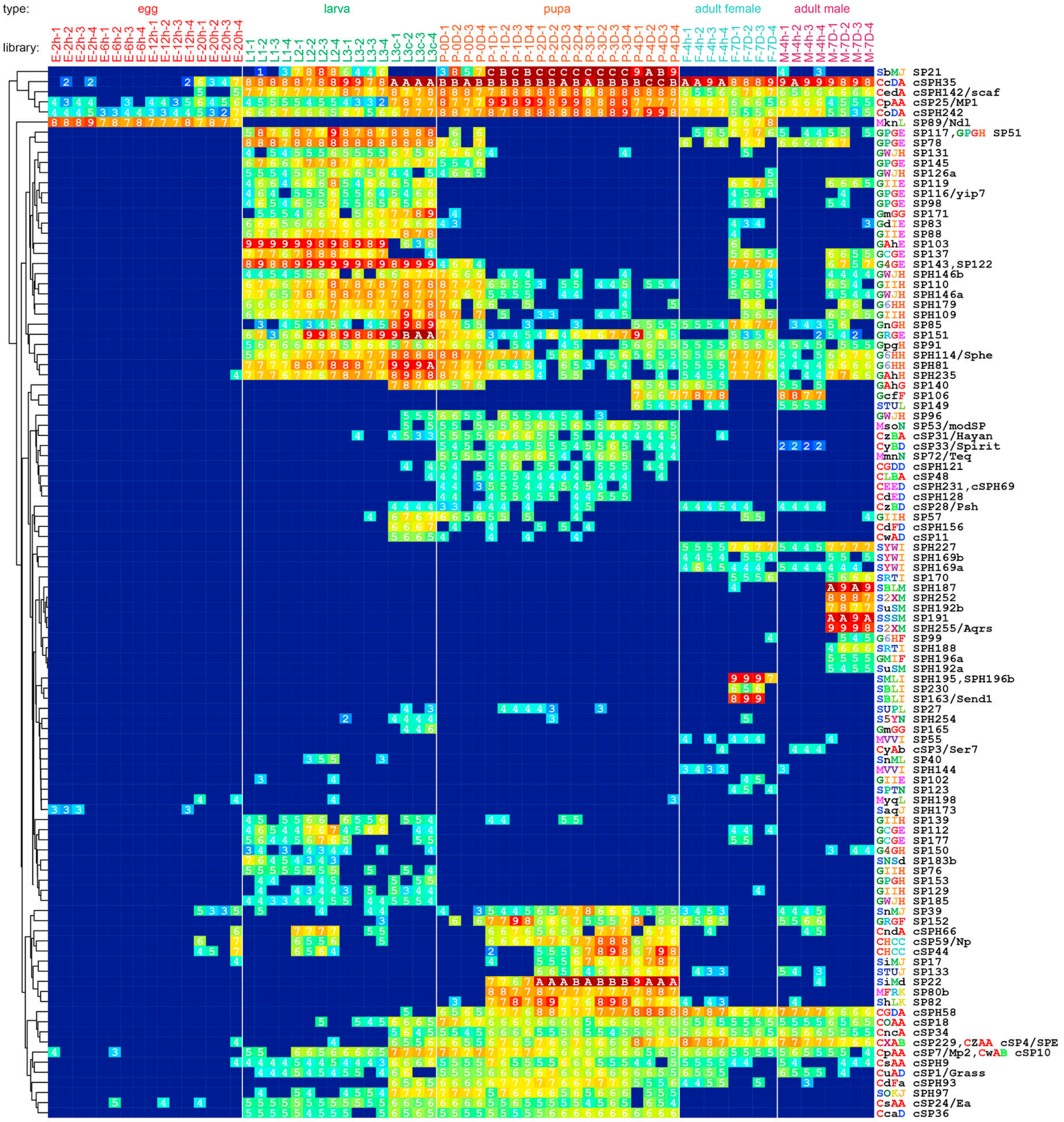


Fig. 8.

