Available online at www.sciencedirect.com
ScienceDirect
Gene 407 (2008) 199–215
GENE
www.elsevier.com/locate/gene

Enhanced function annotations for *Drosophila* serine proteases: A case study for systematic annotation of multi-member gene families

Parantu K. Shah [a,b,c,1], Lokesh P. Tripathi [b,1], Lars Juhl Jensen [a], Murad Gahnim [c,d], Christopher Mason [e], Eileen E. Furlong [a], Veronica Rodrigues [b,f], Kevin P. White [c], Peer Bork [a], R. Sowdhamini [b,*]

[a] *European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany*
[b] *National Centre for Biological Sciences, TIFR, GKVK Campus, Bellary Road, Bangalore 560065, India*
[c] *Department of Human Genetics, The University of Chicago, Cummings Life Science Center, 920 E. 58th St., CLSC 301, Chicago, IL 60637, USA*
[d] *Department of Entomology, Agricultural Research Organization, The Volcani Center, P.O.Box 6, Bet Dagan 50250, Israel*
[e] *Program on Neurogenetics, Yale University School of Medicine, New Haven, CT, USA*
[f] *Department of Biological Sciences, Tata Institute of Fundamental Research, Homi Bhabha Road, Colaba, Mumbai 400 005, India*

Received 8 June 2007; received in revised form 9 September 2007; accepted 7 October 2007
Available online 15 October 2007
Received by J.A. Engler

Abstract

Systematically annotating function of enzymes that belong to large protein families encoded in a single eukaryotic genome is a very challenging task. We carried out such an exercise to annotate function for serine-protease family of the trypsin fold in *Drosophila melanogaster*, with an emphasis on annotating serine-protease homologues (SPHs) that may have lost their catalytic function. Our approach involves data mining and data integration to provide function annotations for 190 *Drosophila* gene products containing serine-protease-like domains, of which 35 are SPHs. This was accomplished by analysis of structure–function relationships, gene-expression profiles, large-scale protein–protein interaction data, literature mining and bioinformatic tools. We introduce functional residue clustering (FRC), a method that performs hierarchical clustering of sequences using properties of functionally important residues and utilizes correlation co-efficient as a quantitative similarity measure to transfer *in vivo* substrate specificities to proteases. We show that the efficiency of transfer of substrate-specificity information using this method is generally high. FRC was also applied on *Drosophila* proteases to assign putative competitive inhibitor relationships (CIRs). Microarray gene-expression data were utilized to uncover a large-scale and dual involvement of proteases in development and in immune response. We found specific recruitment of SPHs and proteases with CLIP domains in immune response, suggesting evolution of a new function for SPHs. We also suggest existence of separate downstream protease cascades for immune response against bacterial/fungal infections and parasite/parasitoid infections. We verify quality of our annotations using information from RNAi screens and other evidence types. Utilization of such multi-fold approaches results in 10-fold increase of function annotation for *Drosophila* serine proteases and demonstrates value in increasing annotations in multiple genomes.

Keywords: Structural properties; Immune response; Evolution; Function annotation; Gene-expression profiling; Enzyme homologues

1. Introduction

Accurate computational annotation of enzyme function and *in vivo* substrate specificities is a difficult task (Rost, 2002). The classical approaches, by homology detection, are mainly suited for predicting approximate molecular function of a protein and should be used in context with other methods (Bork and Koonin, 1998). Many proteins (and domains) encoded in

Abbreviations: SPH, Serine-Protease Homologues; FRC, Functional Residue Clustering; CIR, Competitive Inhibitor Relationships; GO, gene ontology.

* Corresponding author. Tel.: +91 80 23666250; fax: +91 80 23636662.
E-mail address: mini@ncbs.res.in (R. Sowdhamini).

[1] Both authors have contributed equally to this work.

0378-1119/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2007.10.012

eukaryotic genomes are part of multi-member protein families; they participate in variety of cellular processes and are located in different part of cells. In recognition to this, the Gene Ontology (GO) consortium annotates information about molecular function, biological process and cellular component to describe "function" of a protein product (Lewis, 2005).

In the wake of genomic era (Kanehisa and Bork, 2003) many attempts at function annotation that employ multiple data types and statistical frameworks for their integration are underway and have proved to be highly successful in annotating prokaryotic genomes (Date and Marcotte, 2005; von Mering et al., 2005). These frameworks output profiles or clusters of genes and their interpretation is largely dependent on expert knowledge. Moreover, their potential in annotating large protein families in eukryotic genomes has not been assessed. In the current study, we have utilized knowledge from structure–function analysis of 3D-structures and sequences, gene-expression profiling, text-mining, protein–protein interaction and state-of-the art bioinformatic tools to manually assign either of the three GO categories to the members of serine-protease family of trypsin fold encoded in the genome of *D. melanogaster* (Fig. 1a,b).

Analysis of whole metazoan genomes suggested that up to 15% members of all encoded enzymes families may have lost their catalytic activity (Pils and Schultz, 2004) and function annotations for them are very scarce. It is assumed that depending upon the conservation of catalytic pocket, these enzyme homologues may be able to bind specific substrate(s) and compete with active enzymes for substrates (competitive inhibitors) or may evolve newer functions (Pils and Schultz, 2004). However, no computational methods for systematically assigning competitive inhibitor relationships (CIRs) exist in the literature and only one example of function evolution has been proposed before (Pils and Schultz, 2004).

The family of eukaryotic serine proteases of the trypsin fold (referred hereafter only as serine proteases or proteases; Fig. 1a) is one of the largest enzyme families with wide species distribution (Figures S1, S2) and participate in numerous cellular processes (Barrett, 1994; Rawlings and Barrett, 1994; Zdobnov et al., 2002). Despite decades of research, prediction of *in vivo* substrate specificities of proteases remains difficult. *D. melanogaster* genome is estimated to contain roughly half the number of protein coding genes than the human genome (and other mammalian genomes) but both genomes harbour similar numbers of genes coding for serine-protease-like domains (Pugalenthi et al., 2005). Despite experimental evidences for the importance of serine proteases in *Drosophila* development, immune response and other biological processes (Table S1); there is no availability of *in vivo* substrate-specificity data or three dimensional (3D) structural information for them in Protein Data Bank (PDB) (Deshpande et al., 2005). A total of 28% of all serine-proteases-like domains encoded in the *Drosophila* genome are believed to be non-catalytic serine-protease homologues (SPHs) due to mutations in one or more catalytic residues. *Drosophila* proteases and SPHs remain poorly annotated in Flybase (Grumbling and Strelets, 2006) and Swiss–Prot (Apweiler et al., 2004; Bairoch et al., 2004).

Substrate-specificity (molecular function) annotations are plentiful for mammalian serine proteases, but they cannot be transferred reliably to *Drosophila* proteases as mammalian and *Drosophila* proteases have diverged substantially in sequence (see Supplementary material). Therefore transfer of substrate specificities using whole domain similarity would lead to annotation errors (Rost, 2002). To allow a reliable transfer of substrate specificity we have devised a composite procedure termed functional residue clustering (FRC). In its first step, FRC identifies consensus active-site residues that interact with different natural and chemical inhibitors by structure–function analysis of 3 different data sets of proteases. In its second step, FRC performs hierarchical clustering of proteases based on similarity at these residue positions and provides a quantitative measure of similarity between proteases at their active sites, thus allowing for more reliable transfer of molecular function. We benchmark the performance of FRC by applying it to known examples, and by comparing annotations provided by FRC to those provided by Ross et al. (2003) and Swiss–Prot (Apweiler et al., 2004; Bairoch et al., 2004) for *Drosophila* serine proteases. Quantitative measure of similarity at the active site between catalytically active proteases and SPHs provided by the FRC was also utilized to postulate competitive inhibitor relationships (CIRs) between pairs of proteases.

Information about cellular (biological) processes was obtained from the analysis of gene-expression data of *Drosophila* development life cycle and various published studies on immune response, using systematic analysis of functional modules in STRING (von Mering et al., 2005), and high-throughput protein–protein interaction data. We identify a large-scale and dual involvement of proteases in development and immune response. We also provide evidence that the SPH domains have evolved new roles in immune response and that separate down stream protease cascades may be involved in immune response against, bacteria/yeast infection and against parasite/parasitoid infection.

We also predicted sub-cellular localization for the *Drosophila* proteases for providing the third GO category. The verification of our annotations was carried out through literature surveys, protein–protein interaction data and analysis of results of RNAi screens. Finally, we quantitate the annotation increase by comparing our annotations to those in Flybase (Grumbling and Strelets, 2006), the community database for *D. melanogaster*. Supplementary data are also available at http://caps.ncbs.res.in/download/Enhanced_Func_SP/.

2. Methods

2.1. Sequence analysis of Drosophila serine proteases

We carried out domain architecture analysis of 201 proteins with trypsin-like serine-protease domain derived using SMART (Letunic et al., 2006). SMART suggested 144 domains as catalytically active and the remaining 57 (28%) as SPHs due to mutations in one or more catalytic triad residues (Figure S3). At least 7 proteins contain membrane-spanning helices in addition to protease-like domains. Four genes in *Drosophila* encode proteins with two serine-protease-like domains (Additional file 1).

We utilized TargetP (Emanuelsson et al., 2000) and Proteome analyst (Szafron et al., 2004) servers for sub-cellular

