Published in final edited form as:

Insect Biochem Mol Biol. 2017 September ; 88: 48–62. doi:10.1016/j.ibmb.2017.07.008.

Serine protease-related proteins in the malaria mosquito, *Anopheles gambiae*

Xiaolong Cao, Mansi Gulati, and Haobo Jiang*

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

Abstract

Insect serine proteases (SPs) and serine protease homologs (SPHs) participate in digestion, defense, development, and other physiological processes. In mosquitoes, some clip-domain SPs and SPHs (*i.e.* CLIPs) have been investigated for possible roles in antiparasitic responses. In a recent test aimed at improving quality of gene models in the *Anopheles gambiae* genome using RNA-seq data, we observed various discrepancies between gene models in AgamP4.5 and corresponding sequences selected from those modeled by Cufflinks, Trinity and Bridger. Here we report a comparative analysis of the 337 SP-related proteins in *A. gambiae* by examining their domain structures, sequence diversity, chromosomal locations, and expression patterns. One hundred and ten CLIPs contain 1 to 5 clip domains in addition to their protease domains (PDs) or non-catalytic, protease-like domains (PLDs). They are divided into five subgroups: CLIPAs (22). are clip₁₋₅-PLD; CLIPBs (29), CLIPCs (12) and CLIPDs (14) are mainly clip-PD; most CLIPs (33) have a domain structure of PD/PLD-PLD-clip-PLD₀₋₁. While expression of the CLIP genes in group-1 is generally low and detected in various tissue- and stage-specific RNA-seq libraries, some putative GPs/GPHs (*i.e.* single domain gut SPs/SPHs) in group-2 are highly expressed in midgut, whole larva or whole adult libraries. In comparison, 46 SPs, 26 SPHs, and 37 multi-domain SPs/SPHs (*i.e.* PD/PLD-PLD₁) in group-3 do not seem to be specifically expressed in digestive tract. There are 16 SPs and 2 SPH containing other types of putative regulatory domains (*e.g.* LDLa, CUB, Gd). Of the 337 SP and SPH genes, 159 were sorted into 46 groups (2–8 members/group) based on similar phylogenetic tree position, chromosomal location, and expression profile. This information and analysis, including improved gene models and protein sequences, constitute a solid foundation for functional analysis of the SP-related proteins in *A. gambiae*.

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.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

A group of eight CLIPAs with similar phylogenetic tree positions, chromosomal locations, and expression patterns

Keywords

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

1. Introduction

Chymotrypsin-related serine proteases (SPs) form a large family of enzymes that hydrolyze peptide bonds at different rates and with various degrees of specificity (Rawlings and Barrett, 1993). For instance, trypsin cleaves at a high rate, specifically after most Lys and Arg residues, consistent with its role in protein digestion; pancreatic elastase cuts efficiently after any accessible small nonpolar residues (*e.g.* Ala) in many proteins, whereas human coagulation factor Xa cleaves only few protein substrates in plasma, after certain Arg residues and at a low rate (k_{cat}). The S1 pocket of a protease interacts with the P1 residue of a protein substrate, governing its primary specificity (Schechter and Berger, 1967). Regulatory domains or regions in some nondigestive SPs provide additional specificity by localizing enzyme catalysis through specific interactions with activators, substrates, cofactors, and inhibitors (Kanost and Jiang, 2015; Krem and Di Cera, 2002). His, Asp and Ser residues in the active site of SPs are responsible for the acyl transfer mechanism of catalysis, with well-formed substrate binding clefts defining their specificities. SPs often contain a signal peptide guiding them to extracellular or granular locations, where they persist as inactive zymogens and then become activated by proteolytic cleavage at a particular peptide bond. In extracellular spaces, several SPs can constitute a cascade pathway in which one SP activates the zymogen of another in each step to trigger a rapid local response, such as blood coagulation or the complement system in mammals. In addition to active proteases, related serine protease homolog (SPH) genes encode SP-like

sequences lacking one or more of the catalytic triad residues and, thus, proteolytic activity. Some cleaved SPs are active as modulators of interacting SPs (Jiang et al., 2010; Park et al., 2010). While molecular mechanisms for such modulation are unclear, the SP-like fold and associated structural unit (*e.g.* clip domain) of SPs are likely essential for the interactions that determine their biochemical functions.

SP-related proteins mediate insect immune responses (*e.g.* melanotic encapsulation, cytokine activation, and antimicrobial peptide induction) (Jiang et al., 2010). Like human clotting factors, insect SPs and SPs form complex networks to stop bleeding and fight infection. In each insect species with a known genome, SP-related proteins form a large family with 60–400 members (Cao et al., 2015; Christophides et al., 2002; Ross et al., 2003; Waterhouse et al., 2007; Zhao et al., 2010; Zou et al., 2007; Zou et al., 2006). Their roles in defense and development have been explored in *Drosophila melanogaster*, *Manduca sexta*, *Tenebrio molitor*, and other insects (Kanost and Jiang, 2015; Park et al., 2010; Veillard et al., 2016). In mosquitoes, clip-domain SPs/SPs have been named CLIPs (Waterhouse et al., 2007). As summarized by Cao et al. (2015), numbers of the clip-domain SP/SP genes identified in genomes are 63 in *Aedes aegypti*, 55 in *A. gambiae*, 45 in *D. melanogaster*, 42 in *M. sexta*, and 49 in *Tribolium castaneum*.

Accurate gene models form a solid base for protein identification and elucidation of biochemical functions. Continuous efforts have been made to improve quality of the predicted genes after the initial genomes of *D. melanogaster*, *A. gambiae*, *Apis mellifera*, and other insects were published. The *M. sexta* genome project greatly benefited from next-generation sequencing, which provided RNA-seq data for the genome assembly, gene modeling and expression profiling (Kanost et al., 2016). We developed a method to select the best of the models from MAKER, Cufflink, Oases and Trinity programs (*i.e.* MCOT model) (Cao and Jiang, 2015). As this method has been automated and successfully applied in other insect genome projects (Cao and Jiang, 2017), we thought it would be interesting to test whether our method can further improve the latest release of *A. gambiae* gene models using the available RNA-seq data, with a focus on SP-like genes. Numerous discrepancies were identified between AgamP4.5 and corresponding AgMCOT models. To substantiate the observations and promote research on SP-related proteins in this species, we examined and improved the models in the official protein set (OPS), studied their domain organization and sequence diversity, classified them into the groups of CLIPs, GP(H)s and SP(H)s, and established an information system that contains systematic names, putative activation sites, predicted enzyme specificity, genomic locations, expression patterns, and phylogenetic relationships. Through further studies, we hope to establish a platform for comparing SP-related sequences from various insects and suggest functions for orthologs based on genetic and biochemical analyses in a few model species.

2. Materials and methods

2.1. Identification of *A. gambiae* SP-related proteins

OPS AgamP4.4 was downloaded from VectorBase (<https://www.vectorbase.org/>). Protein-coding genes were modeled using the MCOT pipeline (Cao and Jiang, 2015) by selecting the best for each gene from the OPS, TopHat-Cufflinks (Kim et al., 2013; Trapnell et al.,

2012), Trinity (Haas et al., 2013) and Bridger (Chang et al., 2015) assemblies to constitute an AgMCOT protein set (unpublished data). Domains in the AgamP4.4 and AgMCOT sequences were identified by InterProScan5 v5.17 (Jones et al., 2014) in a local supercomputer. Proteins containing a chymotrypsin-like (*i.e.* S1 family), SP-related domain were extracted and pooled. After removal of redundant, alternatively spliced, and severely incomplete genes, the sequences were manually examined and improved according to characteristic features of the S1 SPs, such as signal peptide and conserved regions.

2.2. Properties of *A. gambiae* SP and SPH sequences

Sequences were separated into SPs or SPHs by examining the presence of a His-Asp-Ser catalytic triad as described before (Cao et al., 2015). Signal peptides were 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 were identified by InterProScan5 and others by 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 at particular locations (Cao et al., 2015; Jiang and Kanost, 2000) 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 PD were identified in the aligned sequences for predicting their substrate specificity (Cao et al., 2015).

2.3. Multiple sequence alignment and phylogenetic analysis

Multiple sequence alignments of the entire sequences in the CLIP, GP(H), and SP(H) groups were performed using MUSCLE (Edgar, 2004), one module of MEGA 7.0 (Kumar et al., 2016), under the default setting with maximum iterations changed to 1,000. The classification and naming were based on: 1) clip domain presence or absence, 2) position in a phylogenetic tree of non-CLIP SPs or SPHs, and 3) expression patterns. Neighbor-joining trees were constructed in a preliminary analysis of the SP-related sequences, and reliability of the trees was tested using a bootstrap method with 1,000 trials. Alignments of the three 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 were <0.01 for GP(H)s and CLIPs, and <0.02 for SP(H)s. FigTree 1.4.3 (<http://tree.bio.ed.ac.uk/software/figtree/>) was used to display the phylogenetic trees.

2.4. Chromosomal locations of the SP and SPH genes

For most of the SP-related genes, their genomic locations were available in the information lists of the AgamP4.4 or Cufflinks models. Retrieved position data were plotted using ArkMAP 2.0 (<http://www.bioinformatics.roslin.ed.ac.uk/arkmap/>) and improved using Adobe Illustrator.

2.5. Expression profiling of the SP-related genes

The 113 RNA-seq data sets of *A. gambiae* from previous research (Bonizzoni et al., 2012; Mead et al., 2012; Pinheiro-Silva et al., 2015; Rinker et al., 2013; Vannini et al., 2014) were downloaded from NCBI Sequence Read Archive (SRA) and converted to fastq format using SRA Toolkit. Reads were first trimmed with Trimmomatic (Bolger et al., 2014) to remove adaptors and low quality bases with the setting “SLIDINGWINDOW:4:30 LEADING:20 TRAILING:20 MINLEN:50”. Transcript sequences of the SP-like proteins in AgamP4.4 were replaced with the improved ones (Section 2.1). FPKM (fragments per kilobase of transcript per million 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 that type of samples. Hierarchically clustered gradient heatmaps of $\log_2(\text{FPKM}+1)$ values were plotted using the clustermap function of Seaborn, a Python data visualization library, with the average linkage method and Euclidean matrix.