Abundances of the 110 SP-related proteins in *D. melanogaster* at various developmental stages. Relative protein levels in the 16 egg, 16 larval, 20 pupal, 8 female adult, and 8 male adult samples at different time points, as represented by $\log_2(\text{LFQ}/5\times10^6 + 1)$ values, are shown in the hierarchically clustered gradient heat map from blue (0) to maroon (≥10). The values of 0–0.49, 0.50–1.49, 1.50–2.49, … 8.50–9.49, 9.50–10.49, 10.50–11.49, and 11.50–12.49 are labeled in the color blocks as 0, 1, 2 … 9, A, B, and C, respectively. Proteins identified in two or less of the 68 samples are eliminated. Due to high sequence identity, no distinction can be made in SP51-117, SP122-143, SPH195-196b, cSP7-10, cSP4-229, and cSPH69-231 pairs. The datasets or library names are abbreviated using: E, embryo; L, larval; P, pupal; M, male; F, female; L3c, crawling third instar larva; h, hour; D, day.

Table 1.

Names and key features of the 257 serine protease-like proteins in *D. melanogaster*

name	G.C.T.E	activation cleavage site	specificity	domain	name	G.C.T.E	activation cleavage site	specificity	domain	name	G.C.T.E	activation cleavage site	specificity	domain
cSP1/Grass	CuAD	LSQR*VSNG	T(DGG)	cP	SP126a	GWJH	QSGR*IKGG	C(GGD)	P	SPH169a	SYWI	DEDQ*MLNM	na	H
cSP3/Ser7	CyAb	FSFR*LVGG	T(DGG)	cP	SP126b	GWJH	KDQR*IIGG	C(GGD)	P	SPH169b	SYWI	LDEK*LLYT	na	H
cSP4/SPE	CZAA	FADR*IFGG	T(DGG)	cP	SP129/ιTry	GIIH	ATGR*IIGG	T(DGG)	P	SP170	SRTI	PKDI*IVNG	C(SGS)	P
cSP5	CwAC	VTNR*IYGG	T(DGG)	cP	SP131	GWJH	FQNR*VING	C(GGD)	P	SPH173	SaqJ	QIRR*ETYG	na	H
cSP7/MP2	CpAA	FSNK*VYNG	T(DGG)	cP	SP134/Phae1	GchH	PEGR*VVGG	C(SGS)	P	SP175	SsnL	RGPR*IIAG	C(GGG)	P
cSP8	CuAB	GNPK*VSGG	C(HGG)	cP	SP135/Jon99Fii	GwGE	IQGR*ITNG	C(SVA)	P	SP176a	SNSL	ANSR*IVGG	T(DGG)	P
cSPH9	CsAA	VRWQ*RSND	na	cP	SP137/Jon25Bi	GCGE	INGR*IVNG	E(SVS)	P	SP176b	SNmL	SKTR*IVGG	T(DGG)	P
cSP10	CwAB	IRNR*IYDG	T(DGG)	cP	SP139/θTry	GIIH	REGR*IVGG	T(DGG)	P	SP178	SqUM	STTK*VISF	C(GGG)	P