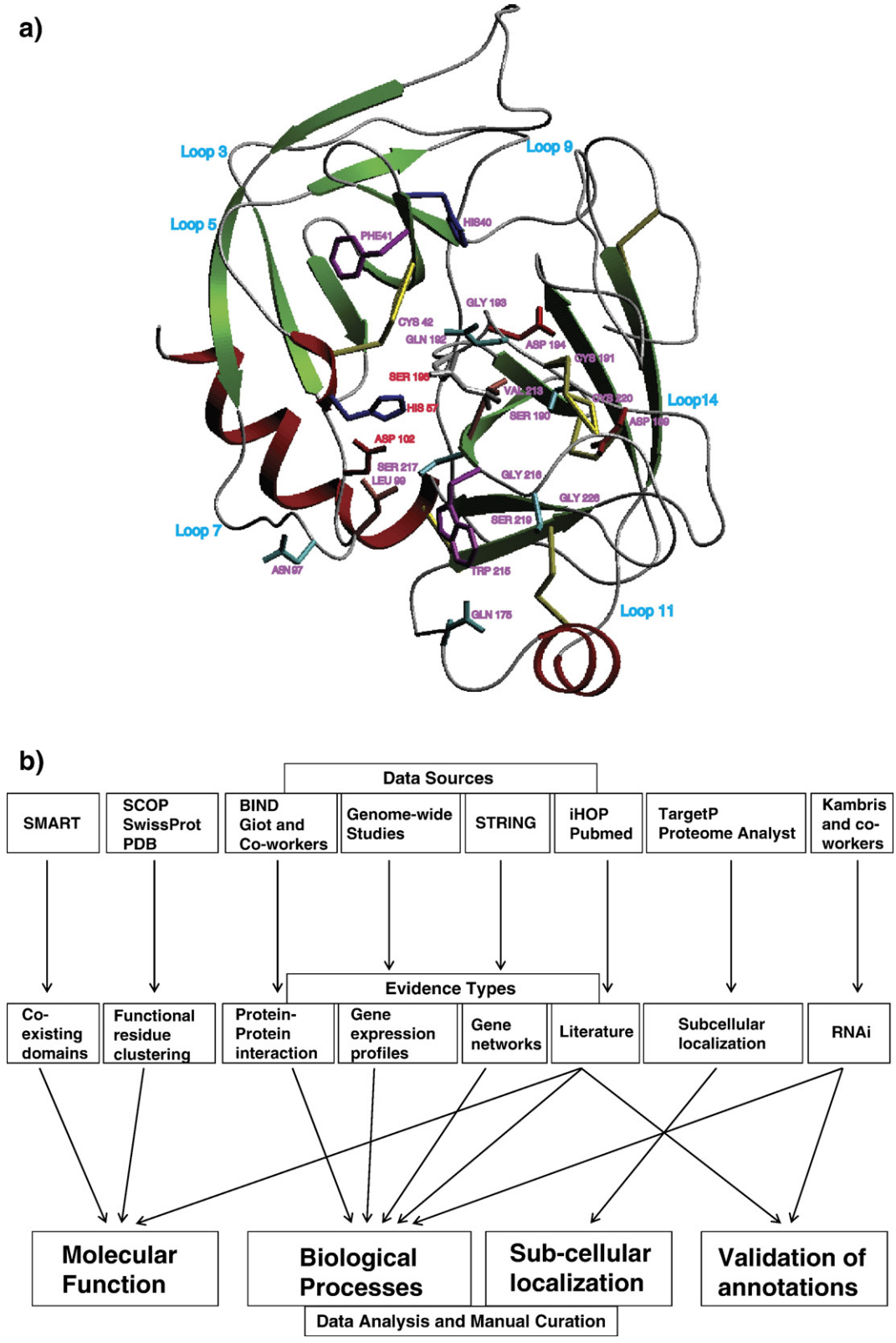


Fig. 1. a) Schematic representation of trypsin fold serine-protease domain (showing bovine chymotrypsin 5ptp-). The catalytic triad residues along with the predicted substrate-interacting residues (see text for details) and the loop regions spatially proximate to the catalytic residues are marked. The figure was prepared with Setor (Evans, 1993). b) A flowchart providing a schematic description of various data sources, and evidence types employed for function annotation of serine-protease-like proteins in *Drosophila* genome. We employed several methods for data analysis and identified molecular function, biological processes and sub-cellular location for serine proteases and SPHs encoded in *Drosophila* genome. The annotations were verified through literature surveys and data from RNAi screens where possible. References are quoted in the main text.

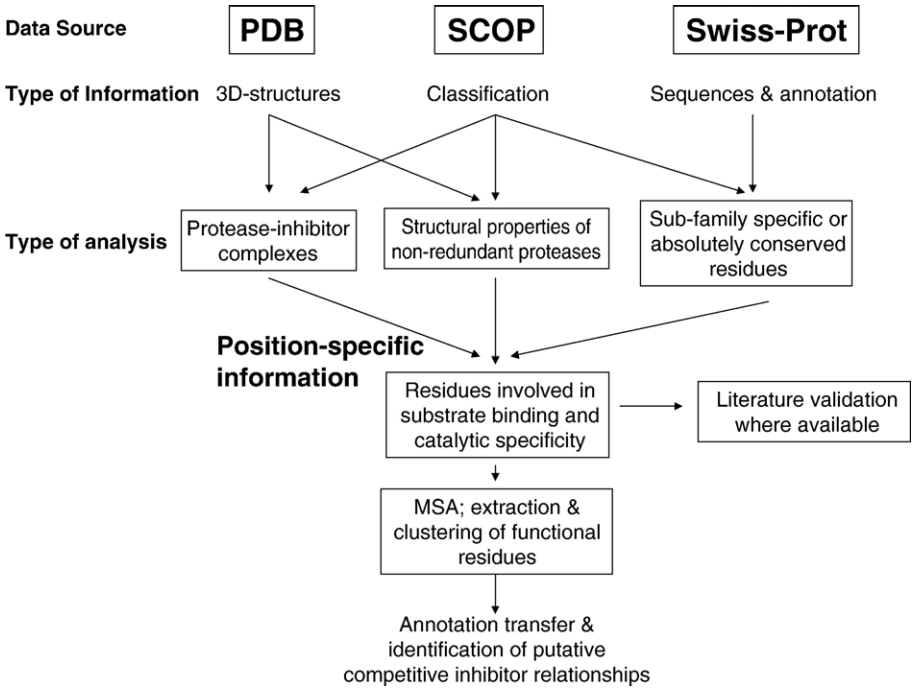


Fig. 2. A flowchart describing the steps of Functional Residue Clustering (FRC) algorithm. In order to identify functionally important residues, the first step of FRC utilizes three different type of information: 1) identification of substrate-interacting residues using 3-D structures from five protease sub-families bound to different chemical/natural inhibitors, 2) representative protease structures to study position-specific properties and 3) protease sequences that are well-annotated by Swiss–Prot (Apweiler et al., 2004; Bairoch et al., 2004) to check for conservation and sub-family specific substitutions. In the second step proteases are clustered according to hydrophobicity and patterns of absence/presence of these amino acids.

location prediction and most proteases were predicted as secreted enzymes (Additional file 1).

2.2. Description of FRC algorithm

FRC algorithm as applied to the serine proteases involves two steps: 1) identification of protease active-site residues involved in substrate binding and catalysis by structure–function analysis of proteases and 2) clustering of proteases using similarities of these residues for quantitative transfer of molecular function from characterized proteases to uncharacterized proteases (Fig. 2).

2.2.1. Identifying protease residues involved in substrate binding and conferring substrate specificity

In this study a list of serine-protease residues that are involved in substrate binding and catalysis was derived using three different datasets (Fig. 2). The first dataset consisted of 55 high-quality 3D structures from five protease sub-families, viz. trypsin, thrombin, elastase, coagulation factors and plasminogen activators, bound to different chemical/natural inhibitors. Here we exploited the knowledge that protease inhibitors bind to proteases at their active sites to inhibit catalytic activities of the proteases and utilized this dataset to identify substrate-interacting residues. Analysis of inhibitor-interacting properties of five different sub-families was expected to provide us with a consensus list of protease residues interacting with substrate/inhibitors. In crystal structures the protease residues with C^{α} atoms at distance less than 4 Å to inhibitor atoms were considered as substrate-interacting residues. For an average length of 227 amino acid residues in the serine-protease domain, each position was presumed as independent and we looked for protease residues that do not interact with inhibitors purely by chance. We calculated frequency with which each protease residue in each sub-family interacted with inhibitors and utilized binomial probability distribution to assign a *p*-value for each residue in each sub-family for interacting with inhibitors (Figure S4).

The second dataset consisted of protease structures representative of different sub-families as classified by SCOP (Murzin et al., 1995). We also utilized SCOP to provide sub-family classification for protease structures in PDB that formed dataset 1. Using structure based sequence alignment of protease structures from dataset 2 we examined various structural properties (e.g. solvent accessibility, hydrogen bonding, packing) of these residues and adjoining secondary structures. While, the third dataset consisted of aligned protease sequences from various species that are well-annotated by Swiss–Prot (Apweiler et al., 2004; Bairoch et al., 2004). The third dataset permitted us to examine the conservation and sub-family specific substitutions in substrate-interacting residues (Figures S5–S7; Additional file 2).

Using these datasets, we identified 29 inhibitor-interacting residues residing on loops 3, 5, 7, 11, 12, and 14 that had *p*-value $< 1e-5$, had conserved structural properties, were solvent accessible (except D102) and were either absolutely conserved or showed sub-family specific conservation (Figure S4; Table 1). See supplementary material for more details on the structure–function analysis.

2.2.2. Hierarchical clustering of functional residues

The substrate-interacting residues identified in the first step and those residing on adjoining loops were extracted from

Table 1
Amino acid side chains that interact with inhibitors

	Trypsin	Thrombin	Co-agulation factor	Elastase	Plasminogen activators
Loop 3	40, 41, 42	38, 40		41	
Loop 5	57	57, 60A–F		57	57
Loop 7	97, 99	97A, 98, 99	97, 98, 99	99	99
Loop 11	175	174	174		174
Loop 12	189–195	189–195	189–195	190–195	189–195
Loop 14	213–217, 219, 220, 226	213–217, 219, 220	213, 215–217, 219, 220, 226	215–218, 220, 226	213–217, 219, 220, 226

multiple sequence alignments and their hydrophobicity values and patterns of absence/presence were employed as features to perform hierarchical clustering of proteases. We employed Kyte–Doolittle hydrophobicity scale (Kyte and Doolittle, 1982) and absolute correlation co-efficient (R) as a similarity measure (Eisen et al., 1998). Since this procedure considers only substrate-interacting residues we can transfer molecular function from a well-characterized protease to an uncharacterized protease with similar active site and utilize correlation co-efficient as a quantitative measure of active-site similarity.

We carried out two clustering experiments. The first clustering experiment included non-redundant set of protease structures termed 'structures set' (Figure S8a) and the second clustering experiment included 'structures set' merged with sequence alignment of all serine proteases in *D. melanogaster* termed 'all proteases set' (Figure S9). The 'structures set' contains proteases with known 3D structures and known substrate specificities. While, 'all proteases set' contains proteases that may or may not have been characterized in terms of substrate specificities.

2.3. Evaluation of FRC algorithm performance

2.3.1. Validation of residues identified by the first step of FRC and comparison to other methods

In the first step of FRC we identified 29 substrate-interacting residues. We carried out an exhaustive literature survey to validate participation of these residues in substrate binding and in conferring specificity as described in various mutagenesis studies. The literature survey validated all the residues identified in our work (see Supplementary material).

We also compared our results to other computational approaches that identify functionally important sites in proteins. The Evolutionary Trace (ET) method (Lichtarge et al., 1996) incorporates structural information, multiple sequence alignments (MSA) and utilizes evolutionary cut-offs for defining sub-groups. ET is closest to the first step of FRC in terms of methodology. However, ET doesn't use sub-family information from SCOP and Swiss–Prot annotations. ET identified 25 residues as the 'surface patch' residues. Only 10 of these 25 residues are in agreement with our results. ET failed to identify many residues identified in our analysis (e.g., 60, 96, 97, 99, 213 and 217) whose roles are verified from mutagenesis reports published in literature. Some residues identified by ET are false positives (e.g. residues forming disulfide bridges) for being classified as functional residues. The procedure described by Hannenhalli and Russell (2000) utilizes information from MSA to predict sequence determinants of protease sub-family specificity. Their method was able to identify only 4 residues, all present in the C-terminal half of the proteases. Methods like TreeDet (Carro et al., 2006) and SPEL (Pei et al., 2006) could not handle multiple sequence alignment of serine proteases used by us.