3. Results

3.1. Generation and classification of 337 reliable SP and SPH models in *A. gambiae*

In an effort to test if the MCOT approach (Cao and Jiang, 2015) is useful for improving the gene models in AgamP4.3 (<https://www.vectorbase.org/>), we upgraded the algorithm, automated data processing (Cao and Jiang, 2017), and generated an AgMCOT set by comparing, selecting, and naming the best models for individual genes from AgamP4.3, Cufflinks, Trinity and Bridger outputs (unpublished results). An InterPro domain scan of AgamP4.3 and AgMCOT sets resulted in a list of 732 SP-related sequences. After removing the redundant hits and sequences whose PD-like regions are shorter than 50% of a typical SP/SPH domain, we manually examined 366 candidates for flaws such as missing a signal peptide, incomplete PD/PLD domain, etc. After removal of isoforms of the same gene, the final list consists of 220 SPs and 117 SPHs (Table 1 and Table S1), one third of which have minor-to-major improvements as compared with the corresponding ones in AgamP4.5. Eighteen were absent in the AgamP4.5 release, and six of these eighteen (SP44, SP142, SPH12, SPH42, SPH113 and CLIPE32) were not detected in the genome assembly. Since the protein models we made are based on experimental data (*i.e.* RNA-seq reads), sequence comparison, model selection and manual curation, quality of the SP-related protein sequences is higher than Agam4.5, the newest official protein set (OPS) released on 2017-2-21.

As CLIPs are key components of insect immune SP-SPH networks (Kanost and Jiang, 2015), we have identified 110 SP-related proteins that contain 1 to 5 clip domains and named the 55 newly discovered CLIPs based on an initial phylogenetic analysis (data not shown). To avoid confusion, we did not change names of the ones reported before (Christophides et al., 2002; Waterhouse et al., 2007), even though six CLIPs (A19, C7, E1, E2, E6, E7) are assigned to different clades in our new analysis. According to a preliminary analysis of the expression patterns and phylogenetic relationships, we named 100 of the other 227 proteins gut proteases (GPs) or homologs (GPHs), based on containing a single PD/PLD with a typical size of 230 residues and higher expression level than CLIPs. The 127

SP(H)s were named by considering their domain structures: SP1–SP81 contain a single PD/PLD, SP101–SP144 contain 2 to 4 PDs/PLDs, and SP201–SPH220 contain a PD/PLD along with other non-clip regulatory domains (Table S1). Consequently, the SPs/SPHs are divided into three groups: 110 CLIPs, 100 GP(H)s and 127 SP(H)s. In the following, “SP-related proteins” or “SPs/SPHs” refer to all or part of the 337 regardless of their groups, whereas “SP(H)s” specify those in the third group.

3.2. General structural features of the 337 SP-related proteins

Consistent with their expected extracellular functions, 324 of the 337 sequences (except for SP133, 214; SPH10, 14, 35, 42, 106, 107, 113, 220; CLIPs B46, D22 and E22) are predicted to have a signal peptide for secretion (Table S1). The presence of catalytic residues His, Asp and Ser in the conserved motifs of TAAHC, DIAL and GDSGGP was used to predict if a protein is an active SP after activation. It is possible that some of the 220 SPs are catalytically inactive due to the lack of other essential structural features not considered. In contrast, none of the 117 SPHs are expected to be active proteases due to substitution of 1–3 of the catalytic residues, even though overall folding of the PDs and PLDs is likely similar due to sequence conservation.

Members of the CLIP, GP(H) and SP(H) groups differ in domain structure. Most of the 100 GP(H)s and 72 SP(H)s (SP1 to SP81) consist of a signal peptide, a pro-region, and a PD/PLD. The CLIPs in subgroups A–D contain a signal peptide, 1 to 5 clip domains, and a PD/PLD (Fig. 1). CLIPD22 has a transmembrane region 20 residues away from its amino terminus. CLIPs, as well as CLIPs C7, D11 and D14, have a structure of signal peptide-PD/PLD-PLD-clip-PLD_{0–1}. Thirty-seven SP(H)s (SP101 to SP144) have two or more PD/PLD domains. In eighteen of the multi-domain SP(H)s (SP201 to SPH220), we identified thirteen types of other domains, namely LDLA for low-density lipoprotein receptor class A (21), CUB for C1r/s, Uegf & Bmp1 (6), Gd for Gastrulation defective (6), SR for scavenger receptor (3), CB for chitin binding (2); LamG for laminin G (2), Fz for frizzled (2), TSP for thrombospondin (2), Ig for immunoglobulin (1), SEA for sperm protein, enterokinase and agrin (1), EGF for epidermal growth factor (1), Sushi (1), and Wonton (1) (Fig. 1). Numbers in parentheses are total numbers of the domains identified in all these proteins. These structural modules probably function in interactions of the proteases with themselves or partners and form SP-SPH cascades to mediate physiological processes and to guide proper domain interactions needed to control catalytic activities and localize proteolytic reactions. This notion is consistent with the conserved domain structures of SP217-ModSP, SP212-Nudel, CLIP15-Masquerade, and several other orthologous groups in a phylogenetically wide range of holometabolous insects, including beetles, moths, bees, mosquitos, and flies (Christophides et al., 2002; Ross et al., 2003; Waterhouse et al., 2007; Zou et al., 2007; Zou et al., 2006). *Drosophila* ModSP, Nudel, and Masquerade are 1:1 orthologs of the mosquito proteins.

3.3. Phylogenetic relationships, genome locations and expression patterns of the 110 CLIPs

Clip domains constitute the largest group of regulatory structures in the SP-related proteins of *A. gambiae*. These disulfide-bridged units exist in insect and crustacean SPs/SPHs

involved in defense, development and other processes (Kanost and Jiang, 2015). In total, 126 clip domains were identified in 63 SPs and 47 SPs – seven CLIPs have 2, 3 or 5 clip domains (Fig. 1). Seventy-one CLIPs have one clip domain at the amino terminus; 23 CLIPs, CLIPs C7, D11 and D14 have a CLIP domain between PLDs; other 5 CLIPs have their clip domain at the carboxyl end. In this study, we have identified 55 CLIPs not previously annotated: 8 CLIPA (19, 20, 26–28, 30–32), 9 CLIPB (3b, 36, 41–47), 4 CLIPC (9, 12–14), 7 CLIPD (9, 11–14, 20, 22), and 27 CLIP E (8–34). Together with those reported before, 22 CLIPAs, 29 CLIPBs, 12 CLIPCs, 14 CLIPDs, and 33 CLIPEs exist in *A. gambiae*. A majority of the mature CLIPs have a distinct domain organization of P(L)D-PLD-clip-PLD₀₋₁. The PD, PLD, and clip domain may organize into higher structures to perform complex functions.

A phylogenetic tree based on alignment of complete sequences of the 110 CLIPs reveals evolutionary relationships among them (Fig. 2A). Separation of the five clades is obvious: all but one CLIPA and CLIPs E1, E2, E6, E7 form a monophyletic group with a probability (P) of 99; most CLIPDs (apart from D11, D14) form two groups (P: 99 and 85); CLIPBs and CLIPA19 forms three groups (P: 88, 100, 94); CLIPCs (except for C7) form three groups (P: 100); most CLIPEs and C7 form seven groups (P: 95–100). The grouping of relevant genes generally agrees with their locations in ten regions (*i.e.* 2E–G, 3D, 3G, 3H, 4E, 4G, 4H, 5B) of the chromosomes 2R (2nd half), 3 and X (Fig. 3). Apparently, rounds of gene duplication have given rise to the clusters of closely related CLIP genes. Since regulatory elements may be duplicated along with the coding regions in members of a gene cluster, we anticipated and then observed a considerable level of consistency in expression patterns (Fig. 4, Table S2) among genes with similar sequence and chromosomal location. For example, genes of CLIPA1, 2, 4, 6, 7, 12, 14, and 30 in tree group “LA” (Fig. 2A) reside in the same region of chromosome 3L (Fig. 3, location group “3D”) and have similar expression profiles (Fig. 4, expression group “CA”). We have observed 14 similar three-way agreements (phylogenetic tree, chromosomal location and expression pattern), each with 2 to 8 genes and involving a total of 46 CLIPs. Transcript levels of the CLIPAs, Bs and Cs in expression group “CA” are much higher than those of CLIPs in “CG” and “CH”, especially in the adults. The profiles of CLIPs A9, A10, B3a, B12, B17, B19, C1, C14, D2, D12, E4, and E8 mRNA levels are distinct in the 45 cDNA libraries, making them attractive targets for functional studies.

3.4. Evolution, location, and expression of the 100 putative GP(H)s in *A. gambiae*

By definition, GP(H)s are serine proteases and their homologs expressed in midgut tissues. While experimental evidence is needed for naming, only two libraries are available for midguts, both from the blood-fed female adults (Mead et al., 2012). As described in *Section 3.1*, we tentatively named them by integrating information from the preliminary analyses of gene expression, sequence similarity and chromosome locations. The profiles of GP(H) mRNA levels demonstrated four major expression groups (Fig. 5): “GB” for 18 GPs and 22 GPHs highly expressed in the larval stages; “GC” for 20 GPs and 10 GPHs mostly expressed at lower levels in larvae; “GD” for 4 GPs and GPH3 expressed in larvae and in adults at lower levels; “GE” for 20 GPs and 4 GPHs expressed at low levels in pupae and male adults and at high levels in female adults. GP19 and GP26 expression in pupae and adults are very high, particularly in the midgut of female adults. GP10, 13–15, 17, 24, 103, GPH16 and 18

transcripts are more abundant in female than male adults and peak in midgut after blood feeding. The expression of GP5, 6, 7, 13, and GPH4 in midgut was higher after feeding on normal blood than infectious blood containing *Plasmodium falciparum* (Mead et al., 2012). GP5 and GPH4 mRNA levels were high in salivary glands, and GPH99 in antennae.