cSP11	CwAD	AYNQ*ITKG	T(DGG)	cP	SP140	GAhG	PEGR*IING	C(GGS)	P	SPH182	SiQM	SSPK*FHGD	na	H
cSP12	CpAB	VESR*LLGG	T(DGG)	cP	SP141	GWJF	TKNR*IVGG	C(GGD)	P	SP183a	SNmL	RQGK*IFGG	T(DGG)	P
cSP14	CZAC	PLYR*MAYG	T(DGG)	4cP	SP143/Jon99Ciii	G4GE	IQGR*ITNG	C(SGA)	P	SP183b	SNSd	IQPR*IIGG	T(DGG)	P
cSP16	CZAB	PAYR*IFGG	T(DGG)	cP	SP145/Jon65Ai	GPGE	IEGR*ITMG	C(GGK)	P	SPH184	SjQM	FEML*ISGG	na	H
cSP18	COAA	YFKK*IVGG	T(DGG)	2cP	SPH146a	GWJH	PQGR*VIGG	na	H	SPH187	SBLM	ADFK*SIGI	na	H
cSP19	CXcC	STGR*IVGG	T(DGG)	cP	SPH146b	GWJH	FSSR*IVGG	na	H	SPH188	SRTI	PDHI*ITNG	na	H
cSP24/Easter	CsAA	LSNR*IYGG	T(DGG)	cP	SP147/Phae2	GchH	PEGR*VVGG	C(SGS)	P	SP189	SuSM	FHPR*IYNG	C(TGT)	P
cSP25/MP1	CpAA	FGDR*VVGG	T(DGG)	cP	SP150/Jon99Ci	G4GH	IEGR*ITNG	E(AVA)	P	SP190/Send2	G4GH	PEER*IIGG	C(AGG)	P
cSP26/Snk	CUBB	SVPL*IVGG	T(DGG)	cP	SP151	GRGE	AEGR*IVNG	E(SAS)	P	SP191	SSSM	HETR*VIGG	T(DGG)	P
cSP28/Psh	CzBD	LVIH*IVGG	T(DGG)	cP	SP152	GRGF	AEGR*IVNG	C(SGS)	P	SPH192a	SuSM	KFRR*VWGG	na	H
cSP31/Hayan	CzBA	LTVH*ILDG	T(DGG)	cP	SP153/Jon65Aii	GPGH	INGR*ITNG	C(GGK)	P	SPH192b	SuSM	PPVR*TLNK	na	H
cSP32	CHCC	RSNR*IVGG	T(DGG)	cP	SPH161	G6HH	PQGR*IAGG	na	H	SP193	SxSM	WGDA*LHRG	C(GGR)	P
cSP33/Spirit	CyBD	FFVS*VVGG	T(DGG)	cP	SP165/Jon66Cii	GmGG	MQGR*ITNG	C(SGA)	P	SPH194a/Spx1	SkTM	LSPR*IAGG	na	H
cSP34	CncA	QFPR*LTGG	T(DGG)	cP	SP171/Jon66Ci	GmGG	IEGR*ITNG	E(GVA)	P	SPH194b/Spx2	SkTM	LSPR*ITGG	na	H
cSPH35	CcDA	VGFK*ITGA	T(DGG)	cP	SP174	GWJF	LEHF*IVGG	E(GFD)	P	SPH195	SMLI	GDQR*IING	na	H
cSP36	CcaD	DQER*IVGG	T(DGG)	cP	SP177/Jon25Biii	GCGE	IEGR*ITNG	C(SVA)	P	SPH196a	SMIF	CNRT*TLGG	na	H
cSP38	CrAB	SRRK*PTKG	T(DGG)	cP	SPH179	G6HH	QSSR*LPAE	na	H	SPH196b	SMLI	AQSR*IIGG	na	H
cSP42	ChBB	TTPF*IVGG	T(DGG)	cP	SP180/γTry	GIIG	LDGR*IVGG	T(DGG)	P	SPH197	SDNK	RCGL*LTNG	na	H
cSP44	CHCC	KSGR*IVGG	T(DGG)	TMcP	SP181/Jon99Fi	GwGE	IQGR*ITNG	C(SVA)	P	SPH199	SnNL	LSPD*IVGP	na	H
cSP48	CLBA	STPF*IVGG	T(DGG)	cP	SP185	GWJH	LDNR*IVGG	C(GGD)	P	SPH200	SnQM	RAKR*LSSP	na	H
cSP54	CHbC	AQRR*IVGG	T(DGG)	cP	SP213/δTry	GIIG	LDGR*IVGG	C(GGD)	P	SPH201a	S1NK	LSND*IIFS	na	H