2.3.2. Validation of the clustering procedure

We are unaware of any computational method analogous to the second step of FRC that provides a quantitative index of similarity between a pair of proteins at their functional site. We employed knowledge of functional specificities in the 'structures set' and Swiss–Prot annotations in the 'all proteases' set, to benchmark the performance of FRC. 'All proteases' set was also employed to determine the threshold value of correlation co-efficient R at which molecular function can be reliably transferred.

First, using the 'structures set', we compared the clusters obtained with FRC, with those obtained from whole domain phylogenetic trees to see whether clusters proposed by FRC proves to be more meaningful than whole domain phylogenetic trees (Figure S8a, b). Whole domain phylogenetic trees were constructed using neighbour-joining (Retief, 2000) and maximum likelihood (Guindon and Gascuel, 2003) methods and compared to FRC clusters. Both FRC and whole domain phylogeny suggest closer relationships between trypsins, thrombins and coagulation factors as compared to elastases and plasminogen activators suggesting that functionally important residues of serine-protease domains identified in the present study are sufficient to group functionally similar proteases together.

On the other hand, FRC groups β-tryptase together with trypsins, which is deep rooted in the whole domain phylogenetic trees. It is known that beta tryptase, like trypsin, preferentially cleaves peptide substrates carboxy-terminal to arginine and lysine residues (Kam et al., 1995), though it forms tetrameric structures and differs substantially from trypsins in sequence.

FRC also clusters caldecrin (1pytc) together with porcine pancreatic elastase (1brup) but not with leukocyte elastase, another elastase, as suggested by SCOP classification. Structural analysis in agreement with FRC results suggests that caldecrin is closer to 1brup and α-chymotrypsinogen (Supplementary material). In fact, it is known that caldecrin possesses characteristics of both elastase and chymotrypsin sub-families; it is closer to elastase sub-family in sequence but shares disulfide bridge pattern and catalytic specificity as in chymotrypsins (Gomis-Ruth et al., 1995). It is not inhibited by classical chymotrypsin inhibitors suggesting that it is different from chymotrypsins (Yoshino-Yasuda et al., 1998). These examples involving tryptase and caldecrin suggest that FRC can provide better insights into similarities and divergence at the functional

sites than the approaches that employ information at the whole domain level.

2.3.3. Benchmarking of FRC annotation transfer

In order to calculate specificity and sensitivity of FRC, we examined FRC clusters at different values of correlation co-efficient R for homogeneity of annotation as provided by Swiss–Prot. FRC clusters at $R>0.3$ for 'all proteases' set yielded the best value for Mathew's correlation co-efficient (MCC). At $R>0.3$ the FRC results included 27 clusters (140 proteins) with more than two proteins in each cluster. These clusters include 87 proteases (true positives) out of total 89 proteases for which Swiss–Prot annotations of substrate specificity are available. A total 22 (98 proteins) of these 27 clusters are homogeneous, 2 clusters (17 proteins) contain 1 member of different molecular function (clusters 23 and 24) and 3 (25 proteins) clusters contained more than one member of different molecular function (clusters 25, 26 and 27; Supplementary tables S2, S3). In summary, at $R>0.3$ the results include 87 true positives, 2 false negatives, and 9 false positives. Considering a total of 144 active proteases in *D. melanogaster*; FRC achieves a specificity of 0.86, sensitivity of 0.98 and MCC of 0.87 (see Supplementary material for definitions of specificity, sensitivity and MCC and Tables S2 and S3 for details on clusters).

We also compared FRC based annotation transfer from proteases in structures set and those annotated in Swiss–Prot to three substrate-specificity class (trypsin, chymotrypsin and elastase) annotations provided to 122 *Drosophila* proteases by Ross et al. (2003). There are 106 identical and 9 differing annotations between the two datasets (87% overlap). The 9 differently annotated proteins were all found in non-homogeneous FRC clusters that contain *Drosophila* gene products with differing substrate specificities (Tables S2, S3). Moreover, FRC also provided annotations for 20 proteases that were classified as 'unassigned' by Ross et al. (Tables S2, S3). Thus, FRC provides many more annotations of substrate specificity and supersedes the analysis by Ross et al., which assigns substrate specificity on the basis of only 3 C-terminal amino acid residues (Perona et al., 1995).

A careful examination of the deviations from FRC annotations provides some interesting observations. For instance, a comparison of sequence alignment of trypsin, and chymotrypsin structures with 2 *Drosophila* members (CG17234/Q9VQ99 and CG11911/Q9VPN8) from non-homogeneous clusters (clusters 22 and 24 respectively, Supplementary Tables S2, S3), through visual inspection reveals a closer relationship with trypsins, though annotated as chymotrypsins (Ross et al., 2003). Similarly, CG8215, annotated as a trypsin by Ross et al. (2003) is actually a SPH due to mutation in His-57. The co-clustering of chymotrypsins and elastases (clusters 25 and 26) may be explained due to observed similarities in their active sites.

2.4. Analysis of gene-expression data

Gene-expression data for 30 time points taken at regular intervals during the embryogenesis for ~13,000 *Drosophila* genes (Hooper et al., 2007) and from Arbeitman et al. (2002) were analysed. The dataset included 191 genes whose products contain serine-protease-like domains. The procedure of microarray construction and data acquisition has been described elsewhere in detail (Hooper et al., 2007). The expression data were normalized using non-linear normalization protocol with cubic splines (Workman et al., 2002). Peak-finding procedure is as described by Arbeitman et al. (2002). For each gene the dynamic range of gene expression was defined as the third highest ratio minus the third lowest ratio of expression. A gene was considered activated if the ratios at two successive time points fall into upper half of the dynamic range of the gene expression. The two highest and lowest ratios were discarded to avoid counting artifactual extremes.

A similarity matrix based on absolute correlation and hierarchical clustering algorithm was used to cluster genes (Eisen et al., 1998). We utilized $R>0.8$ for examining gene-expression clusters (von Mering et al., 2005). Genes that share correlated expression patterns with SPHs during *Drosophila* embryogenesis were examined systematically for physical associations curated in BIND (Gilbert, 2005) and functional associations using STRING (von Mering et al., 2005). The co-expressed genes were further mapped to their putative functional categories in GO, which provides a unified gene function classification system across genomes (Additional file 3).

2.5. Analysis of other data types

The high-throughput interaction data of Giot et al. (2003) provided 144 putative interactions for *Drosophila* proteases. Only 33 out of 144 interactions have been rated as high-confidence interactions and the rest as low-confidence interactions by BIND curators (Giot et al., 2003; Gilbert, 2005). We filtered interaction data using predicted sub-cellular location of interaction partners and expert judgement (Additional file 1). Literature surveys were performed using iHOP server (Hoffmann and Valencia, 2004) and PubMed (Fig. 1b).

3. Results

3.1. Analysis of domain architectures

Domain architecture analysis of all 201 proteins with trypsin-like serine-protease domain derived using SMART (Letunic et al., 2006) suggested 144 domains as catalytically active and the remaining 57 (28%) as SPHs due to mutations in one or more catalytic triad residues. Most of the catalytically active proteases and SPHs occur as single domain proteins while a few are present in multi-domain proteins. Protease domain resides at the C-terminal end in a majority of these multi-domain proteins (Figure S3; Additional file 1). Four genes in *Drosophila* encode proteins with two serine-protease-like domains. Interestingly, each of them is a combination of an active protease-like and a SPH domain (Additional file 1).

At least 7 proteins contain membrane-spanning helices in addition to protease-like domains. Most proteases are secreted enzymes and therefore these domains may be cleaved from

mature protein products to perform its role but otherwise be resident in membrane. For example, CG8464 encodes a gene product with a transmembrane helix, an active protease domain, and a C-terminal PDZ domain and is predicted to localize to mitochondria (Additional file 1). Its human orthologue HtrA2/Omi is also a mitochondrial protein and is involved in cytochrome c dependent apoptosis after autocatalytic processing (Hegde et al., 2002). Thus, CG8464 may also serve the same role and it is the first mitochondrial protease identified in *D. melanogaster*.

Drosophila masquerade (*mas*) gene codes for a 1047 amino acids long protein that contains an N-terminal domain containing disulfide knotted motifs and a C-terminal SPH domain (Murugasu-Oei et al., 1995). The sequence similarity between *mas* and its homologues such as GRAAL and Sp22D, extended to regions beyond the C-terminal protease domain. Several conserved Cys-rich motifs were observed in the N-terminal region (Fig. 3a). Masquerade also contains poly-threonine stretches like GRAAL and Sp22D (data not shown). We propose the possibility of chitin binding function at the N-terminal region of *Drosophila mas* and its homologues based on similarities with known chitin binding motifs (Suetake et al., 2000). An alignment including trypsin, thrombin, some active serine-proteases and mas-like sequences (Fig. 3b) suggested that the functional residues over the entire catalytic pocket have undergone mutations, suggesting that these proteins are unlikely to be active proteases or act as a competitive inhibitor of Easter (Moussian and Roth, 2005), as originally suggested (Murugasu-Oei et al., 1995). Patterns of disulfide bridges also suggested *mas* to be an intermediate sequence to both trypsin and thrombin sub-families.

Analysis of domain co-occurrences revealed that proteins containing CLIP domain invariably contain a serine-protease-like domain for all such genes in *Drosophila* genome. The consistent co-existence of these two domains suggests synergistic function for them in immune response (see below).

3.2. Identifying the determinants of substrate specificity

By performing structure–function analysis on 3 datasets of proteases; we identified 29 protease residues involved in catalysis and substrate specificity (Table 1). These residues include well-known catalytic triad residues (H57, D102 and S195) and others that line the catalytic pocket (Fig. 1a). Patterns of solvent accessibility, hydrogen bonding and side-chain packing suggested that residues of the catalytic triad are rigidly held by the adjacent secondary structures to maintain the geometry (Figures S5–S7). Loop12 and Loop14 contributed 16 out of 29 substrate-interacting residues contributing 7 and 9 residues, corresponding to the classical view that the C-terminal of proteases contains determinants of catalysis and substrate specificity. However, we also identified substrate-interacting residues reside in loops 3, 5, 7 and 11. Exhaustive survey of mutagenesis results reported in the literature supports our claims (see Supplementary material). We utilized these residues in the second step of FRC to efficiently transfer annotations of substrate specificity from well-characterized mammalian and other proteases to uncharacterized *Drosophila* proteases.

3.3. FRC is an efficient method for transfer of substrate specificity

Substrate specificity for a majority of the serine proteases in *Drosophila* genome is unknown. In the clustering experiments described in the Methods section FRC achieved a specificity of 0.86, sensitivity of 0.98 and MCC of 0.87 at $R > 0.3$. Therefore, we explored the possibility of transferring substrate specificities to *Drosophila* proteases from structurally characterized but distantly related mammalian proteases and existing Swiss–Prot annotations based on more reliable FRC approach.