Most GP(H) genes are located in 14 regions on the left (location IDs: 1D, 1F, 1G, 1H, 1J, 1L, 1N, 1P, 1R, 3F) and right (2A, 2B, 4C, and 4F) arms of chromosomes 2 and 3 (Fig. 3). Genes located in each of these regions are generally consistent with their positions (“TA”–“TE”, “TG”, “TH”, and “TK”) in the phylogenetic tree (Fig. 2B). It appears that extensive gene duplication has resulted in large clusters of GP(H) genes, whose transcription is regulated in a similar manner for each gene group. We have identified 16 such three-way agreement groups, each involving 2 to 7 members whose gene locations, tree positions, and expression patterns are the same. Among the 65 in these groups, the GPH28, 29, 35, 37, 38, 101 and 102 genes (tree ID: “TG”, Fig. 2B) reside in the “1D” region of chromosome 3L (Fig. 3), have similar expression profiles (expression ID: “GC”, Fig. 5). GP6, 7, 9, 10, 24, 26 and 103 genes have the tree ID “TD”, location ID “4C”, and expression ID “GE”; GPH61, 69, 77, 91, 94, and 95 are in “TK”, “1H” and “GB”.

3.5. Features of the 127 SP(H) genes in *A. gambiae*

Most SP(H) genes are found in ten regions of chromosomes 2 (location groups: 1G, 2H, 2N, 2P) and 3 (3B, 3E, 3J, 3K, 4B) (Fig. 3). In region “3E”, a recently evolved cluster of six genes encode CUB-domain SPs 201–206, five of which are identical in tree position (“PJ”) (Fig. 2C) and expression group (“SD”) (Fig. 6). The two gene doublets encoding Gd-domain SPs 207–210 are likely products of two rounds of gene duplication, even though they are 5.2 Mb apart (Fig. 3). Such evolutionary events are proposed based on structural similarity and phylogenetic relationships of these SPs (Fig. 2C). There are twelve gene dyads, three triads, one tetrad, one pentad, and one hexad, based on similar gene locations, tree positions, and expression groups. For instance, SPs 74–77 genes in tree group “PB” (Fig. 2C) and region “4B” of chromosome 3R (Fig. 3) are mainly expressed in adult males (Fig. 6, group “SJ”).

Judged on the basis of their $\log_2(\text{FPKM}+1)$ values, most of the SP(H) genes are expressed at low levels in the RNA samples of whole insects (Fig. 6). Transcript levels of group-SA and -SB genes are moderate-to-high in most libraries, whereas high mRNA abundances are detected in a few tissue types for genes in groups Sc, SD, Sf, SG, SH, and SJ. High expression of SP2, SP3 and SP4 in adult females (but not males) led us to consider their possible involvement in reproduction.

4. Discussion

4.1. Improving the AgamP4.3 gene models using AgMCOT

Correct modeling of protein-encoding genes based on genome and cDNA sequences is important for guiding functional studies of their protein products. Several programs have been developed to fulfil this goal, with varying degrees of success. We took an integrated approach that compares and selects the best from models predicted by different programs for a single and then all genes in *M. sexta* (Cao and Jiang, 2015). In this study, we employed the

same method to improve AgamP4.3 and generated AgMCOT, which represents a collection of the selected protein models. We then focused on SP-related proteins by manually examining the corresponding ones and validating improvements in 117 of the 337 SP-like sequences in AgamP4.5. Since parallel study of the *Drosophila* SP-like genes resulted in fewer than 10 such corrections (data not shown), we think the room is still large for improving AgamP4.5 and even the genome assembly. It is possible that some of the SP-like genes not detected in the genome assembly but with good evidence from RNA-seq data are located near their close relatives in a chromosomal region that has not been assembled. Models for proteins other than SPs/SPHs in AgMCOT should be used to validate and improve the respective sequences in the latest OPS. The power of AgMCOT stems from genome-independent assemblies of RNA-seq data that are integrated with the MAKER/OPS or Cufflink models during selection and manual curation.

4.2. Functional importance of the *A. gambiae* CLIPs

Specific genetic traits of *A. gambiae* cause developmental arrest and melanotic encapsulation of *Plasmodium cynomolgi* ookinetes (Collins et al., 1986). Since phenoloxidases (POs) are key enzymes that catalyze melanization, the proteolytic activation of PO zymogens (*i.e.* proPOs) by an SP-SPH system has been studied by reverse genetic methods. Knocking down CLIPs B4 or B8 led to reduced melanization of Sephadex beads, and silencing CLIP B1, B9 or B10 had lesser effects (Paskewitz et al., 2006). RNAi silencing of CLIP A8/B4/B8/B14/B15/B17 and A2/A5/A7 decreased and increased melanization of *Plasmodium berghei* ookinetes and oocysts (Volz et al., 2006; Volz et al., 2005; Zhang et al., 2016), respectively. Recombinant CLIPB9_{Xa}, activated by bovine clotting factor Xa, cleaved *M. sexta* proPOs and generated POs with a low specific activity (An et al., 2011). Melanization has a functional link to TEP1 activation via CLIPA2 (Yassine et al., 2014) and CLIPA30 (*i.e.* SPCLIP1) (Povelones et al., 2013). Formation of a complex of TEP1, LRIM1 and APL1C is required in defense against malaria parasites and bacteria in *A. gambiae*. Transcriptome analysis showed that CLIPC2 was preferentially induced in midgut of *A. gambiae* by *P. fuscarius* infection (Blumberg et al., 2013). RNAi screening revealed CLIPA26's role in transcriptional regulation and SPH51's role in phagocytosis (Lombardo et al., 2013). Together, these studies have begun to elucidate CLIPs' functions in the mosquito immune responses.

Interestingly, CLIPs A2, A5, A7 and A30, encoded by the same cluster of genes in region “3D” on chromosome 3L, all regulate melanization and/or TEP1 activation. Such functional relatedness also exists in the gene triplet of CLIPs B8, B9 and B10, as well as in the doublet of CLIPs B1 and B4. One explanation is that during neo- or sub-functionalization, duplicated genes may maintain certain levels of their ancestor's original functions. On the other hand, if dosage increase of the gene copies is detrimental to the host, one or more of the copies may encounter functional loss. Since molecular functions are mostly unclear for proteins encoded by the 15 CLIP genes in “3D”, it would be exciting to find out what functions these copies of original genes have. With the tree, location, and expression IDs available, RNAi of entire gene clusters may produce mutant phenotypes that are masked by functional redundancy in single knockdown tests. Once a strong phenotype is found, scaling down the targets should reveal the culprit(s).

4.3. Functions and expression regulation of the putative GPs and GPHs in *A. gambiae*

GPs had been studied for years before the *A. gambiae* genome was published. These include GP5 (Sp24D), GP6 (Antryp6), GP7 (Antryp5), GP9 (Antryp3), GP10 (Antryp7), GP12 (ISP13), GP13 (AgChyL), GP22 (Anchym1), GP23 (Anchym2), GP24 (Antryp2), GP26 (Antryp1), GP97 (AgESP) and GP103 (Antryp4) (Dimopoulos et al., 1997; Han et al., 1997; Muller et al., 1995; Muller et al., 1993; Rodrigues et al., 2012; Shen et al., 2000; Vizioli et al., 2001). Expression patterns of GP5, GP13, GP22 and GP26, for example, vary dramatically, due to the regulatory elements in their genes (Giannoni et al., 2001; Shen and Jacobs-Lorena, 1998; Skavdis et al., 1996). While these studies support the naming, mRNA profiles and structure features (*e.g.* size, domain, similarity) of the 100 G(P)Hs provide additional evidence for the classification. Nonetheless, we must point out that experimental data are necessary to validate their identities as GPs and GPHs. For instance, GP5 mRNA level is much higher in thorax than gut and in adult males than females (Han et al., 1997) and GP97 is involved in the *Plasmodium* invasion of midgut and salivary glands (Rodrigues et al., 2012).

In the larvae, dietary proteins are likely processed by the 37 GPs in expression groups GB and GC (Fig. 5). The transcript levels are much higher for genes in group GB (18 GPs, 22 GPHs) than in group GC (20 GPs, 10 GPHs). We do not know their protein levels or catalytic activities to estimate their relative contributions to digestion but, if all things (*e.g.* translation, stability) are the same, why would the larvae make similar or more GPHs than GPs in the midgut (Fig. 5, Table S2)? In other words, what physiological roles do these non-catalytic proteins play in the midgut? Do they protect the host cells from damage caused by excessive GPs or toxic molecules taken up from the environment? *Bacillus thuringiensis israelensis*, a naturally occurring soil bacterium, is used as a biological control agent to kill mosquito larvae in water (Shaalán and Canyon, 2009). Its insecticidal crystal proteins require proper cleavage by GPs to form active toxins. Under- or over-processing of the protoxins by GPs may both impact their effectiveness and, therefore, call for further studies of the mixture of GPs and GPHs.

It is also possible that GPs and GPHs in expression group GC serve a function different from those in group GB. The proteins in group GC may be constitutively synthesized at low levels and released as samplers to produce a basal level of amino acids from ingested food. If the level exceeds a threshold when dietary proteins are present, GPs and GPHs in the GB group are then expressed at high level and released for the bulk digestion and protection of larval tissues, respectively. Such a scenario was reported in adult females of *Aedes aegypti* (Noriega and Wells, 1999).

GP1, GP2, GP96, GP97 and GPH3 in group D, GP5, GP88 and GP98 are expressed in larvae, pupae and adults, whereas the other 20 GPs and 3 GPHs in group GE are mainly expressed in pupae and adults (Fig. 5). It is clear that different gene sets are employed by the mosquito for digestion in larvae and adults. The GP(H) transcript levels in pupae are generally low except for GP19, GP20 and GP26. Roles of these putative GPs in tissue remodeling need exploration in the pupae, and so do the tissue specificity and sex dichotomy of GPs in the adults.