name	G.C.T.E	activation cleavage site	specificity	domain	name	G.C.T.E	activation cleavage site	specificity	domain	name	G.C.T.E	activation cleavage site	specificity	domain
cSP56/Sb	CsCC	PETR*IVGG	T(DGG)	cP	SP215	GIIG	LDGR*IVGG	C(GGD)	P	SPH201b	S1NJ	PHQD*VFKE	na	H
cSPH58	CGDA	DNDKFPYS	na	2cH	SPH235	GAhH	IQPL*IIDG	na	H	SPH202/Intr	S2XM	IETL*LTDG	na	H
cSP59/Np	CHCC	PEPR*IVGG	T(DGG)	cP	SP15	SfjL	PSSY*IVGG	T(DGG)	P	SPH206	ShQL	YHQN*VVSI	na	H
cSP61	CZAb	PVFR*DRGA	T(DGG)	cP	SP17	SiMJ	IQKR*IVGG	T(DGG)	P	SPH207	SEKL	VQFN*VTEG	na	H
cSPH64	CdAB	VQGH*FYKG	na	cH	SP20	SiML	TLYK*IVGG	T(DGG)	P	SPH214	SmQM	RVKR*LSDG	na	H
cSPH66	CndA	KNPV*YVDG	na	cH	SP21	SbMJ	NVNR*IVGG	T(DGG)	P	SP216	SJNL	PTNR*IVGG	T(DGG)	P
cSP67	CHbB	LQKR*IIGG	T(DGG)	cP	SP22	SiMd	TRHR*IVGG	T(DGG)	P	SP217	SJNK	ISPK*IMHG	T(DGG)	P
cSPH69	CEED	FSFR*EEDT	na	cH	SP23	SuVN	EEIR*IVGG	T(DGG)	P	SP218	SJNL	TAMR*VVNG	T(DGG)	P
cSPH79/Mas	COdC	RRAR*VVGG	na	5cH	SP27	SUPL	HDDF*NGRS	T(DGG)	P	SP219	SJNK	VATR*IVRG	T(DGG)	P
cSPH93	CdFa	NGLQ*MVEG	na	cH	SP29	SpPJ	YVER*IFPN	T(DGG)	P	SP220	SJNN	IAFK*IIGG	T(DGG)	P
cSPH94	CbFD	GSPQ*VFGD	na	cH	SP37	SKOJ	SQFK*ILGG	C(NGG)	P	SP221	ShNL	FRIR*VIGG	C(YGG)	P
cSPH101	CpdB	AAPG*QASF	na	cH	SP39	SnMJ	DESR*IVGG	T(DGG)	P	SP223	SKNJ	NEEH*QAHI	T(VGG)	P
cSP115	CUBB	SQNL*LVGG	T(DGG)	cP	SP40	SnML	NVNR*IVGG	T(DGG)	P	SP224	SKOJ	GLYR*VING	C(SGG)	P
cSPH121	CGDD	YQLD*GYNN	na	2cH	SPH41	SJNK	QCGL*MREE	na	H	SP225	SKOK	YRAR*IDGG	T(DGG)	P
cSPH125	CbeD	TVEE*VVDQ	na	cH	SP43	S3PL	YTPL*IVGG	T(DGG)	P	SP43	S3PL	YVPN*IFGG	T(DGG)	P