FRC assigned diverse functions to *Drosophila* proteases. For example, *Drosophila* Corin was assigned plasminogen activity and CG10472 and CG8329 (with CG18180 and CG18179) were identified as collagenases (Table 2). FRC also permitted us to annotate substrate-specificity information for additional 38 active proteases in comparison with existing Swiss–Prot

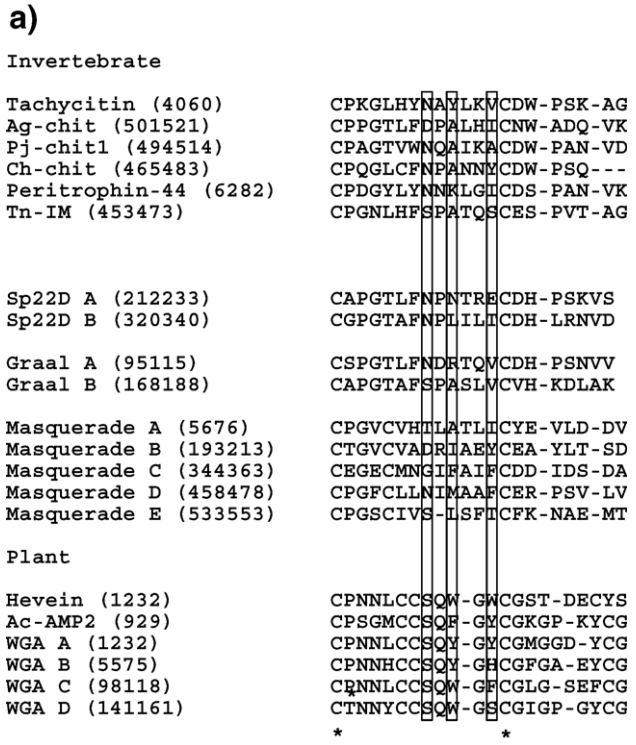


Fig. 3. a) Alignment of N-terminal chitin binding repeats in masquerade (Murugasu-Oei et al., 1995) with those in GRAAL gene product of *Drosophila melanogaster* and Sp22D protein of *Anophilis gambiae* are shown with those identified from plants and other invertebrates (Suetake et al., 2000) (see Supplementary material for details). Presence of sequence repeats are indicated by English uppercase letters (A to E). Proposed chitin binding residues are boxed and conserved cysteines are indicated with an asterisk (see Supplementary material for details). b) Restricted alignment of serine-protease domain of masquerade and homologues with other proteins involved in patterning in *Drosophila* and closest known structural homologues (bovine trypsin [5ptp-] and human thrombin [1c1uh]) of masquerade. Residues conserved in masquerade like sequences are boxed and functional residues identified are marked with an asterisk. Secondary structures are shown according to bovine trypsin (5ptp-). The loop regions are marked. Loop nomenclature was adopted from Peisach et al. (1999).

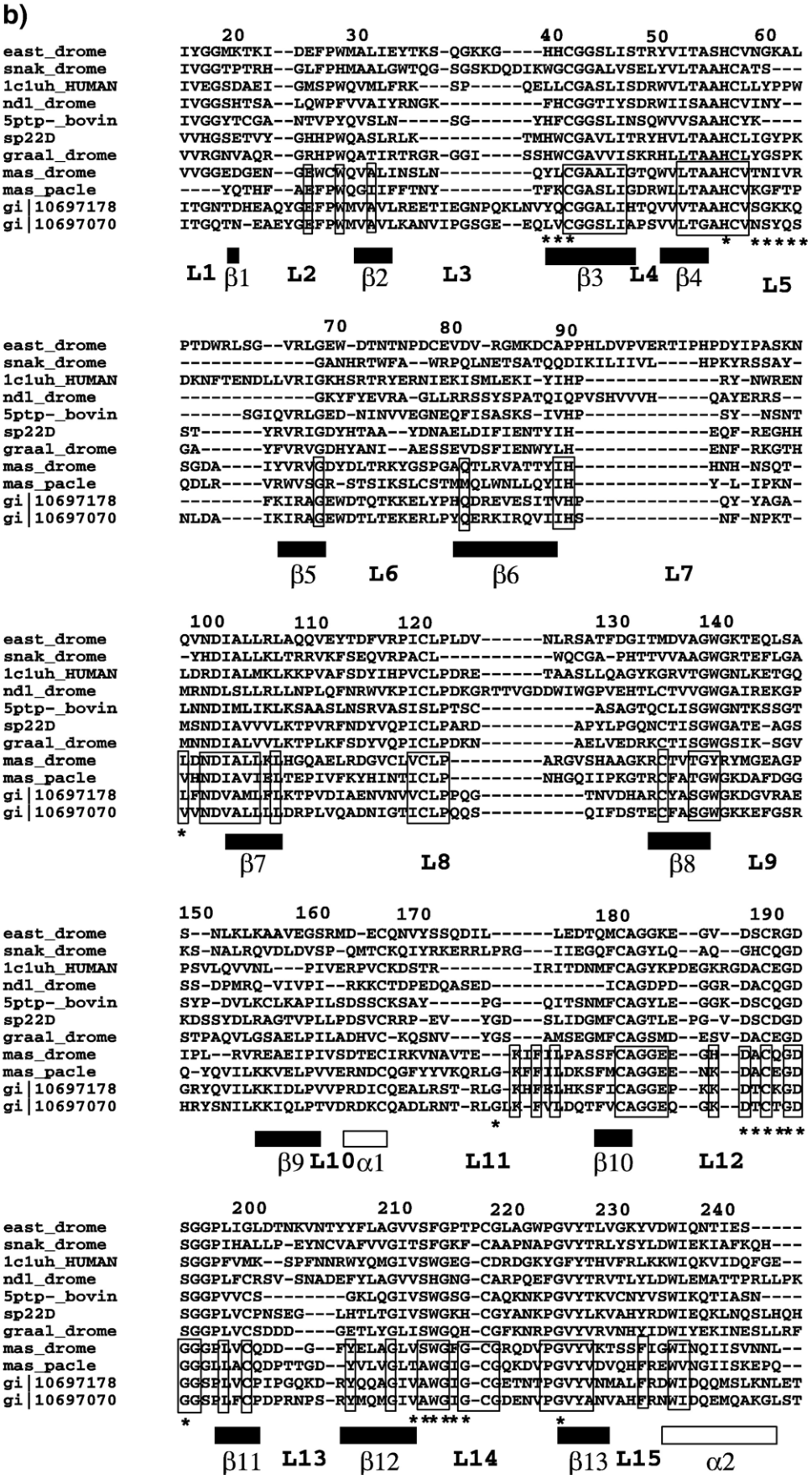


Fig. 3 (*continued*).

Table 2
Transferring function based on functional site similarities

PDB structure	Function	*Drosophila* SP	Correlation co-efficient (*R*)	Specific comments
1aola (*Homo sapiens*)	Beta-tryptase	CG16998	0.59	CG16998 implicated in immune response (Tables S5–S7)
1ekbb (*Bos taurus*)	Enteropeptidase	CG14760	0.59	
1a5ia (*Desmodus rotundus*); **1a5ha** (*Homo sapiens*)	**Single-chain plasminogen activator**	**Corin (CG2105)**	**0.56**	**CG2105 implicated in immune response (Tables S5–S7)**
1ddja (*Homo sapiens*)	Plasminogen	CG11664 (SPH)	0.55	
1agja (*Staphylococus aureus*)	Epidermetolytic toxin	CG4815	0.53	
1pytd (*Bos taurus*)	Chymotrypsin	CG5240 (CG31267; SPH)	0.53	
4chaa (*Bos taurus*)	Chymotrypsinogen	CG8528 (CG32374)	0.53	
1azza (*Uca pugilator*)	**Collagenase**	**CG7542, CG10472**	**0.52**	**CG18180 implicated in**
2hlca (*Hypoderma lineatum*)	**HL Collagenase**	**CG18180, CG18179, CG8329**	**0.5**	**immune response (Tables S5–S7)**
1elt (*Salmo salar*), 1qnj (*Sus sucrofa*), 1brup (*Sus sucrofa*)	Elastase	CG1497 (CG32523), Ser6 (CG2071), CG1304, CG9676, CG9675	0.49	All implicated in immune response (Tables S5–S7)
1aut (*Homo sapiens*), 1fxy (*Homo sapiens*), 1kig (*Bos taurus*), 1hcg (*Homo sapiens*), 1ucy (*Bos taurus*), 1clu (*Homo sapiens*)	Activated protein C; coagulation proteases	CG11530 (CG32270), CG17239	0.45	
1ton (*Rattus rattus*)	Tonin	CG6069 (SPH)	0.44	CG6069 implicated in immune response and chitin metabolism (Tables S5–S7)
1ddja (*Homo sapiens*)	Plasminogen	CG8528 (CG32374)	0.4	
1ppf (*Homo sapiens*), 1a7s (*Homo sapiens*), 1fuj	Leukocyte elastase; heparin binding protein (thrombin); myeloblastin (PR3; thrombin)	CG11529	0.38	
1aola (*Homo sapiens*)	Beta-tryptase	Ranbp11(CG33139); CG10764	0.37	

Some of the examples discussed in the text are highlighted in bold.

annotations (Fig. 4; Table S3). This represents more than 50% increase in function annotation.

3.4. Putative competitive inhibitors of active proteases can be recognised

Competitive inhibitors bind to substrates of actual proteases, thus preventing proteolysis (Jackson, 1999; Tesch et al., 2005). It is thus imperative for them to have catalytic pocket very similar to the active proteases. Since, FRC was applied to all the proteases and SPHs in the *Drosophila* genome, it is possible to assign CIRs to protease-SPH groups that share significant similarities at the catalytic pocket. Some examples of CIRs include relationships between CG5246 and *mas*, *Easter* and CG3505, and *Tequila* and CG12388 (Table 3).

Interestingly, functional sites of two pairs of SPHs (CG5390, CG4998 and CG4653, CG9673) were found to be most similar to each other. This perhaps indicates some functional redundancy in SPHs (Table 3). Further evidences supporting annotations from FRC and CIRs are discussed below.

3.5. Mining for functional interactions

STRING (von Mering et al., 2005) provides protein functional association derived from high-throughput experimental data, from the mining of other databases, literature, and from predictions based on genomic context analysis. We systematically searched STRING to retrieve high-confidence functional associations (STRING score >0.7) to assign molecular functions or biological processes to serine proteases. Out of 201 *Drosophila* serine proteases, STRING provides functional associations (clusters) for only 25 proteases (21 catalytically active proteases and 4 SPH) with high confidence (Additional file 3). These included serine proteases Gd, Snake, Easter and Nudel; members of the well-known protease cascade in Toll pathway involved in development and in immune response (LeMosy et al., 1999; Han et al., 2000; Dissing et al., 2001; Rose et al., 2003).

We found SPH CG6069, clustered with CG7663 (STRING score 0.788). CG7663 is a structural component of cuticle and contains Pfam (Finn et al., 2006) domain 'Chitin_bind_4' (PF00379), which is suggestive of its function as a chitin binding protein (Additional file 3). Thus, association of CG6069 with CG7663 indicates that it may function during chitin metabolism, possibly by recruiting substrate(s) for active proteases. It is also implicated in immune response, which may also be possible due to its role in chitin metabolism (Table S5).