4.4. Functions and transcription of the *A. gambiae* SPs and SPHs

Even though some of the 126 SP(H)s have interesting domain structures (Fig. 1, Table S1), their functions are poorly explored, except for SP2, SP3, SP4, SPH51, and SP213 (Danielli et al., 2000; Gorman et al., 2000; Lombardo et al., 2013; Mancini et al., 2011). Consistent with their specific expression in the adult females (*Section 3.5*), SP2, SP3 and SP4 proteins are detected in the lower reproductive tissues to process transferred male proteins. SP213 (GRAAL or Sp22D) may mediate immune responses, as its constitutive expression in adult hemocytes, fat body and midgut epithelial cells is induced 1.5 fold after wounding or bacterial infection. SP212 and SP217 are orthologs of *Drosophila* Nudel and ModSP, which are involved in embryonic development and immune responses, respectively.

Of the 25, 44, 40 and 17 SPs/SPHs in expression groups SA–SB, Sc–SJ, Sk–SN and Sq–SR 13, 20, 9, and 13 have two or more domains (Fig. 6). The transcript levels in groups Sk–SN were the lowest, slightly higher and more evenly distributed in these libraries for groups Sq–SR, a lot higher in some tissue types for groups Sc–SJ, and the highest in most of the libraries for groups SA–SB. Specific expression of the genes in Sc–SJ (*e.g.* SP2, SP3 and SP4) in RNA-seq libraries is interesting, which may provide clues for their functional elucidation.

5. Conclusions

Serine proteases and their homologs constitute a large family of proteins in *A. gambiae*. We generated the AgMCOT gene set and made improvements in the SP and SPH sequences of AgamP4.5. Extensive RNA-Seq data not only enhanced the quality of AgMCOT models but also revealed the expression patterns of 220 SPs and 117 SPHs. We also identified close connections among phylogenetic relationships, chromosomal locations, and expression profiles for 159 genes in 46 groups. Structural features and other information of the SP-related proteins are provided to facilitate research on their physiological functions. We have identified thirteen types of cysteine-stabilized domains in 127 SP(H)s, which may allow molecular recognition to occur among members of SP-SPH cascade pathways in the malaria mosquito.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Dr. Michael Kanost at Kansas State University for his insightful comments, which greatly helped the manuscript improvement. This work was supported by NIH grants AI112662 and GM58634. We would like to thank the mosquito scientists for producing the RNA-seq data, especially the researchers in Pirbright Institute who have deposited their data in NCBI SRA but not yet published their analyses. Computation for this project was done at OSU High Performance Computing Center, supported in part through the NSF grant OCI-1126330. This work was approved for publication by the Director of Oklahoma Agricultural Experimental Station and supported in part under project OKLO2450.

Abbreviations

SP serine protease

SPH	(non-catalytic) serine protease homolog
PD	SP catalytic domain
PLD	protease-like domain in SPH
LDLa	low-density lipoprotein receptor class A repeat
SR	scavenger receptor
TSP	thrombospondin
CUB	C1r/C1s, Uegf, Bmp1
MSP	modular serine protease
CLIP	clip-domain SP or SPH
GP and GPH	gut serine protease and gut serine protease homolog
PO and proPO	phenoloxidase and its precursor
PAP	proPO activating protease

References

- An CJ, Budd A, Kanost MR, Michel K. Characterization of a regulatory unit that controls melanization and affects longevity of mosquitoes. *Cellular and Molecular Life Sciences*. 2011; 68:1929–1939. [PubMed: 20953892]
- Blumberg BJ, Trop S, Das S, Dimopoulos G. Bacteria- and IMD Pathway-Independent Immune Defenses against *Plasmodium falciparum* in *Anopheles gambiae*. *Plos One*. 2013;8.
- Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*. 2014; 30:2114–2120.
- Bonizzoni M, Afrane Y, Dunn WA, Atieli FK, Zhou G, Zhong D, Li J, Githeko A, Yan G. Comparative transcriptome analyses of deltamethrin-resistant and -susceptible *Anopheles gambiae* mosquitoes from Kenya by RNA-Seq. *PloS one*. 2012;7.
- Cao X, He Y, Hu Y, Zhang X, Wang Y, Zou Z, Chen Y, Blissard GW, Kanost MR, Jiang H. Sequence conservation, phylogenetic relationships, and expression profiles of nondigestive serine proteases and serine protease homologs in *Manduca sexta*. *Insect biochemistry and molecular biology*. 2015; 62:51–63. [PubMed: 25530503]
- Cao X, Jiang H. Integrated modeling of protein-coding genes in the *Manduca sexta* genome using RNA-Seq data from the biochemical model insect. *Insect biochemistry and molecular biology*. 2015; 62:2–10. [PubMed: 25612938]
- Cao X, Jiang H. Integrated modeling of structural genes using MCuNovo. *Insect Genomics, Methods in Molecular Biology*. 2017 (submitted).
- Chang Z, Li G, Liu J, Zhang Y, Ashby C, Liu D, Cramer CL, Huang X. Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. *Genome biology*. 2015; 16:30. [PubMed: 25723335]
- 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, Muller HM, Osta MA, Paskewitz SM, Reichhart JM, Rzhetsky A, Troxler L, Vernick KD, Vlachou D, Volz J, von Mering C, Xu JN, Zheng LB, Bork P, Kafatos FC. Immunity-related genes and gene families in *Anopheles gambiae*. *Science*. 2002; 298:159–165. [PubMed: 12364793]

- Collins FH, Sakai RK, Vernick KD, Paskewitz S, Seeley DC, Miller LH, Collins WE, Campbell CC, Gwadz RW. Genetic Selection of a Plasmodium-Refractory Strain of the Malaria Vector *Anopheles-Gambiae*. *Science*. 1986; 234:607–610. [PubMed: 3532325]
- Danielli A, Loukeris TG, Lagueux M, Muller HM, Richman A, Kafatos FC. A modular chitin-binding protease associated with hemocytes and hemolymph in the mosquito *Anopheles gambiae*. *Proceedings of the National Academy of Sciences of the United States of America*. 2000; 97:7136–7141. [PubMed: 10860981]
- Dimopoulos G, Richman A, Muller HM, Kafatos FC. Molecular immune responses of the mosquito *Anopheles gambiae* to bacteria and malaria parasites. *Proc Natl Acad Sci U S A*. 1997; 94:11508–11513. [PubMed: 9326640]
- Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*. 2004; 32:1792–1797. [PubMed: 15034147]
- Giannoni F, Muller HM, Vizioli J, Catteruccia F, Kafatos FC, Crisanti A. Nuclear factors bind to a conserved DNA element that modulates transcription of *Anopheles gambiae* trypsin genes. *Journal of Biological Chemistry*. 2001; 276:700–707. [PubMed: 11016929]
- Gorman MJ, Andreeva OV, Paskewitz SM. Sp22D: a multidomain serine protease with a putative role in insect immunity. *Gene*. 2000; 251:9–17. [PubMed: 10863091]
- 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. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols*. 2013; 8:1494–1512. [PubMed: 23845962]
- Han YS, Salazar CE, Reese-Stardy SR, Cornet A, Gorman MJ, Collins FH, Paskewitz SM. Cloning and characterization of a serine protease from the human malaria vector, *Anopheles gambiae*. *Insect Mol Biol*. 1997; 6:385–395. [PubMed: 9359580]
- Jiang H, Kanost MR. The clip-domain family of serine proteinases in arthropods. *Insect Biochem Mol Biol*. 2000; 30:95–105. [PubMed: 10696585]
- Jiang H, Vilcinskas A, Kanost MR. Immunity in lepidopteran insects. *Advances in experimental medicine and biology*. 2010; 708:181–204. [PubMed: 21528699]
- 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. InterProScan 5: genome-scale protein function classification. *Bioinformatics*. 2014; 30:1236–1240. [PubMed: 24451626]
- Kanost MR, Arrese EL, Cao X, Chen YR, Chellapilla S, Goldsmith MR, Grosse-Wilde E, Heckel DG, Herndon N, Jiang H, Papanicolaou A, Qu J, Soulages JL, Vogel H, Walters J, Waterhouse RM, Ahn SJ, Almeida FC, An C, Aqrabi P, Bretschneider A, Bryant WB, Bucks S, Chao H, Chevignon G, Christen JM, Clarke DF, Dittmer NT, Ferguson LC, Garavelou S, Gordon KH, Gunaratna RT, Han Y, Hauser F, He Y, Heide-Fischer H, Hirsh A, Hu Y, Jiang H, Kalra D, Klinner C, Konig C, Kovar C, Kroll AR, Kuwar SS, Lee SL, Lehman R, Li K, Li Z, Liang H, Lovelace S, Lu Z, Mansfield JH, McCulloch KJ, Mathew T, Morton B, Muzny DM, Neunemann D, Onger F, Pauchet Y, Pu LL, Pyrousis I, Rao XJ, Redding A, Roesel C, Sanchez-Gracia A, Schaack S, Shukla A, Tetreau G, Wang Y, Xiong GH, Traut W, Walsh TK, Worley KC, Wu D, Wu W, Wu YQ, Zhang X, Zou Z, Zucker H, Briscoe AD, Burmester T, Clem RJ, Feyereisen R, Grimmelikhuijzen CJ, Hamodrakas SJ, Hansson BS, Huguet E, Jermini LS, Lan Q, Lehman HK, Lorenzen M, Merzendorfer H, Michalopoulos I, Morton DB, Muthukrishnan S, Oakeshott JG, Palmer W, Park Y, Passarelli AL, Rozas J, Schwartz LM, Smith W, Southgate A, Vilcinskas A, Vogt R, Wang P, Werren J, Yu XQ, Zhou JJ, Brown SJ, Scherer SE, Richards S, Blissard GW. Multifaceted biological insights from a draft genome sequence of the tobacco hornworm moth, *Manduca sexta*. *Insect Biochem Mol Biol*. 2016; 76:118–147. [PubMed: 27522922]
- Kanost MR, Jiang HB. Clip-domain serine proteases as immune factors in insect hemolymph. *Curr Opin Insect Sci*. 2015; 11:47–55. [PubMed: 26688791]
- Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*. 2013; 14:R36. [PubMed: 23618408]