cSPH128	CdED	HVNR*IGVG	na	cH	SP45	S3PI	SRPL*IVDG	T(DGG)	P	SPH227	SYWI	NSDN*IIAE	na	H
cSPH142/Scaf	CedA	TKPT*GVKD	na	3cH	SP62	SjOJ	FIPM*ITGG	T(DGG)	P	SP228	SXjL	STYR*MVGG	T(DGG)	P
cSPH156	CdFD	TERT*QPGG	na	cH	SP63	SnnL	RNPK*IVGG	T(DGG)	P	SP230	SBLI	PEER*IVGG	T(DGG)	P
cSPH166	CcDB	LRPL*GYKQ	na	cH	SP68	S3PI	YAPL*IIGG	T(DGG)	P	SP233	S5YL	ILPK*IVGG	T(DGG)	P
cSP229	CXAB	TTNR*VIGG	C(GGG)	cP	SP70	ShkJ	QDGR*IVGG	T(DGG)	P	SP234	SxpL	EDGK*IVNG	C(GGS)	P
cSPH231	CEED	FSFR*EEDT	na	cH	SP73a	SsNK	HRTR*IIGG	C(NGG)	P	SPH236	SiON	RIRR*VVGG	na	H
cSP232	COAB	TSNR*VVGG	T(DGG)	cP	SP73b	SsNK	SVPR*VKNG	C(NGG)	P	SP237	ShNL	YTYR*ITGG	E(YGL)	P
cSPH242	CoDA	FTLS*GVSQ	na	cH	SP74	SfNN	LRRR*ITGG	T(DGG)	P	SP238	ShNL	FRMR*IFGG	C(FGG)	P
SP6	GIIF	LDGR*IVGG	T(DGG)	P	SP82	ShLK	GDGR*IVGG	T(DGG)	P	SP239	SJNL	LSYK*IING	E(NGY)	P
SP46	GzHF	LNGR*VVGG	E(GSD)	P	SP84	SnTc	TDSY*AVGQ	T(DGG)	P	SP240	SJNK	ISER*SVNA	E(QGK)	P
SP47/Ser6	GzHH	LNGR*VVGG	E(GVD)	P	SP86	S5YN	IEPK*IVGG	T(DGG)	P	SP243	SKOJ	VREQ*ILGG	C(SGG)	P
SP50	GzHF	IEPR*IVGG	C(GGD)	P	SP92/Ser12	SBLL	SPER*IVGG	T(DGG)	P	SPH244	SdQJ	TTIK*INHY	na	H
SP51	GPGH	IDGR*ITNG	C(SVA)	P	SP95	SxYL	QQSR*IING	T(DGG)	P	SP245	Sund	VGGR*IVST	T(DGG)	P
SP52	GyGH	IDNR*IVSG	E(SVS)	P	SPH97	SOKJ	PNGL*VANV	na	H	SP246	SfNI	IRFM*ITGG	T(DGG)	P
SP57/ηTry	GIIH	SDGR*IVGG	T(DGG)	P	SP104	SQLL	PQER*IVGG	T(DGA)	P	SPH249	S1NK	HMER*INGS	na	H
SP65	GhgF	ISTH*IVGG	T(DGG)	P	SP105	SPTL	AMDR*IFGG	C(SGG)	P	SPH250	S1NK	FLEQ*NCGK	na	H
SP71	GnGH	VEPY*ITNG	C(SGA)	P	SP108	S5YN	DPGR*IING	T(DGG)	P	SPH252	S2XM	SKEP*VVTL	na	H