Manual examination of the STRING clusters revealed correlated expression of genes with cytochrome P450 domains in 8 out of 24 clusters. A recent microarray analysis of immune challenged *Drosophila* hemocytes, reported presence of cytochrome P450 family members (Johansson et al., 2005) in high-confidence clusters indicating that these proteases may be involved in detoxification pathways. One cytochrome P450 enzyme (shade) is known to carry out the hydroxylation of

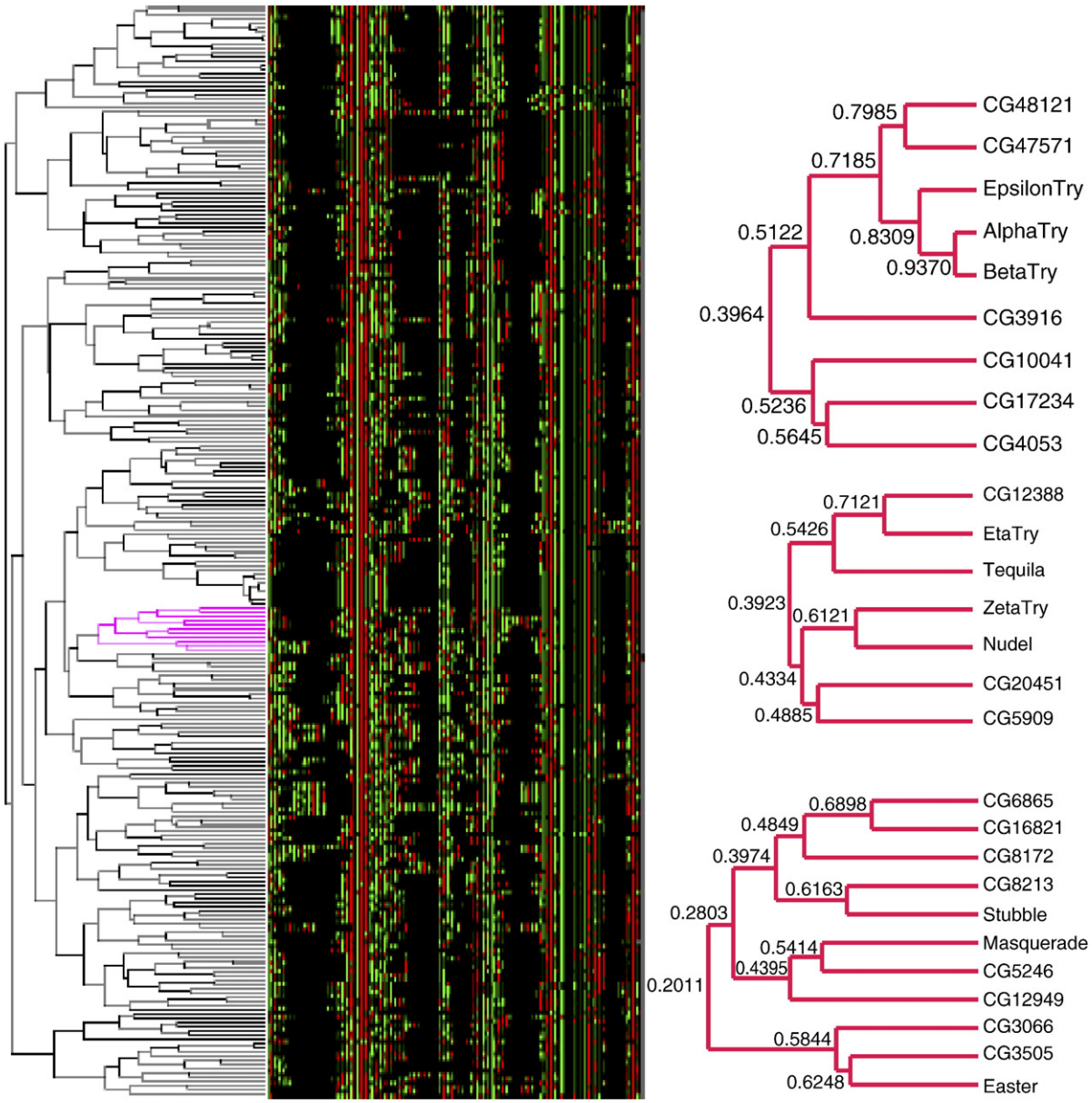


Fig. 4. Clustering of all *Drosophila* serine-protease-like proteins based on active-site similarities. The inset shows similarities at the active sites measured in terms of correlation co-efficient *R* between *Drosophila* serine-protease-like gene products masquerade and CG5246, and Easter, CG3505 and CG3066.

ecdysone to the 20-hydroxyecdysone mediating the developmental transition (Petryk et al., 2003) which triggers cascades of serine proteases (see below).

3.6. Large-scale protease expression during Drosophila embryogenesis

In order to gain further insight into the possible functional associations of *Drosophila* serine proteases, we examined the gene-expression profiles of 191 proteases (Hooper et al., 2007) during embryogenesis and profiles of 41 proteases reported in the literature (Arbeitman et al., 2002). We found 191 proteases expressed during embryogenesis, most of them at the end of the stage (Fig. 5a). While, gene expression itself may not quantify functional contribution, clearly visible *cis*-regulation of gene-expression clusters points to the involvement of proteases during development. Out of 30 time points taken during the embryogenesis, no protease is expressed continuously in even half of them (Fig. 5b; left panel) and there is a clear surge in protease expression visible towards the end of embryogenesis (Fig. 5b; right panel).

The proteases show regulated expression at various stages of embryogenesis. For example, expression of genes involved in early dorso-ventral patterning (e.g., Easter, Snake and gastrulation-defective) can be seen in the beginning of the embryogenesis. *Drosophila* gene Stubble (Bayer et al., 2003) can

Table 3
Competitive inhibitor relationship based on functional site similarities

Drosophila SPH	*Drosophila* SP	Correlation-co-efficient (*R*)	Specific comments
CG4998	CG5390 (SPH)*	0.74	Both implicated in immune response (Tables S5–S7)
CG8586	AAG22434.1 (CG18478)*	0.73	Both implicated in immune response (Tables S5–S8)
CG17477	CG17475	0.72	CG17477 co-expresses with genes possibly involved in metabolism and defense response (Additional file 3)
CG12388	EtaTry (CG12386)	0.71	
CG6639	CG8586, AAG22434.1 (CG18478)*	0.68	All implicated in immune response (Tables S5–S7)
CG3505	**Easter (CG4920)**	**0.62**	**CG3505 functions in the same immune response pathway as Easter (see text)**
CG4653	CG9673 (SPH)*	0.58	CG9673, possibly associated with drug metabolism (Tables S5–S7)
CG13527	CG15873 (SPH)*	0.58	
CG4259	CG11531, CG11532	0.56	CG4259 and CG11532 implicated in immune response (Tables S5–S7)
CG9897	Ser89E (CG31217)	0.56	CG31217 implicated in immune response (Tables S5–S7)
CG8738	CG10586	0.55	CG8738 co-expresses with genes possibly involved in signal transduction (Additional file 3)
CG12388	**Tequila (CG4821)**	**0.54**	**CG12388 co-expresses with synapsin, a neuronal phosphoprotein implicated in associative learning (see text)**
mas (CG15002)	**CG5246**	**0.54**	**See text**
CG3088	Ser99Da (CG7877), CG2229, CG18030	0.52	All implicated in immune response (Tables S5–S7)
CG9377	CG10469	0.51	CG9377 implicated in immune response (Tables S5–S7); CG10469 interacts with pip, which is implicated in immune response (Additional file 3)
CG3117	CG9898 (CG32834; SPH)*	0.49	CG3117 implicated in immune response, co-expresses with several genes implicated in metabolism and signaling (Additional file 3)
CG1632	CG4259 (SPH)*	0.47	Predicted to have low-density lipoprotein receptor activity (Additional file 1)
CG4271	CG9897 (SPH)*, Ser89E (CG31217)	0.47	CG31217 implicated in immune response (Tables S5–S7)
CG18563	AAG22436.1 (CG4793; SPH); AAG22433.1 (CG18477; SPH)*	0.46	CG18563 implicated in immune response (Tables S5–S7)
CG18557	CG6048	0.46	Both implicated in immune response (Tables S5–S7)
CG10450 (C-term; CG30286)	CG16749; CG4998 (SPH)*; Ser4 (CG8867; Jon25Bi)	0.44	All implicated in immune response; Ser4 also implicated in digestion (Tables S5–S7)
mas (CG15002)	CG12949	0.44	
CG13527	CG15873 (SPH); CG17837 (SPH); CG3795 (EG:9D2.4)	0.4	CG17837 implicated in immune response (Tables S5–S7)
CG11664	CG8528 (CG32374)	0.4	
Gd (CG1505)	CG9649, CG9645 (CG31326; CG33109; SPH)*	0.37	CG1505 implicated in immune response (Tables S5–S7)
CG9672	CG17572 (SPH)*	0.36	CG9672 implicated in immune response (Tables S5–S7)
CG4650	CG16918 (SPH)*	0.36	CG4650 implicated in immune response (Tables S5–S7)
CG17242	CG8170, CG13744	0.34	
CG8555 (CG32382)	CG16731 (C-term; CG31219)	0.34	
CG14990	CG9631	0.32	Both implicated in immune response (Tables S5–S7)

The examples discussed in the text are highlighted in bold. SPH pairs are indicated with asterisk (*).