- Krem MM, Di Cera E. Evolution of enzyme cascades from embryonic development to blood coagulation. *Trends Biochem Sci.* 2002; 27:67–74. [PubMed: 11852243]
- Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol.* 2016; 33:1870–1874. [PubMed: 27004904]
- Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods.* 2012; 9:357–359. [PubMed: 22388286]
- 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]
- Lombardo F, Ghani Y, Kafatos FC, Christophides GK. Comprehensive genetic dissection of the hemocyte immune response in the malaria mosquito *Anopheles gambiae*. *PLoS pathogens.* 2013;9.
- Mancini E, Tammamo F, Baldini F, Via A, Raimondo D, George P, Audisio P, Sharakhov IV, Tramontano A, Catteruccia F, della Torre A. Molecular evolution of a gene cluster of serine proteases expressed in the *Anopheles gambiae* female reproductive tract. *Bmc Evol Biol.* 2011;11. [PubMed: 21226948]
- Mead EA, Li M, Tu Z, Zhu J. Translational regulation of *Anopheles gambiae* mRNAs in the midgut during *Plasmodium falciparum* infection. *BMC Genomics.* 2012; 13:366. [PubMed: 22857387]
- Muller HM, Catteruccia F, Vizioli J, della Torre A, Crisanti A. Constitutive and blood meal-induced trypsin genes in *Anopheles gambiae*. *Exp Parasitol.* 1995; 81:371–385. [PubMed: 7498434]
- Muller HM, Vizioli I, della Torre A, Crisanti A. Temporal and spatial expression of serine protease genes in *Anopheles gambiae*. *Parassitologia.* 1993; 35(Suppl):73–76.
- Noriega FG, Wells MA. A molecular view of trypsin synthesis in the midgut of *Aedes aegypti*. *Journal of Insect Physiology.* 1999; 45:613–620. [PubMed: 12770346]
- Park JW, Kim CH, Rui J, Park KH, Ryu KH, Chai JH, Hwang HO, Kurokawa K, Ha NC, Soderhall I, Soderhall K, Lee BL. Beetle Immunity. *Adv Exp Med Biol.* 2010; 708:163–180. [PubMed: 21528698]
- Paskewitz SM, Andreev O, Shi L. Gene silencing of serine proteases affects melanization of Sephadex beads in *Anopheles gambiae*. *Insect Biochemistry and Molecular Biology.* 2006; 36:701–711. [PubMed: 16935219]
- Perona JJ, Craik CS. Structural basis of substrate specificity in the serine proteases. *Protein Sci.* 1995; 4:337–360. [PubMed: 7795518]
- Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature methods.* 2011; 8:785–786. [PubMed: 21959131]
- Pinheiro-Silva R, Borges L, Coelho LP, Cabezas-Cruz A, Valdes JJ, do Rosario V, de la Fuente J, Domingos A. Gene expression changes in the salivary glands of *Anopheles coluzzii* elicited by *Plasmodium berghei* infection. *Parasit Vectors.* 2015; 8:485. [PubMed: 26395987]
- Povelones M, Bhagavatula L, Yassine H, Tan LA, Upton LM, Osta MA, Christophides GK. The CLIP-Domain Serine Protease Homolog SPCLIP1 Regulates Complement Recruitment to Microbial Surfaces in the Malaria Mosquito *Anopheles gambiae*. *Plos Pathogens.* 2013;9.
- Rawlings RD, Barrett AJ. Evolutionary families of peptidases. *Biochem J.* 1993; 290:205–218. [PubMed: 8439290]
- Rinker DC, Pitts RJ, Zhou X, Suh E, Rokas A, Zwiebel LJ. Blood meal-induced changes to antennal transcriptome profiles reveal shifts in odor sensitivities in *Anopheles gambiae*. *Proceedings of the National Academy of Sciences of the United States of America.* 2013; 110:8260–8265. [PubMed: 23630291]
- Rodrigues J, Oliveira GA, Kotsyfakis M, Dixit R, Molina-Cruz A, Jochim R, Barillas-Mury C. An Epithelial Serine Protease, AgESP, Is Required for *Plasmodium* Invasion in the Mosquito *Anopheles gambiae*. *Plos One.* 2012;7.
- Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Hohna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol.* 2012; 61:539–542. [PubMed: 22357727]
- Ross J, Jiang H, Kanost MR, Wang Y. Serine proteases and their homologs in the *Drosophila melanogaster* genome: an initial analysis of sequence conservation and phylogenetic relationships. *Gene.* 2003; 304:117–131. [PubMed: 12568721]

- Schechter I, Berger A. On the size of the active site in proteases. I. Papain. *Biochem Biophys Res Commun.* 1967; 27:157–162. [PubMed: 6035483]
- Shaan EAS, Canyon DV. Aquatic insect predators and mosquito control. *Trop Biomed.* 2009; 26:223–261. [PubMed: 20237438]
- Shen HB, Chou KC. Signal-3L: A 3-layer approach for predicting signal peptides. *Biochem Biophys Res Commun.* 2007; 363:297–303. [PubMed: 17880924]
- Shen Z, Edwards MJ, Jacobs-Lorena M. A gut-specific serine protease from the malaria vector *Anopheles gambiae* is downregulated after blood ingestion. *Insect Molecular Biology.* 2000; 9:223–229. [PubMed: 10886405]
- Shen ZC, Jacobs-Lorena M. Nuclear factor recognition sites in the gut-specific enhancer region of an *Anopheles gambiae* trypsin gene. *Insect Biochemistry and Molecular Biology.* 1998; 28:1007–1012. [PubMed: 9887516]
- Skavdis G, Sidenkiamos I, Muller HM, Crisanti A, Louis C. Conserved function of *Anopheles gambiae* midgut-specific promoters in the fruitfly. *Embo J.* 1996; 15:344–350. [PubMed: 8617209]
- Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley D, Pimentel H, Salzberg S, Rinn J, Pachter L. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature protocols.* 2012; 7:562–578. [PubMed: 22383036]
- Vannini L, Dunn WA, Reed TW, Willis JH. Changes in transcript abundance for cuticular proteins and other genes three hours after a blood meal in *Anopheles gambiae*. *Insect Biochemistry and Molecular Biology.* 2014; 44:33–43. [PubMed: 24269292]
- Veillard F, Troxler L, Reichhart JMM. *Drosophila melanogaster* clip-domain serine proteases: Structure, function and regulation. *Biochimie.* 2016; 122:255–269. [PubMed: 26453810]
- Vizioli J, Catteruccia F, della Torre A, Reckmann I, Muller HM. Blood digestion in the malaria mosquito *Anopheles gambiae* - Molecular cloning and biochemical characterization of two inducible chymotrypsins. *European Journal of Biochemistry.* 2001; 268:4027–4035. [PubMed: 11453997]
- Volz J, Muller HM, Zdanowicz A, Kafatos FC, Osta MA. A genetic module regulates the melanization response of *Anopheles* to *Plasmodium*. *Cell Microbiol.* 2006; 8:1392–1405. [PubMed: 16922859]
- Volz J, Osta MA, Kafatos FC, Muller HM. The roles of two clip domain serine proteases in innate immune responses of the malaria vector *Anopheles gambiae*. *Journal of Biological Chemistry.* 2005; 280:40161–40168. [PubMed: 16188883]
- 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. Evolutionary dynamics of immune-related genes and pathways in disease-vector mosquitoes. *Science.* 2007; 316:1738–1743. [PubMed: 17588928]
- Yassine H, Kamareddine L, Chamat S, Christophides GK, Osta MA. A Serine Protease Homolog Negatively Regulates TEP1 Consumption in Systemic Infections of the Malaria Vector *Anopheles gambiae*. *Journal of Innate Immunity.* 2014; 6:806–818. [PubMed: 25012124]
- Zhang X, An CJ, Sprigg K, Michel K. CLIPB8 is part of the prophenoloxidase activation system in *Anopheles gambiae* mosquitoes. *Insect Biochemistry and Molecular Biology.* 2016; 71:106–115. [PubMed: 26926112]
- Zhao P, Wang GHH, Dong ZMM, Duan J, Xu PZZ, Cheng TCC, Xiang ZHH, Xia QYY. Genome-wide identification and expression analysis of serine proteases and homologs in the silkworm *Bombyx mori*. *BMC genomics.* 2010; 11:405. [PubMed: 20576138]
- Zou Z, Evans JD, Lu Z, Zhao P, Williams M, Sumathipala N, Hetru C, Hultmark D, Jiang H. Comparative genomic analysis of the *Tribolium* immune system. *Genome Biology.* 2007; 8.
- Zou Z, Lopez DL, Kanost MR, Evans JD, Jiang H. Comparative analysis of serine protease-related genes in the honey bee genome: possible involvement in embryonic development and innate immunity. *Insect Mol Biol.* 2006; 15:603–614. [PubMed: 17069636]

Identify 337 SP/SPH genes, improve 117 of their models, and classify them into 110 CLIPs, 100 GPs/GPHs and 127 SP(H)s

Analyze the domain organization of CLIPs A–E and identify 13 other types of putative regulatory domains in 18 SP(H)s

Reveal relationships among phylogenetic tree positions, chromosomal locations and expression patterns of 159 SPs/SPHs

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Fig. 1.

Domain organization of 128 multi-domain SPs and SPHs in *A. gambiae*. Signal peptide and other structural elements (see symbols in inset) were predicted as described in *Section 2.2*. The schematic diagrams are not drawn to scale.

Fig. 2.