name	G-C-T-E	activation cleavage site	specificity	domain	name	G-C-T-E	activation cleavage site	specificity	domain	name	G-C-T-E	activation cleavage site	specificity	domain
SP76λTry	GIIH	LDGR*IVGG	T(DGG)	P	SPH111	SDNL	LGPR*IVNG	na	H	SPH253	S2XM	QIVR*IINK	na	H
SP78	GPGE	PSGR*ITGG	C(SGG)	P	SP113	SBLI	WEQR*IIGG	C(TGG)	P	SPH254	S5YN	FQLK*MVAP	na	H-TM
SPH81	G6HH	PQGR*ILGG	na	H	SP118	ShkL	EESR*IIGG	C(SGG)	P	SPH255/Aqrs	S2XM	YNRE*ALEE	na	H
SP83	GdIE	PFGR*IVNG	T(DGG)	P	SP123	SPTN	GSLR*IMNG	C(LGG)	P	SP30	MnnJ	NMLK*IIGG	T(DGG)	2TSP-P
SP85/Jon74E	GnGH	IGGR*IAGG	C(SGS)	P	SP124a	SNSL	IQSR*IVGG	T(DGG)	P	SP49	MJNK	LIPK*IVGG	T(DGG)	PH
SP87	GpgF	SISR*VVNG	C(GSG)	P	SP124b	SNSL	RSPR*IVGG	T(DGG)	P	SP53/ModSP	MsoN	IKQF*SSGG	C(SGN)	Fig. 1
SP88/εTry	GIIE	LDGR*IVGG	T(DGG)	P	SP127	SPTL	VRTR*IAGG	C(GGG)	P	SP55	MVVI	QTPF*IHNG	C(GGG)	GdP
SP90/Ser8	GgIH	LGGR*IVGG	T(DGG)	P	SP130	SDNL	TAYR*IING	C(NGN)	P	SP60	MVVI	FSPL*QIGG	E(GGP)	GdP
SP91	GpgH	QMGR*VVNG	C(GSG)	P	SP132	StnJ	WTKK*FLAK	C(SGG)	P	SP72/Tequila	MmnN	REER*VVRG	T(DGG)	Fig. 1
SP96	GWJH	PETR*VIGG	C(GGD)	P	SP133	STUJ	SQER*VVGG	C(GGG)	P	SP75/Corin	MeiL	PSRR*IIGG	T(DGG)	Fig. 1
SP98/Jon65Aiii	GPGE	IEGR*ITNG	E(SVA)	P	SP136	SvkL	PDPR*IVGG	T(DGG)	P	SP77	MhNK	TRPK*ISGG	T(DGG)	PH
SP99	G6HF	VEPRI*VGG	E(GVD)	P	SP138	SBLI	IPER*IVGG	T(DGG)	P	SP80a	MFRc	ATTR*IANG	T(DGG)	CUBP
SP100/Jon44E	GGGH	IEGR*ITMG	C(SGA)	P	SP148a	SQLL	LPTR*IVNG	T(DGG)	P	SP80b	MFRK	FPNR*IANG	T(DGG)	CUBP
SP102/αTry	GIIE	LDGR*IVGG	T(DGG)	P	SP148b	SQLL	FPTR*IVNG	T(DGG)	P	SP89/Ndl	MknL	GDGR*IVGG	T(DGG)	Fig. 1
SP103	GAhE	ATGF*VING	C(SGS)	P	SP149	STUL	TPHR*IVGG	C(GGG)	P	SP120	MFRL	RIPR*IASP	T(DGG)	CUBP
SP106/Try29F	GcfF	LDGR*IVGG	T(DGG)	P	SP154	STUL	SPTR*INGG	C(GGG)	P	SPH144	MVVI	TTPL*IFQG	na	GdH
SP107	GbfF	LDGR*IVGG	T(DGG)	P	SP155	SsoJ	PDSR*IVNG	T(DGG)	P	SP186/Gd	MyVL	SLPS*ITRG	C(SGA)	GdP
SPH109/κTry	GIIH	PEGR*IIMG	na	H	SPH157	SEKL	VQFN*VTEG	na	H	SPH198	MyqL	SLPK*MSAP	na	Fig.1
SP110/ζTry	GIIH	PDGR*IVGG	T(DGG)	P	SP158	SzOL	AKLT*WWNY	C(HGG)	P	SP212	MVVL	TTPF*IVRG	C(GGG)	2Gd-P
SP112/Jon25Bii	GCGE	IEGR*ITNG	E(SVD)	P	SP159	SjSL	IQPR*IVGG	T(DGG)	P	SP222	MLNK	IALK*ITGG	T(DGG)	PH
SPH114/Sphe	G6HH	AQGR*IMGG	na	H	SP160	SxQM	FQFL*VTGG	T(DGG)	P	SPH241	MhNK	KTSE*NINF	na	HH
SP116/yip7	GPGE	ITGR*ITNG	C(SGA)	P	SP163/Send1	SBLI	PSER*IIGG	C(GGG)	P	SPH247	MLNJ	CGAP*ISNQ	na	HH
SP117/Jon65Aiv	GPGE	IGGR*ITGG	C(SGA)	P	SPH164	SBiM	SSNG*IYNG	na	H	SP248	MhNJ	PVPK*IISG	T(DGG)	PH
Sp119/βTry	GIIE	LDGR*IVGG	T(DGG)	P	SP167	SSSM	FQTR*VVGG	T(DGG)	P	SP251	MiNL	ITYR*VANG	T(DGG)	PH
Sp122/jon99cii	G4GE	IQGR*ITNG	C(SGA)	P	SP168/Sems	SSSM	YQTR*VIGG	T(DGG)	P					