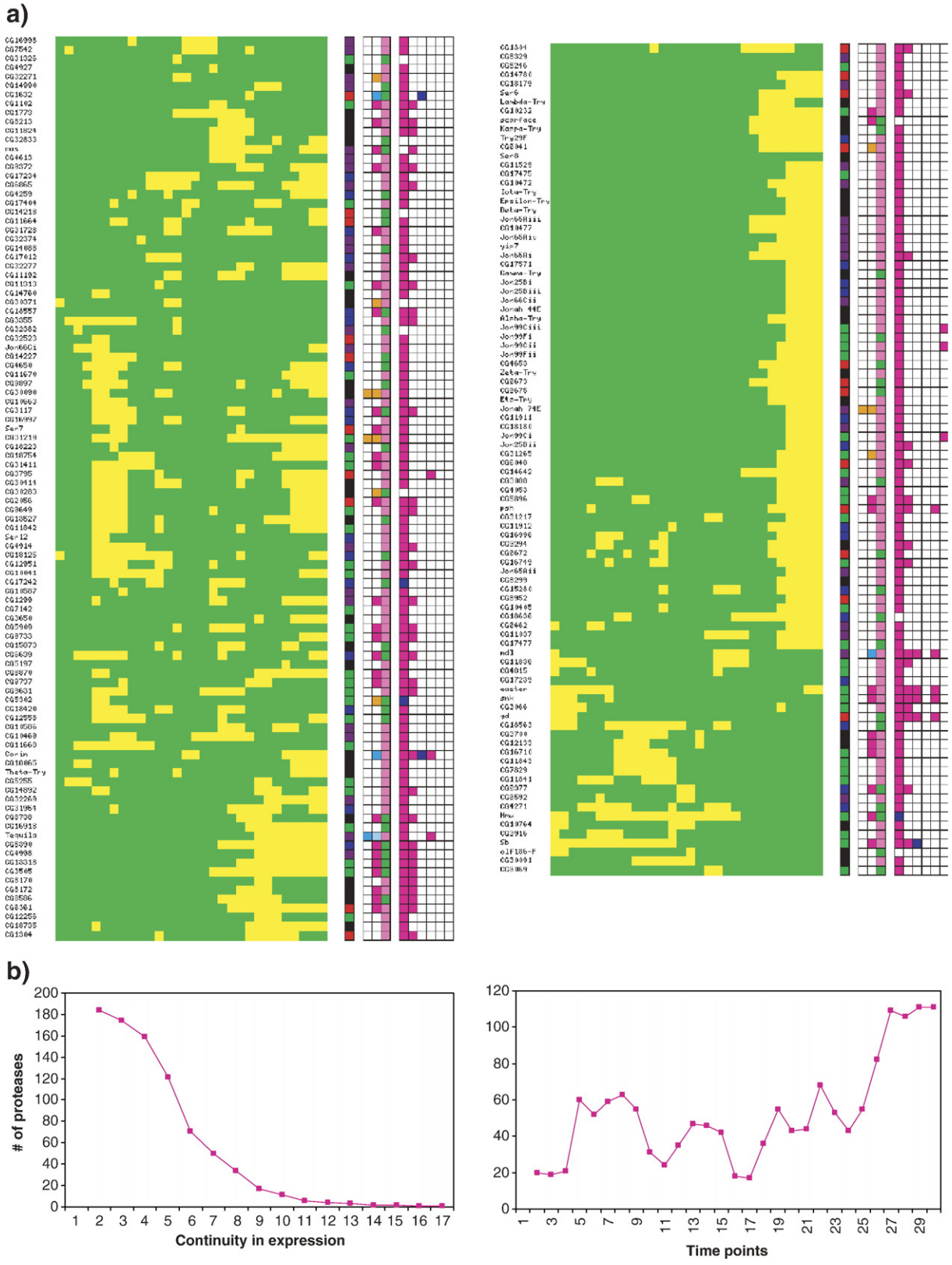
be observed in the first half of the embryogenesis. We could see protease cascade including genes of *Jonah* and Trypsin subgroups at the end of the embryogenesis (Fig. 5a) suggesting that they participate in the large-scale tissue re-modelling at the end of embryogenesis. We found conserved motifs in 5′-regions of trypsin and *Jonah* gene clusters using Meta-MEME (Grundy et al., 1997) and CIS-ANALYST (Berman et al., 2004) (data not shown). Thus, *cis*-regulation visible in the time-series data is biologically meaningful. Ecdysone-dependent regulation of *Jonah* genes at pupariation was also reported recently (Beckstead et al., 2005).

Time-series data suggest that expression of CG3066 is correlated with Easter and Snake ($R=0.69$; Fig. 5a,b). CG3066 is a monophenol monooxygenase activator involved in activation of melanization chiefly in response to fungal infection. It is also believed to be involved in a possible cross-talk between melanization and Toll pathway (Tang et al., 2006). The FRC ($R=0.62$) also suggests it to be the closest active-site homologue of Easter (Fig. 4; Figure S9; Table S3). Thus, the expression and function site profiles permit us to postulate an important role for this gene in early embryogenesis too.

A search on LIFECYCLE database (Arbeitman et al., 2002) for genes whose expression patterns are best correlated with *mas* resulted in the identification of gene products whose functions are implicated either in skeletal development (CG5656) or in defense response (CG5772) or both (CG15151).

3.7. Dual roles of proteases in development and immune response

D. melanogaster is a good model for studying mechanisms of immune response. Induction of serine proteases has been noticed in many classical as well as genome-wide studies of *Drosophila* immunity. Flybase currently annotates only 22 genes with GO terms 'serine-type endopeptidase and defense response'. Therefore, the extent of involvement of serine proteases in immunity has been under-appreciated. We carried out a survey of *Drosophila* genes implicated in immunity by



literature searches and in four published genome-wide studies of *Drosophila* immunity in a search for serine-protease genes that display significant change in expression levels (De Gregorio et al., 2001; Irving et al., 2001; Roxstrom-Lindquist et al., 2004; Wertheim et al., 2005). We found a total of 94 (out of 201; 47%) trypsin-like serine proteases involved in immune response in *Drosophila* (Table S5).

These four genome-wide surveys involve two studies of *Drosophila* innate immunity against bacteria and fungi and the remaining two assess immune response against parasites/parasitoids. There are 21 genes (termed as the core set) common between these two different types of immune responses (Table S6). These genes contain 11 members of *Jonah* gene family, and proPO-activating enzymes among others. As shown *Jonah* genes have been known to express coordinately in response to the steroid hormone ecdysone during the end of embryogenesis. Genes in core set as well as those involved in innate immune response contain many well-characterized genes belonging to protease cascade involved in *Drosophila* development and *Toll* pathway. On the other hand, proteases implicated in parasite response alone are not well-characterized.

Flybase annotates 560 genes (out of 14888 *Drosophila* genes) with the GO term 'defense response' but that includes only one SPH. Moreover, in most genome-wide studies, presence of SPHs was not distinguished from the catalytically active proteases. We found significant change in expression for 27 out of 57 SPH coding genes (47%) in those studies. This suggests specific recruitment of SPH domain in immune response (p-value 6.4×10^{-18}; Fishers exact test) and perhaps evolution of a new function for SPHs.

Similarly, we also found all 17 genes encoding for proteins with CLIP domains that are always accompanied by serine-protease-like domains in *Drosophila* genome, induced during the immune response. It also suggests specific recruitment of CLIP domain containing proteases in immune response (p-value 1.14×10^{-24}; Fisher's exact test).

3.8. Functional inferences on the basis of co-expression and physical interactions

We identified gene clusters (Bansal et al., 2007) with highly correlated expression patterns across 30 time points and containing *Drosophila* proteases and SPHs (correlation coefficient >0.8; Additional file 3; Figure S10). *Drosophila* SPHs appear to co-express and in some cases physically interact, with gene products involved in diverse physiological processes, but overall, the analysis showed no enrichment in protein–protein interactions (Additional file 3). We found a single cluster (see Supplementary material) with an over-representation of a single GO annotation (diazepam binding/valium binding activity; p-value 6.5×10^{-05}). Thus, SPH CG9673 is likely to be associated with the cluster that is involved in the drug metabolism process.

Gene expression may or may not be related to functional specificity. Some of the reasons may be 1) incomplete gene-expression data 2) differences between mRNA and protein expression levels 3) noise in the existing data and 4) lack of information about physiological cleavage sites in substrates for the enzymes. While we have tried to predict *in vivo* substrate specificity, it is beyond the scope of current analysis to study the presence of cleavage sites in proteins that belong to these gene-expression clusters.

3.9. Verification of function annotations

We utilized results from a recent *in vivo* RNAi screen for 37.5% (75) of all serine proteases for studying regulation and activation of Toll pathway by gram-positive bacteria/fungi in *Drosophila* (Kambris et al., 2006), as an independent source of information for verifying our annotations. A total of 29 out of 75 genes tested with RNAi showed significant changes in induction level or increased susceptibility to bacterial/fungal infection (p-value 4.4×10^{-13}; Fisher's exact test). This study screened for 13 of the 21 genes belonging to the core set (see above) of immune response and found 11 of them responding to the bacterial/fungal infection (high rate of true positives). An additional 5 genes annotated for immune response against bacteria/fungi only were also found responsive in RNAi studies (Table S7). The RNAi study also included 21 SPHs, of which 10 showed response against bacterial/fungal infection (p-value 3.4×10^{-13}; Fisher's exact test). Moreover, none of the eight genes annotated as responsive against parasite/parasitoid infection using gene-expression data and tested in RNAi screen for response to bacterial/fungal RNAi screen, showed any response in those screens (no false positives). Thus, the RNAi screen not only provides further evidence for our hypothesis of large-scale utilization of active proteases and SPHs in *Drosophila* immune response but also reflects on high quality of annotations provided in current analysis (Tables S5–S7).

3.10. Evidences supporting annotation of substrate specificity and CIRs

We performed literature surveys and scanned various databases to verify some annotations of substrate specificity using FRC. CG8329 (with CG18180 and CG18179) annotated as a collagenase (Table 2) is encoded by the *furry* locus which is important for maintaining integrity of cellular extensions during morphogenesis (Cong et al., 2001). Moreover, CG8329 interacts

Fig. 5. a) Illustrative gene-expression patterns for 191 *Drosophila* SP-like proteins during *Drosophila* embryogenesis. Gene names, chromosomal locations and domain assignments were taken from Ross et al. (2003). The columns from left to right show: serine gene names; Expression profile at different time points of embryogenesis (yellow — activated, green — not activated); Chromosomal localization (5 colours, each representing a single chromosome); Domain composition (three columns, LDLR, SR, C, PC and SP, SPH. SPH domains are shaded green while SP domains are shaded magenta in the third column); GO annotation (Lewis, 2005) in six columns: 1) Proteolysis and peptidolysis, 2) Immune response, 3) Development and patterning, 4) Toll pathway signalling, 5) Transport, and 6) Digestion). b) Patterns of serine-protease expression during *Drosophila* embryogenesis. The graphs show the number of serine proteases versus the continuity in gene expression over different stages of *Drosophila* embryogenesis (left panel) and the number of serine proteases expressed at any given time point in *Drosophila* embryogenesis (right panel) indicating regulated gene expression. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

with *esn* gene which is assigned the GO term of 'structural constituent of cytoskeleton' (Giot et al., 2003) (Additional file 3). RNAi knockdown for CG18180 results in cuticular tumors; a phenotype that is compatible with it being collagenase involved in cytoskeleton re-modelling (Table S7). Thus, CG8329, CG18180 and CG18179 may function in pathways for the re-modelling of cytoskeleton and could be collagenases.

CG10472 was also annotated as a collagenase (Table 2). Large-scale interaction data for *Drosophila* genome suggested interaction between CG10472 and CG31120, a dioxygenase (Additional file 3). A dioxygenase enzyme carries out hydroxylation of proline and lysine side chains in collagen and other animal glycoproteins (Aravind and Koonin, 2001). Thus, CG10472 may be a putative collagenase. *Drosophila* Corin (Table 2) whose human orthologue has known fibrinolysis activity (plasminogen activator) (Knappe et al., 2003) clustered with plasminogen activators.