Phylogenetic relationships of the 110 CLIPs (A), 100 GP(H)s (B) and 127 SP(H)s (C) in *A. gambiae*. Entire sequences of the proteins in each group were aligned and the phylogenetic tree was constructed using MrBayes as described in *Section 2.3*. Probability values for branches are indicated near the branching points, with “*” representing 100. Subtrees of similar sequences in various colors are assigned with a series of group IDs, which start with L for CLIPs, T for gut protease (homologs), or P for serine protease (homologs). Those on red background with the 2nd letter in capital represent reliable monophyletic groups. Other branches are assigned with IDs on blue background with the 2nd letter in lowercase. Chromosomal location (Fig. 2) and expression (Fig. 4) IDs are listed next to gene names, in various colors based on the IDs. In panel A, the letter A, B, C, D, or E in CLIP gene names are marked with different colors. In panel C, SP1xx are colored red and SP2xx blue. CLIP32, SP44, SP142, SPH12, SPH42, and SPH113 gene are not identified in the genome assembly and, thus, without location IDs.

Fig. 3.

Chromosomal locations of the 331 genes coding for *A. gambiae* 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. Note that CLIP32, SP44, SP142, SPH12, SPH42, and SPH113 genes are not found in the genome assembly. Names of CLIPs (red), GP(H)s (green) and SP(H)s (blue) are linked to their locations by straight lines. Adjacent genes with high sequence similarities are grouped by lines in the same color and marked by a series of location IDs on red background for different chromosomal segments. Regions in between are labeled with IDs on blue background. In these IDs, the 1st letter (1–6) represent chromosome (Ch.) 2L, 2R, 3L, 3R, X and UNKN (unknown), respectively, and the 2nd letter in capital for gene clusters or in lowercase for other regions.

Fig. 4.

Transcript profiles of the 110 CLIPs in *A. gambiae*. The CLIP mRNA levels in 45 types of the tissue samples, as represented by $\log_2(\text{FPKM}+1)$ values, are shown in the hierarchically clustered gradient heatmap from blue (0) to maroon (10). The values, rounded to the closest integers and, if equal to 10, converted to A, are used to label the color blocks. Subtrees of genes with similar expression patterns are in different colors and assigned with group IDs beginning with C (for CLIP) on red background. The 2nd letter in capital indicates a reliable monophyletic group. Group IDs with the 2nd letter in lowercase are assigned to other branches. The letter A, B, C, D, or E in CLIP gene names are marked on the left with different colors. See Fig. 4 legend for descriptions of library types, chromosomal location IDs, and phylogenetic tree IDs.

Fig. 5.

Transcript profiles of the 100 GP(H)s in *A. gambiae*. mRNA levels of the gut serine proteases or their homologs in 45 types of tissue samples, represented by $\log_2(\text{FPKM}+1)$ values, are shown in the hierarchically clustered gradient heatmap from blue (0) to maroon (10). The values, rounded to the closest integers and, if 10, converted to A (10), B (11) ... G (16), are used to label the color blocks. Subtrees of genes with similar expression patterns are in different colors and assigned with group IDs GA through GE (G for gut) on red background. In the library names, E, egg, L, larva, P, pupa, A, adult; h, hour, d, day; F, female, M, male; An, antenna, S, salivary gland; NB, no blood feeding; B, blood fed; C, control, and I, infected. Chromosomal location IDs (loc.) (Fig. 2) and phylogenetic tree IDs (tree) (Fig. 3) are labeled on the right next to the gene names in different colors based on their IDs.

Fig. 6.

Transcript profiles of the 127 SP(H)s in *A. gambiae*. The serine protease (homolog) mRNA levels in 45 types of tissue samples, as represented by $\log_2(\text{FPKM}+1)$ values, are shown in the hierarchically clustered gradient heatmap from blue (0) to maroon (10). The values, rounded to the closest integers and, if equal to 10 and 11, converted to A and B, are used to label the color blocks. Subtrees of genes with similar expression patterns are in different colors and assigned with group IDs beginning with S (for serine in SPs or SPHs) on red background. The 2nd letter in capital indicates a reliable monophyletic group. Group IDs with the 2nd letter in lowercase are assigned to other branches. SP1xx are colored red and SP2xx blue on the left. See Fig. 4 legend for descriptions of library types, chromosomal location IDs, and phylogenetic tree IDs.

Table 1Key structural features of the 337 SP-related proteins in *A. gambiae*

name	T-C-E	activation cutting site	specificity	domain
CLIPA1	LA3DCA	VVSH*SGEN	na	cH
CLIPA2	LA3DCA	VVQR*TINE	na	2cH
CLIPA3	LA6bCC	LDVR*IVSN	na	cH
CLIPA4	LA3DCA	EQNK*FNEI	na	cH
CLIPA5	LA3DCK	KNGR*GVID	na	cH
CLIPA6	LA3DCA	IGFR*ITGS	na	cH
CLIPA7	LA3DCA	VGFR*ITGD	na	cH
CLIPA8	LA3aCA	IMLR*FGEE	na	cH
CLIPA9	LA3aCb	?	na	cH
CLIPA10	LA1qCE	KNPV*YVDG	na	cH
CLIPA12	LA3DCA	VGFR*IGAG	na	cH
CLIPA13	LA3DCf	IDVR*VGEE	na	cH
CLIPA14	LA3DCA	IDIR*VGED	na	cH
CLIPA15	LA2dCE	RKGR*VVGG	na	5cH
CLIPA19	LG2GCK	PRAL*STD	na	cH
CLIPA20	LA3aCf	?	na	cH
CLIPA26	LA2cCK	DEVH*LDFF	na	cH
CLIPA27	LA5aCK	ISQA*VAGP	na	cH
CLIPA28	LA3aCA	IGLR*AGLD	na	cH
CLIPA30	LA3DCA	IEPR*LLND	na	cH
CLIPA31	LA3DCG	ISLR*LNPE	na	cH
CLIPA32	LA3DCJ	ISLR*LNPE	na	cH
CLIPB1	LG2GCA	QMDR*IVGG	T(DGG)	cP
CLIPB2	LG2GCK	VTDR*IIGG	T(DGG)	cP
CLIPB3a	LG2GCd	LADR*VIGG	T(DGD)	cP
CLIPB3b	LG2GCA	LADR*VIGG	T(DGG)	cP
CLIPB4	LG2GCA	LTDR*VIGG	T(DGG)	cP
CLIPB5	LG2mCA	TSDR*IFGG	T(DGG)	cP
CLIPB6	LG2GCA	YTDR*IIGG	T(DGG)	cP
CLIPB7	LG2dCA	LLMK*QKHS	na	cH
CLIPB8	LG2FCA	YVAK*IRGG	T(DGG)	cP
CLIPB9	LG2FCA	IGMR*IYGG	T(DGG)	cP
CLIPB10	LG2FCA	LADR*IIGG	T(DGG)	cP
CLIPB11	LF4GCA	SEDR*IAFG	T(DGG)	cP
CLIPB12	LF4GCK	SNSR*ATWT	na	cH
CLIPB13	LF1aCA	TVNR*IAHG	T(DGG)	cP
CLIPB14	LG3aCK	FGVR*IIGG	T(DGG)	cP
CLIPB15	LG4jCA	LADR*IYFG	T(DGG)	cP
CLIPB16	LD4jCC	QCYR*GDFS	na	cH

name	T-C-E	activation cutting site	specificity	domain
CLIPB17	LF2cCK	LLVK*IQDG	C(NGS)	3cP
CLIPB18	LF4GCK	SADR*MAYG	T(DGG)	cP
CLIPB19	LG2GCK	TNTR*LIGS	T(EGG)	cP
CLIPB20	LE3HCK	MDSG*SIGR	T(DGV)	cP
CLIPB36	LG2GCA	TLRK*DTLT	na	cH
CLIPB41	LD2mCf	VNTR*IIGG	T(DGG)	cP
CLIPB42	LF4GCG	RVQL*IAYG	C(AGG)	cP
CLIPB43	LF4GCJ	?	na	cH
CLIPB44	LF4GCC	TDDK*ISFG	T(DGG)	cP
CLIPB45	LF3aCJ	IEEK*IANG	T(DGG)	cP
CLIPB46	LG4jCK	ADEF*SFDS	T(DGG)	cP
CLIPB47	LF2mCA	TVNK*IAFG	E(LGG)	5cP
CLIPC1	Lj4dCf	AVEL*IVDG	T(DGA)	cP
CLIPC2	Lj2mCK	DQNL*IVGG	T(DGG)	cP
CLIPC3	Lj2mCA	VVKL*IVGG	T(DGA)	cP
CLIPC4	LH5BCA	IIDH*ISGR	T(DGG)	cP
CLIPC5	LH5BCJ	LAYH*IIAG	T(DGS)	cP
CLIPC6	LH5BCA	LTFH*IIDG	T(DGS)	cP
CLIPC7	Lj2mCA	RQLK*GKGR	na	HcH
CLIPC9	LH1aCA	KQFQ*IMHG	T(DGG)	cP
CLIPC10	LH5BCf	LADH*IFNG	T(DGG)	cP
CLIPC12	LE3HCK	PSWN*VWSD	T(DGG)	cP
CLIPC13	LE3HCK	PLAL*VFSK	T(DGG)	cP
CLIPC14	LE3HCK	PLAL*VYSE	T(DGG)	cP
CLIPD1	LC2dCK	QLSK*IAGG	T(DGG)	cP
CLIPD2	LC4aCf	DTER*IVGG	T(DGG)	cP
CLIPD3	LC2cCE	SSGR*IVGG	T(DGG)	cP
CLIPD4	LC2ECf	EHNH*VVGG	T(DGG)	cP
CLIPD6	LC2ECA	QHNR*VVGG	T(DGG)	2cP
CLIPD7	LB4ECE	LQKR*IIGG	T(DGG)	cP
CLIPD8	LB2dCE	PETR*IVGG	T(DGG)	cP
CLIPD9	LB4ECE	RTNR*IVGG	T(DGG)	cP
CLIPD11	Lj1aCC	DYYL*IYPI	T(DGA)	PHcH
CLIPD12	LB4ECE	KSGR*VVGG	T(DGG)	cP
CLIPD13	LB4ECE	AQRR*IVGG	T(DGG)	cP
CLIPD14	Lj4dCf	?	na	2HcH
CLIPD20	LC2ECG	THTR*VVGG	T(DGG)	cP
CLIPD22	LB4ECE	PEPR*IVGG	T(DGG)	cP
CLIFE1	LA4aCA	VISK*TPVV	na	3cH
CLIFE2	LA3DCH	VNDR*VSGT	na	3cH
CLIFE4	Lm3aCC	RLRK*GERV	na	2HcH
CLIFE5	Lm3aCC	IRQR*MSNG	T(DGG)	PHcH