Names, gene symbols (if available), G-C-T-E IDs, putative activation sites (*), predicted enzyme specificity (T for trypsin, C for chymotrypsin, and E for elastase) and its key determinants (in parentheses), as well as domain structures are enlisted. The 1st letters in the G-C-T-E IDs are group names: C for CLIPs, G for gut, M for other multi-domain, and S for other single-domain SPs/SPHs. The 2nd letters show chromosomal locations: A–Z and 1–6 for gene clusters and a–z for other genes (Fig. 3). The 3rd letters indicate positions (A–Y and a–q) in the phylogenetic trees (Fig. 2). The 4th letters mark major and minor expression assemblies A–N and a–d (Fig. 4). In domain structure, P stands for SP catalytic domain, H for non-catalytic SP-like domain, "c" for clip domain(s), and "TM" for transmembrane region. "CUB" and "Gd" domains are also indicated. More complicated domain structures are presented in Fig. 1.

Table 2.

Domain organization of the non-clip, multi-domain serine proteases and their homologs in the five model insects

Name/feature	Domain structure	*Dm*	*Ms*	*Ag*	*Am*	*Tc*
Nudel	TM-3LDL-SP-2–4LDL-SPH-3/4LDL	Ndl/SP89	SP50	SP212	SP20	mSP19
Tequila	S-2/15CBD-LDL-SR-LDL-SR-SP	Tequila/SP72	-	SP213	SP23[*]	mSP7[*]
Corin	TM-Fz-2LDL-SR-SP	Corin/SP75	SP56	SP214	SP30	mSP15
/SEA	TM-SEA-Fz-2LDL-SPH	SPH198	SPH145	SPH216	SPH56	mSPH6
ModSP	S-4/5LDL-Sushi-Wonton-SP	ModSP/SP53	HP14a, HP14b	SP217	SP49	mSP3, 13
/TSP	S-2TSP-SP	SP30	SP55	SP218	SP38	mSP2
Masquerade	S-5/4Clip-SPH	Mas/cSPH79	SPH53	CLIPA15	cSPH41, 39	cSPH51
/CUB	S-CUB-SP	SP120, SP80a-b	HP27	SP201–6	SP28, 34	mSP9, 10
/Gd	S-1/2Gd-SP	SP186/Gd, 55, 60, 212, SPH144,	SP45, 138, HP19	SP207–10	SP46	mSP1, 4, 5, 11, 12
GRAAL	TM-SEA-EGF-5LDL-SR-SPH			SP220	SPH54	mSPH16
/SEA	TM-SEA-3LDL-Coil-SP				SP45	
/SEA	TM-SEA-3LDL-SP-LDL					mSP8
/LamG	S-SP-Ig-2LamG			SP219		
/Sushi	S-1/2/3Sushi-SP(H)		SP25, 112, SPH33			
/2Sushi	S-2Sushi-SP-2Sushi-SP-2Sushi-SP					mSP14

CBD, type-2 chitin binding domain; CTL, C-type lectin domain; Coil, coiled coil region; CUB, a domain identified in complement 1r/s, uegf, and bmp1; EGF, a Ca^{2+}-binding domain in epidermal growth factor; Gd, a conserved region first identified in DmSP186/Gd; Fz, frizzled domain; Ig, immunoglobulin domain; LamG, a domain in laminin γ-subunit; LDL, low-density lipoprotein receptor class A repeat; S, signal peptide; SEA, a domain identified in a sperm protein, enterokinase and agrin; SP, serine protease catalytic domain; SPH, non-catalytic serine protease homolog domain; SR, scavenger receptor Cys-rich domain; Sushi, Sushi domain, also known as CCP or SCR; TM, transmembrane region; TSP, type-1 repeat in thrombospondin-1; Wonton, a Sushi-like module first identified in *M. sexta* HP14 with six instead of four Cys residues.

[*] AmSP23 and TcmSP7 are more complex in domain organization than AgSP213 or DmSP72/Tequila. In between the last CBD and the first LDL of Tequila, AmSP23 has SR-CTL-Kringle-LDL-Apple domains whereas TcmSP7 has only LDL-Apple domains.