Similarly, evidences supporting CIRs were also found. For example, FRC suggested SPH *mas* (a muscle attachment protein) to be competitive inhibitor of CG5246 (Fig. 4; Figure S9; Table 3). The protease CG5246 interacts with CG9319; an alpha-methylacyl-CoA racemase involved in fatty acid metabolism (Additional file 3). This is in agreement with original observations that a total loss of *mas* function causes defective muscle attachment (Murugasu-Oei et al., 1995). Thus, *mas* functions in stabilizing cell–matrix interaction and is a limiting component in the adhesion process. SPH CG12388 is considered to be a competitive inhibitor of Tequila (Table 3), a neurotrypsin implicated in long term memory formation in *Drosophila* (Didelot et al., 2006). Interestingly, during *Drosophila* embryogenesis, CG12388 was found to be co-expressed (Additional file 3) with CG3985 (Synapsin), a neuronal phosphoprotein implicated in associative learning in *Drosophila* (Michels et al., 2005). This suggests that CG12388 may be associated with cellular pathways associated with learning and memory in *Drosophila*. SPH CG3505 (CIR with Easter; Fig. 4; Figure S9; Table 3) is already known to have a function in immune response, a pathway in which Easter also functions (De Gregorio et al., 2002).

3.11. Annotating Drosophila serine proteases

Flybase (Grumbling and Strelets, 2006) is the reference database for the biological community that pursues research using *Drosophila* as a model organism. Serine proteases have been assigned the GO terms suitable for proteolysis and peptidolysis by transferring molecular function based on whole domain similarity. In some cases, SPHs are assigned GO terms that suggests its participation in proteolysis. Proteases like CG10232 (active) and CG3088 (SPH) have been assigned wrong GO terms based on dubious domain predictions. Only 43 proteases carry meaningful GO terms (Fig. 6a). By employing multi-fold approaches including literature searches, functional information could be obtained for 190 gene products containing serine-protease-like domains. Moreover, it includes putative functional associations to 35 of the 57 gene products containing SPH domains in *Drosophila* genome (Tables S8, S9). These annotations represent a ten-fold increase in annotation for serine proteases in *D. melanogaster* genome (Fig. 6a,b; Tables S8, S9), *albeit* with the help of new data, with substantial manual intervention and biological knowledge of these systems. Moreover for 30 genes all 3 GO categories were assigned, 26 of which were not present in the Flybase. Literature curation supports 10 of them, while the other 20 are completely supported by analysis reported in this work (Table S8).

FRC approach was ~10 times more powerful in function annotation than transferring information based on domains (Fig. 6b). Gene-expression data during development and immune response provided information far greater than that available from studies on individual genes. However, information provided by different approaches is mutually exclusive and therefore necessitates data-integration approaches (Jensen and Steinmetz, 2005) such as STRING. However, STRING identifies a functional module for only 25 proteases suggesting that data-integration tools for eukaryotic genomes are still evolving.

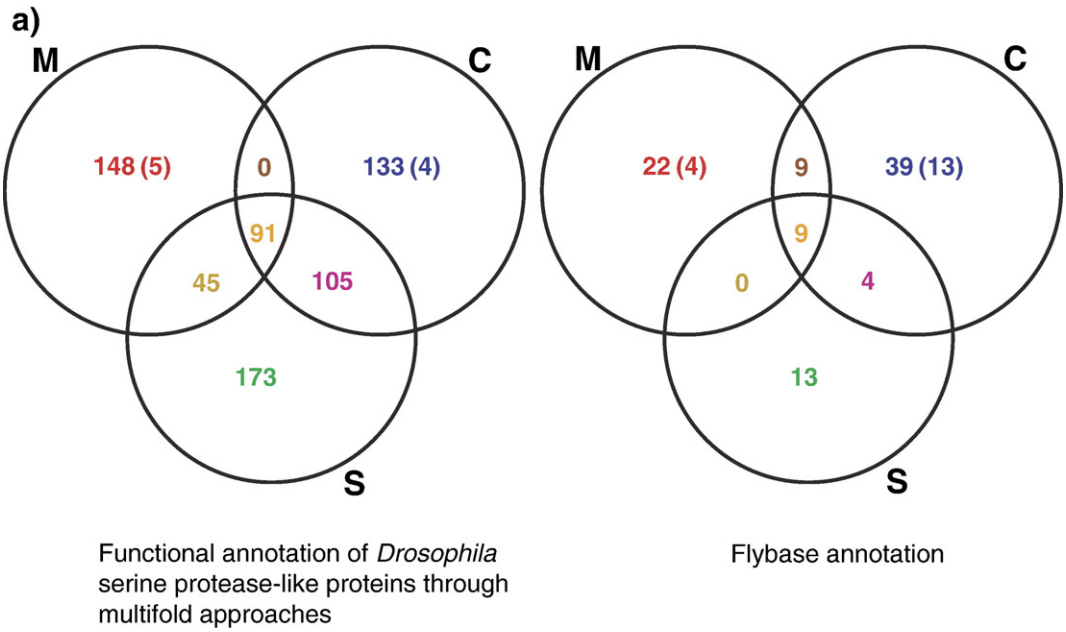
4. Discussion

Function annotation of proteins by computational approaches has remained difficult for a large set of proteins belonging to a family encoded in a single genome. There is no previous effort on systematic *in silico* annotation of all three GO terms for proteins belonging to a large multi-member family encoded in a single eukaryotic genome. Similarly, there are not many attempts to systematically identify function(s) for catalytically inactive enzyme sequences. Our analysis on *Drosophila* serine proteases describes such an approach with significant implications in genome annotation.

We establish FRC as a powerful tool for transferring *in vivo* substrate specificity (molecular function) in a quantitative manner and with high values of sensitivity and specificity. A number of computational methods have been developed to identify functionally important sites within protein families. These include approaches that employ parameters such as sequence and phylogenetic patterns (Lichtarge et al., 1996; Pei et al., 2006), 3D structures (Jones and Thornton, 2004; Jambon et al., 2005; Brylinski et al., 2007) and residue physical properties (Elcock, 2001). But none of these approaches explicitly utilize structural classification and function annotation present in sequence databases. Results from these methods could be used instead of the first step of FRC to provide a set of functionally important residues.

Another innovation of FRC lies in the clustering step, which provides a quantitative similarity measure between given proteins at their functional sites. In case of *Drosophila* serine proteases, FRC was found to be ~10-times more sensitive for transfer of annotation using similarities of functional residues than the whole domain similarities, from distantly related (and well-characterized) but divergent mammalian proteases to uncharacterized *Drosophila* serine proteases. It was also possible to postulate CIRs from the active-site similarity between pairs of active protease and SPHs using FRC. FRC results could be sensitive to availability of structural information and alignment errors.

FRC could potentially be employed to recognise CIRs with pathogenic proteases that are believed to have a bearing in host–pathogen interactions. To the best of our knowledge, this is the



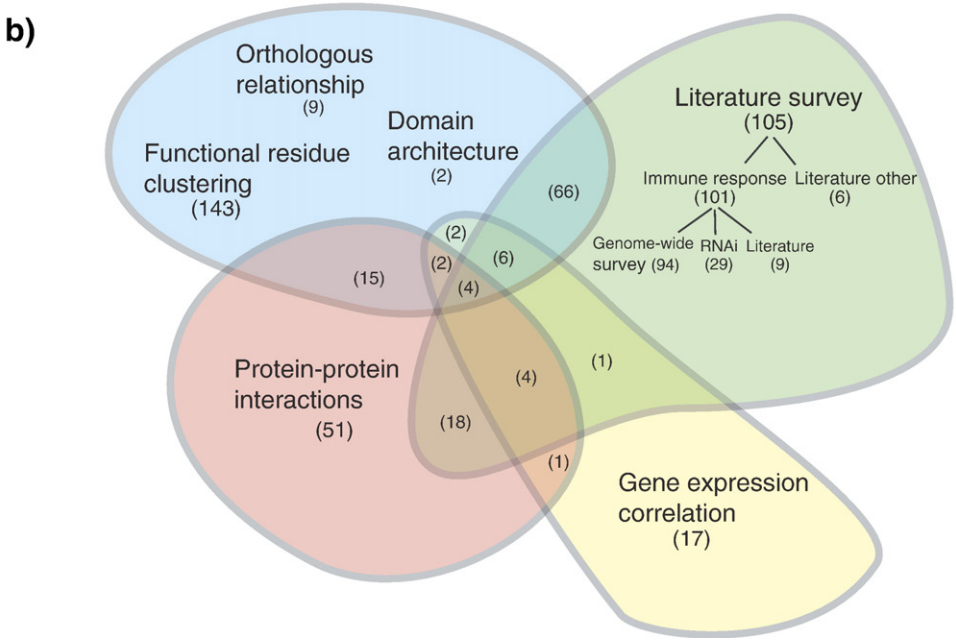


Fig. 6. a) A comparative representation of functional annotations for *Drosophila* serine-protease-like proteins corresponding to GO functional categories (Lewis, 2005) gathered through multi-fold approaches employed in our analysis and those provided for in Flybase (Grumbling and Strelets, 2006). Functional categories are represented in abbreviated form as follows: M — GO molecular function; C — GO cellular process (biological process); C — GO sub-cellular localization. The serine-protease-like gene products for which functional annotation was derived through either of the GO categories are indicated numerically in circles corresponding to each GO category. The figures in parentheses indicate the number of gene products for which function annotation was derived exclusively by the corresponding GO category. The numbers in intersects indicate gene products for which functional annotation was derived under more than one GO category (see Tables S8, S9). b) Multifold approaches employed for obtaining functional information resulted in function annotation for 190 serine-protease-like gene products in *Drosophila melanogaster*. The numbers in parentheses below the different approaches indicate the number of gene products for which possible functional information was obtained (see Table S8).

first computational method for identifying CIRs using structural data at a genome-wide level. The annotations are of high quality and are accompanied by evidences supporting our transfer of substrate specificities and CIRs from literature and other data types. However, there are no *in vivo* competitive inhibition data available for *Drosophila* proteases and the difficulties in identification of *in vivo* substrates for proteases at genomic scale (Hashimoto C, personal communication) makes it difficult to experimentally validate our results (on CIR) at present.

We observed large-scale and dual involvement of several active proteases and SPHs in *Drosophila* development and immune response. This large-scale involvement in immune response may explain rapid divergences of *Drosophila* proteases from their mammalian counterparts as compared to

other proteins in *Drosophila* genome. Moreover, analysis of gene-expression data provided evidences supporting evolution of new functions for SPHs in immune response. These results were further validated by results obtained from genome-wide *in vivo* RNAi screens of serine proteases for immune response (Kambris et al., 2006). To the best of our knowledge, this is the first report that identifies evolution of a new function for enzyme homologues (and SPHs) by utilizing high-throughput data. This finding is also significant as SPHs have wide-spread species distribution including in mosquito and humans (Zdobnov et al., 2002). Integration of gene-expression profiling data for multiple immune response experiments allowed identification of a shared core set as well as separate set of proteases with high confidence suggesting separate downstream cascades for specific immune response. However, unlike proteases involved in bacterial and fungal response, proteases participating only in parasite/parasitoid response are uncharacterized.