name	T-C-E	activation cutting site	specificity	domain
CLIP6	LA3DCK	LGKR*FVPD	na	cH
CLIP7	LA3DCG	RNDH*GIGF	na	cH
CLIP8	Lr3aCb	ERRR*IESA	na	2Hc
CLIP9	LN3GCG	IHAR*LQNK	T(DGG)	PHcH
CLIP10	Lm3aCC	GNTG*IVGP	T(DGG)	PHcH
CLIP11	LK4HCK	MLYR*FQQN	T(DGG)	PHcH
CLIP12	Lj2mCC	QEAK*QGKP	na	2HcH
CLIP13	Lj2mCH	LGFF*IFGG	T(DGG)	PHc
CLIP14	LQ6aCG	ADER*IPPS	na	2HcH
CLIP15	LK4dCA	LPLP*AFGR	T(DGG)	PHcH
CLIP16	LK4HCC	RYDF*SVNR	na	2HcH
CLIP17	LK4HCC	RYDF*SVNR	na	2HcH
CLIP18	Lj4jCK	SQCL*IFGG	T(DGG)	PHc
CLIP19	LP3aCG	YADI*SVGF	T(DGG)	PHcH
CLIP20	LQ6aCG	ADKL*IPPF	na	2HcH
CLIP21	LP3DCE	PDQF*ISSG	T(DGG)	PHcH
CLIP22	LN3GCG	IHAR*LQNK	T(DGG)	PHcH
CLIP23	LN3GCG	INAR*LQNN	na	2HcH
CLIP24	LN3GCH	IHTR*LQNN	T(DGG)	PHcH
CLIP25	LP6aCG	SSAT*AIGN	T(DGG)	PHcH
CLIP26	LQ3GCH	RLDA*SKFI	na	2HcH
CLIP27	LQ3GCG	QRLK*DSKL	na	2HcH
CLIP28	LQ3GCG	QRLK*DSKL	na	2HcH
CLIP29	LK4HCK	MLYR*FQQN	na	2HcH
CLIP30	Lr3aCE	VRQR*MENN	T(EAD)	PHc
CLIP31	Lr3aCG	ISK*AGKS	na	2Hc
CLIP32	LNnaCH	IHAR*LQSK	T(DGG)	PHcH
CLIP33	LN6aCH	IHTR*LQNN	T(DGG)	PHcH
CLIP34	LN3GCG	LQNK*QQIS	na	cH
GP1	TD1aGD	DQSK*IVNG	C(GSG)	P
GP2	TA3cGD	WFPR*IIIG	T(DGG)	P
GPH3	TC1aGD	?	na	H
GPH4	TC4jGE	EVRS*IVGG	na	H
GP5	TC1kGE	QGAR*IVGG	E(GID)	P
GP6	TD4CGE	AGKR*IVGG	C(GSG)	P
GP7	TD4CGE	AGKR*IVGG	T(DGG)	P
GP8	TD6bGE	AGKR*IVGG	T(DGG)	P
GP9	TD4CGE	VGHR*IVGG	T(DGG)	P
GP10	TD4CGE	SGHR*IVGG	T(DGG)	P
GP11	TD4dGE	RRAQ*IVGG	T(DAG)	P
GP12	TA2AGE	SRPK*IVGG	C(SGG)	P
GP13	TA2AGE	IRPP*IIIEG	E(SGI)	P

name	T-C-E	activation cutting site	specificity	domain
GP14	TD4jGE	KTYR*IVGG	C(GGD)	P
GP15	TC1PGE	DGYR*VVGG	C(GGD)	P
GPH16	TE1qGE	SENV*TANG	na	H
GP17	TC1aGE	IWNR*IVGG	E(GID)	P
GPH18	TC1cGE	VVGR*VADG	na	H
GP19	TE1kGE	GGMR*VVNG	C(SGA)	P
GP20	TA2AGE	TVNR*IIGG	C(SGN)	P
GP21	TC1PGE	YVNR*VVGG	C(GGD)	P
GP22	TC1PGE	YVNR*VVGG	C(GGD)	P
GP23	TC1PGE	YVNR*VVGG	C(GGD)	P
GP24	TD4CGE	NGHR*VVGG	T(DGG)	P
GP25	TC1qGE	DSGR*IVGG	E(GAD)	P
GP26	TD4CGE	VGQR*IVGG	T(DGG)	P
GP27	TD1aGC	TTQR*IVGG	T(DRG)	P
GPH28	TG1DGC	ENRL*ATYG	na	H
GPH29	TG1DGC	TNRL*ATNG	na	H
GP30	TB6aGC	HSGR*IVNG	T(DAS)	P
GP31	TB6bGC	QSGR*IVNG	T(DSA)	P
GP32	TB3aGC	QSGR*IVNG	T(DSA)	P
GP33	TB3aGC	QSGR*IVNG	T(DTA)	P
GPH35	TG1DGC	NQVR*IVSE	na	H
GPH36	TE2HGC	?	na	H
GPH37	TG1DGC	ENRL*STYG	na	H
GPH38	TG1DGC	PSPL*ATDG	na	H
GP39	Tf2dGC	PNRR*IVNG	C(AGT)	P
GP40	Tf1RGC	PNRR*IVNG	C(SGT)	P
GP41	TH1GGC	RTNR*ITNG	E(SVS)	P
GP42	TH1GGC	PSHR*ITNG	C(SGS)	P
GP43	TH1GGC	PSHR*ITNG	C(SGS)	P
GP44	TB2BGC	RADR*IVGG	T(DGG)	P
GP45	TE1aGC	LLAK*VVNG	E(NVS)	P
GPH46	Tj1FGB	PSAR*IVGG	na	H
GPH47	TE1aGB	RNSR*IVNG	na	H
GPH48	TA3cGB	VSPF*LVGG	na	H
GP49	TH1FGB	PTHR*IVNG	C(SGS)	P
GPH50	Tj1NGB	NNQR*VFGG	na	H
GP51	Tj1RGB	DNAR*IVNG	C(GGS)	P
GP52	TB2BGB	YNGR*IVGG	T(DGG)	P
GPH53	TA1LGB	FLPF*IAGG	na	H
GPH54	TA1LGB	AGPR*VTGG	na	H
GPH55	TA1LGB	RSPR*LIGG	na	H
GP56	TB1qGB	TSGR*IVGG	T(DGG)	P

name	T-C-E	activation cutting site	specificity	domain
GPH57	TB2BGB	FQGR*IFGG	na	H
GP58	TH1FGC	PSHR*VTNG	C(SGS)	P
GP59	TC3FGB	WAGR*IVGG	C(GGD)	P
GP60	TH1NGB	PSQR*IVNG	E(SVS)	P
GPH61	TK1HGB	PDRR*INNG	na	H
GP62	TC3FGB	WEGR*IVNG	C(GGD)	P
GP63	TB2BGB	SLKK*IVGG	T(DGA)	P
GPH64	TC3FGB	PNGR*IVGG	na	H
GP65	TH1FGB	PTHR*ITNG	C(SGS)	P
GP66	TH1NGB	PSAR*IVNG	E(SVS)	P
GPH67	Tj1NGB	PRGR*VVGG	na	H
GPH68	TM1NGB	KTPR*IRGG	na	H
GPH69	TK1HGB	PSSR*ISNG	na	H
GP70	TB2BGB	QNGR*IVGG	T(DGG)	P
GP71	TH1FGB	PSHR*ITNG	C(SGS)	P
GP72	TB2BGB	FSGR*IVGG	T(DGG)	P
GP73	TH1FGB	PSHR*VVNG	C(SGS)	P
GP75	TC2mGB	KGGR*IVGG	E(GVD)	P
GPH76	TC3FGB	WKGR*IVGG	na	H
GPH77	TK1HGB	RSSR*ISDG	na	H
GPH78	TA1LGB	TLLR*DTIW	na	H
GP80	TB6bGC	GLGR*IVNG	T(DGS)	P
GP81	TH1NGB	PTRR*ITNG	E(SVS)	P
GP82	TH1FGB	PSHR*IVNG	C(SGS)	P
GP83	TB2BGC	VTGR*IFGG	T(DGG)	P
GPH84	TA1LGB	TVVR*NVGF	na	H
GP85	Tf1NGC	SLSK*VAGG	C(NGG)	P
GP86	TG1RGC	PSGR*ITNG	C(SGT)	P
GPH87	TC2mGB	FSPR*IAGG	na	H
GP88	TB2BGC	ATGR*IVGG	E(SLA)	P
GPH89	TK1HGC	RSQR*ILNG	na	H
GPH90	TK1HGC	LNAR*ISGG	na	H
GPH91	TK1HGB	PDRR*INNG	na	H
GP92	TA4FGB	PSGR*VVGG	C(SGS)	P
GPH93	TA4FGB	PQQR*LIGG	na	H
GPH94	TK1HGB	RTGR*INNG	na	H
GPH95	TK1HGB	RSAR*IADG	na	H
GP96	TD3cGD	NMAR*VVGG	T(DGS)	P
GP97	TE4jGD	RSSR*IVNG	E(GVS)	P
GP98	TB2BGC	PSPF*IFGG	T(EGG)	P
GPH99	TA4FGA	PERR*IFGG	na	H
GP100	TB2BGC	KSAR*IVGG	T(DGS)	P

name	T-C-E	activation cutting site	specificity	domain
GPH101	TG1DGC	GNKL*ATDG	na	H
GPH102	TG1DGC	PYTY*SATY	na	H
GP103	TD4CGE	SNHR*IVGG	T(DGG)	P
SP1	PC2cSq	QQQR*IVGG	T(DGG)	P
SP2	PD1BSG	PMAL*IIGG	E(GAD)	P
SP3	PD1BSG	AANY*IVDG	E(GVD)	P
SP4	PD1BSG	PLAP*IIGG	E(GSD)	P
SPH5	PG1aSB	?	na	H
SPH6	PG2GSA	APER*LITS	na	H
SP7	PG2HSA	VVDL*IVGG	T(DGA)	P
SPH8	Pk1Jsk	FDCG*VRGQ	na	H
SPH10	PA4jSN	QSGR*ILNG	na	H
SP11	Ph6bSN	GSQR*IIGG	T(DGG)	P
SPH12	PAAnaSN	QSGR*IVNG	na	H
SP13	PA6aSN	QSGR*IING	T(DTA)	P
SPH14	PA3aSN	RSGR*ITNS	na	H
SPH15	PA6bSN	QSGR*IFNG	na	H
SPH17	PF2NSE	ASFR*VLGG	na	H
SP18	PF2NSE	TSSR*IVGG	T(DGG)	P
SP19	PG2HSA	VVQL*IVGG	T(DGA)	P
SP20	PA3aSN	QSGR*IVNG	T(EVA)	P
SP21	PF2NSc	DASR*IVGG	T(DGG)	P
SP22	PF2NSf	QEIR*IVGG	T(DGG)	P
SPH23	PC2qSf	RNPK*IMHG	na	H
SP24	PF2NSf	NNSK*IVGG	T(DGG)	P
SP25	PF2NSB	RLTR*IVGG	T(DGG)	P
SP26	PF2NSf	INER*IVGG	T(DGG)	P
SPH27	PG3BSR	VTGV*VSFG	na	H
SP28	Pm2cSL	YNNL*ILGG	C(GAT)	P
SPH29	PP4dSN	GRRK*VQTN	na	H
SP30	Ph1GSN	PFQR*ITNG	E(SVS)	P
SP31	Ph1GSN	STDR*ITNG	E(SVS)	P
SP33	Ph1GSN	STDR*VVNG	C(SGS)	P
SPH34	PF2NSm	HGQR*IVAG	na	H
SPH35	PA3aSN	QSGR*IING	na	H
SPH36	PA3aSN	QSGR*IING	na	H
SPH37	PA6aSN	QSGR*LVNG	na	H
SPH38	Ph1DSN	GNWL*DTYG	na	H
SP39	PA6aSL	HSGR*IVNG	T(DGS)	P
SP40	PA6aSL	HSGR*IVNG	T(DGS)	P
SPH42	PEnaSL	AFNR*TLFN	na	H
SP44	PDnaSA	PTDA*IVRG	T(DGG)	P

name	T-C-E	activation cutting site	specificity	domain
SP45	PG3mSR	MDNR*IVGG	T(DGA)	P
SPH46	PE2PSA	?	na	H
SPH47	PE2PSA	?	na	H
SPH48	PE2PSA	EDGR*IFEN	na	H
SP49	PA3aSL	HSGR*IING	T(DVS)	P
SPH50	PJ3aSE	RDKR*MAAG	na	H
SPH51	PG5aSA	EQEP*ECGD	na	H
SPH52	Pk2mSA	?	na	H
SP53	PG2HSB	VVQL*IVGG	T(DGA)	P
SP54	PL1kSA	KQGL*IFGG	E(SSV)	P
SP55	PG4HSN	FYSF*GSGG	T(DGG)	P
SP56	PG2dSN	VVGA*IVGG	T(DGG)	P
SP57	PA3aSN	SSYR*IVNG	T(DGG)	P
SP58	PA3aSN	MSFR*IVNG	T(DAG)	P
SP61	PF4dSk	NSLR*VIGG	T(DGG)	P
SP62	PD1qSc	GNDR*VVGG	C(GGD)	P
SP63	PA1mSk	PDRR*IVNG	C(GSG)	P
SP64	PF2NSR	RTNR*IVGG	T(DGG)	P
SP66	Ph4GSN	HNNE*TLGN	T(DGS)	P
SP67	Ph4GSc	QNNE*TLGN	T(DGV)	P
SP68	PG3BSH	ISER*IIAY	T(DGG)	P
SP70	PP1cSH	SEYL*IQNG	C(STT)	P
SP71	PD2dSJ	DDSK*IVGG	C(GGD)	P
SP72	PA4jSN	NGER*IVGG	T(DGG)	P
SP73	PD1kSN	LGER*IVGG	T(DGG)	P
SP74	PB4BSJ	SGNL*IVGG	T(DGG)	P
SP75	PB4BSJ	ATNM*IVGG	T(DGG)	P
SP76	PB4BSJ	WNNM*IVGG	T(DGG)	P
SP77	PB4BSJ	TANM*VVGG	T(DGG)	P
SPH78	PB1JSJ	AKRL*QIGG	na	H
SPH79	PB1JSJ	GLGQ*AQNG	na	H
SPH80	PB1JSf	AKGQ*QIGG	na	H
SP81	PD3mSJ	IKPD*VANG	E(TGI)	P
SP101	PQ3cSq	TIPL*IRKG	C(STT)	P3H
SP105	PL3aSB	REGL*VKGG	E(GSI)	PH
SPH106	PE3JSE	?	na	2H
SPH107	PE3JSE	?	na	>3H
SP108	PP3JSE	SVYL*IHNG	C(STT)	PH
SP109	PP3JSJ	SVYL*IHNG	C(STT)	P2H
SP111	PG3BSf	DLTV*AYGG	T(DGG)	PH
SP112	Pm1qSA	PMGL*VTKG	E(NAA)	PH
SPH113	PEnaSJ	KYSC*IVGQ	na	3H

name	T-C-E	activation cutting site	specificity	domain
SP114	PL3aSN	RQEL*VKGG	E(GSI)	PH
SP115	PN3KSR	SVYL*IHNG	E(SIT)	P2H
SP116	PN6bSR	TIHL*VQNG	E(SIT)	P3H
SP117	PN3KSR	SVYL*IHNG	E(SIT)	P3H
SP118	PN6bSq	SVYL*IHNG	E(SIT)	PH
SP119	PP3aSq	SIYL*IHNG	C(STT)	P3H
SP120	PG3BSA	SHIE*SYGG	T(DGG)	PH
SP122	Pm2mSA	VNLL*ITNG	C(STT)	PH
SP123	PG3BSR	GFYG*AFGG	T(DGG)	PH
SPH124	PG3BSN	VGYH*SFNA	na	2H
SP125	PP5aSL	TTYL*IHNG	C(STT)	P3H
SP126	PL3aSR	RKGL*VKGG	E(GSI)	PH
SP127	PN3KSL	TIHL*VHNG	E(SIT)	P3H
SP128	Pm2cSL	RVNL*ILGG	C(GAA)	PH
SP130	PP4dSL	SVFL*IHNG	C(GTT)	P3H
SPH132	PE3aSR	ERRK*INGM	na	2H
SP133	PA3aSL	GFLR*IVNG	T(DGS)	PH
SP134	Pm2mSA	RKVK*TIYL	E(SMT)	PH
SP135	Pk3aSA	ALPK*QPSE	C(NTA)	PH
SP136	PG4HSq	FYSF*GSGG	T(DGE)	P2H
SP137	Pm4dSR	YAKL*ILGG	C(TSA)	PH
SP138	Pk4aSN	RPIR*VAGS	C(NTA)	PH
SP139	PG3BSH	ISSY*AFGG	T(DGG)	PH
SP140	PP3cSH	SQYL*IHNG	C(STT)	P3H
SP141	PP1cSH	SEYL*IQNG	C(STT)	P3H
SP142	PPnaSH	SVYL*IHQG	C(SST)	P3H
SP143	PG4dSH	IVPA*VSGG	T(DGG)	PH
SP144	PP3JSH	TQYY*IHNG	C(SST)	P3H
SP201	PJ3ESD	RTRK*IVGG	T(DGS)	CUBP
SP202	PJ3ESm	KLNR*IVNG	T(DGS)	CUBP
SP203	PJ3ESD	KTPT*IVNG	T(DGS)	CUBP
SP204	PJ3ESD	RTPT*IVNG	T(DGS)	CUBP
SP205	PJ3ESD	RTAK*IVGG	T(DGS)	CUBP
SP206	PJ3ESD	RTSK*IVNG	T(DGS)	CUBP
SP207	PL2cSB	ANPL*VTHG	E(SSV)	GdP
SP208	Pk2cSA	IRSR*IIGG	C(SSS)	2GdP
SP209	Pk2cSE	FSHY*SING	C(VGV)	GdP
SP210	Pk2cSA	FNRL*SING	C(VGV)	2GdP
SP212	PF1sSq	ESVR*IVGG	T(DGG)	Fig1
SP213	PG1eSA	YGAR*VVHG	T(DGG)	Fig1
SP214	PA1mSJ	ATKR*IVGG	T(DGG)	Fig1
SPH216	Ph5aSf	PTSQ*NIGL	na	Fig1

name	T-C-E	activation cutting site	specificity	domain
SP217	Pk2cSA	AEAY*IIGG	C(SAV)	Fig1
SP218	PC1kSB	NMLR*IIGG	T(DGG)	Fig1
SP219	Pk1aSR	LRSR*ITDG	E(SSV)	Fig1
SPH220	Pk2cSA	LEQR*IAGG	na	Fig1

T-C-E: identification codes for the phylogenetic trees (T) (Fig. 2), chromosomal (C) locations (Fig. 3), and expression (E) profiles (Figs. 4–6), as indicated in the figure legends. Activation cleavage sites (*) are predicted based on the domain scan results for SPs, usually before the conserved IVGG motif. For most clip-domain SPs, activation cleavage sites are predicted to be next to R/K between Cys-3 and Cys-4 in the clip domain, based on the existing biochemical data. Enzyme specificity of SPs is predicted based on Perona and Craik (1995). T, trypsin; C, chymotrypsin; E, elastase; na, not applicable. Letters in parentheses are residues that determines the primary specificity. In the domain column, “c” stands for clip, “P” for serine protease or PD, and “H” for serine protease homolog or PLD. For SP201 to SPH220, detailed domain structures are shown in Fig. 1.