Different datasets may often provide overlapping or complementary information (Jensen and Steinmetz, 2005) due to hierarchy in definition of function of a gene. It is evident that structure–function analysis and gene-expression profiling data provide complementary information for function annotation. On the other hand, gene-expression profiling and RNAi screens may provide overlapping information, which could be used to increase confidence levels of function annotation. Our analysis also suggests that results from expression profiling can also be utilized for selecting targets for RNAi screening. Our efforts resulted in a ten-fold increase in function annotation for *Drosophila* proteases compared to Flybase at very fine-grained level (e.g. *Drosophila* Masquerade) by integration of multiple sources of data and manual intervention. We also provide annotations for more than half the SPHs identified in the *Drosophila* genome, which represents a significant progress in annotating function for enzyme homologues.

Finally, our results have established how knowledge-based computational tools can be reliably exploited in systematic annotation of function of gene products belonging to multi-member families in eukaryotic genomes. The efforts described here are limited only by the availability of suitable type of data and errors in annotation introduced by different databases and bioinformatics tools. Similar approaches may prove useful for attempting systematic function annotation of other large multi-member gene families such as protein kinases and G-Protein coupled receptors. Our results emphasize significance and usefulness of still evolving data-integration tools for large-scale function annotation, since enormous amounts of data from multiple sources are likely to become available in the near future.

Acknowledgements

R.S. was a Senior Research Fellow funded by the Wellcome Trust (U.K.). We would also like to acknowledge financial and infrastructural support of NCBS (TIFR). L.T. is a Senior Research Fellow of the Council of Scientific and Industrial Research (CSIR), India. This work was supported in part by the GeneFun project funded by the European Commission FP6 Programme, contract number LSHG-CT-2004-503567. Authors would also like to thank members of Bork group, and Dr. Toby Gibson for their constructive suggestions.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2007.10.012.

References

Apweiler, R., Bairoch, A., Wu, C.H., 2004. Protein sequence databases. Curr. Opin. Chem. Biol. 8, 76–80.

Aravind, L., Koonin, E.V., 2001. The DNA-repair protein AlkB, EGL-9, and leprecan define new families of 2-oxoglutarate- and iron-dependent dioxygenases. Genome Biol. 2 RESEARCH0007.

Arbeitman, M.N., et al., 2002. Gene expression during the life cycle of *Drosophila melanogaster*. Science 297, 2270–2275.

Bairoch, A., Boeckmann, B., Ferro, S., Gasteiger, E., 2004. Swiss–Prot: juggling between evolution and stability. Brief. Bioinform. 5, 39–55.

Bansal, M., Belcastro, V., Ambesi-Impiombato, A., di Bernardo, D., 2007. How to infer gene networks from expression profiles. Mol. Syst. Biol. 3, 78.

Barrett, A.J., 1994. Classification of peptidases. Methods Enzymol. 244, 1–15.

Bayer, C.A., Halsell, S.R., Fristrom, J.W., Kiehart, D.P., von Kalm, L., 2003. Genetic interactions between the RhoA and Stubble-stubbloid loci suggest a role for a type II transmembrane serine protease in intracellular signaling during *Drosophila* imaginal disc morphogenesis. Genetics 165, 1417–1432.

Beckstead, R.B., Lam, G., Thummel, C.S., 2005. The genomic response to 20-hydroxyecdysone at the onset of *Drosophila* metamorphosis. Genome Biol. 6, R99.

Berman, B.P., et al., 2004. Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*. Genome Biol. 5, R61.

Bork, P., Koonin, E.V., 1998. Predicting functions from protein sequences—where are the bottlenecks? Nat. Genet. 18, 313–318.

Brylinski, M., et al., 2007. Prediction of functional sites based on the fuzzy oil drop model. PLoS Comput. Biol. 3, e94.

Carro, A., et al., 2006. TreeDet: a web server to explore sequence space. Nucleic Acids Res. 34, W110–W115.

Cong, J., Geng, W., He, B., Liu, J., Charlton, J., Adler, P.N., 2001. The furry gene of *Drosophila* is important for maintaining the integrity of cellular extensions during morphogenesis. Development 128, 2793–2802.

Date, S.V., Marcotte, E.M., 2005. Protein function prediction using the Protein Link EXplorer (PLEX). Bioinformatics 21, 2558–2559.

De Gregorio, E., Spellman, P.T., Rubin, G.M., Lemaitre, B., 2001. Genome-wide analysis of the *Drosophila* immune response by using oligonucleotide microarrays. Proc. Natl. Acad. Sci. U. S. A. 98, 12590–12595.

De Gregorio, E., Spellman, P.T., Tzou, P., Rubin, G.M., Lemaitre, B., 2002. The Toll and Imd pathways are the major regulators of the immune response in *Drosophila*. EMBO J. 21, 2568–2579.

Deshpande, N., et al., 2005. The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. Nucleic Acids Res. 33, D233–D237.

Didelot, G., et al., 2006. Tequila, a neurotrypsin ortholog, regulates long-term memory formation in *Drosophila*. Science 313, 851–853.

Dissing, M., Giordano, H., DeLotto, R., 2001. Autoproteolysis and feedback in a protease cascade directing *Drosophila* dorsal-ventral cell fate. EMBO J. 20, 2387–2393.

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. U.S.A. 95, 14863–14868.

Elcock, A.H., 2001. Prediction of functionally important residues based solely on the computed energetics of protein structure. J. Mol. Biol. 312, 885–896.

Emanuelsson, O., Nielsen, H., Brunak, S., von Heijne, G., 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J. Mol. Biol. 300, 1005–1016.

Evans, S.V., 1993. SETOR: hardware-lighted three-dimensional solid model representations of macromolecules. J. Mol. Graph. 11, 134–138.
Finn, R.D., et al., 2006. Pfam: clans, web tools and services. Nucleic Acids Res. 34, D247–D251.
Gilbert, D., 2005. Biomolecular interaction network database. Brief. Bioinform. 6, 194–198.
Giot, L., et al., 2003. A protein interaction map of *Drosophila melanogaster*. Science 302, 1727–1736.
Gomis-Ruth, F.X., Gomez, M., Bode, W., Huber, R., Aviles, F.X., 1995. The three-dimensional structure of the native ternary complex of bovine pancreatic procarboxypeptidase A with proproteinase E and chymotrypsinogen C. EMBO J. 14, 4387–4394.
Grumbling, G., Strelets, V., 2006. FlyBase: anatomical data, images and queries. Nucleic Acids Res. 34, D484–D488.
Grundy, W.N., Bailey, T.L., Elkan, C.P., Baker, M.E., 1997. Meta-MEME: motif-based hidden Markov models of protein families. Comput. Appl. Biosci. 13, 397–406.
Guindon, S., Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 696–704.
Han, J.H., Lee, S.H., Tan, Y.Q., LeMosy, E.K., Hashimoto, C., 2000. Gastrulation defective is a serine protease involved in activating the receptor toll to polarize the *Drosophila* embryo. Proc. Natl. Acad. Sci. U. S. A. 97, 9093–9097.
Hannenhalli, S.S., Russell, R.B., 2000. Analysis and prediction of functional sub-types from protein sequence alignments. J. Mol. Biol. 303, 61–76.
Hegde, R., et al., 2002. Identification of Omi/HtrA2 as a mitochondrial apoptotic serine protease that disrupts inhibitor of apoptosis protein–caspase interaction. J. Biol. Chem. 277, 432–438.
Hoffmann, R., Valencia, A., 2004. A gene network for navigating the literature. Nat. Genet. 36, 664.
Hooper, S.D., et al., 2007. Identification of tightly regulated groups of genes during *Drosophila melanogaster* embryogenesis. Mol. Syst. Biol. 3, 72.
Irving, P., et al., 2001. A genome-wide analysis of immune responses in *Drosophila*. Proc. Natl. Acad. Sci. U. S. A. 98, 15119–15124.
Jackson, R.M., 1999. Comparison of protein–protein interactions in serine protease-inhibitor and antibody–antigen complexes: implications for the protein docking problem. Protein Sci. 8, 603–613.
Jambon, M., Andrieu, O., Combet, C., Deleage, G., Delfaud, F., Geourjon, C., 2005. The SuMo server: 3D search for protein functional sites. Bioinformatics 21, 3929–3930.
Jensen, L.J., Steinmetz, L.M., 2005. Re-analysis of data and its integration. FEBS Lett. 579, 1802–1807.
Johansson, K.C., Metzendorf, C., Soderhall, K., 2005. Microarray analysis of immune challenged *Drosophila* hemocytes. Exp. Cell Res. 305, 145–155.
Jones, S., Thornton, J.M., 2004. Searching for functional sites in protein structures. Curr. Opin. Chem. Biol. 8, 3–7.
Kam, C.M., et al., 1995. Mammalian tissue trypsin-like enzymes: substrate specificity and inhibitory potency of substituted isocoumarin mechanism-based inhibitors, benzamidine derivatives, and arginine fluoroalkyl ketone transition-state inhibitors. Arch. Biochem. Biophys. 316, 808–814.
Kambris, Z., et al., 2006. *Drosophila* immunity: a large-scale in vivo RNAi screen identifies five serine proteases required for Toll activation. Curr. Biol. 16, 808–813.
Kanehisa, M., Bork, P., 2003. Bioinformatics in the post-sequence era. Nat. Genet. 33, 305–310 Suppl.
Knappe, S., Wu, F., Masikat, M.R., Morser, J., Wu, Q., 2003. Functional analysis of the transmembrane domain and activation cleavage of human corin: design and characterization of a soluble corin. J. Biol. Chem. 278, 52363–52370.
Kyte, J., Doolittle, R.F., 1982. A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. 157, 105–132.
LeMosy, E.K., Hong, C.C., Hashimoto, C., 1999. Signal transduction by a protease cascade. Trends Cell Biol. 9, 102–107.
Letunic, I., Copley, R.R., Pils, B., Pinkert, S., Schultz, J., Bork, P., 2006. SMART 5: domains in the context of genomes and networks. Nucleic Acids Res. 34, D257–D260.
Lewis, S.E., 2005. Gene ontology: looking backwards and forwards. Genome Biol. 6, 103.
Lichtarge, O., Bourne, H.R., Cohen, F.E., 1996. An evolutionary trace method defines binding surfaces common to protein families. J. Mol. Biol. 257, 342–358.
Michels, B., Diegelmann, S., Tanimoto, H., Schwenkert, I., Buchner, E., Gerber, B., 2005. A role for Synapsin in associative learning: the *Drosophila* larva as a study case. Learn. Mem. 12, 224–231.
Moussian, B., Roth, S., 2005. Dorsoventral axis formation in the *Drosophila* embryo-shaping and transducing a morphogen gradient. Curr. Biol. 15, R887–R899.
Murugasu-Oei, B., Rodrigues, V., Yang, X., Chia, W., 1995. Masquerade: a novel secreted serine protease-like molecule is required for somatic muscle attachment in the *Drosophila* embryo. Genes Dev. 9, 139–154.
Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247, 536–540.
Pei, J., Cai, W., Kinch, L.N., Grishin, N.V., 2006. Prediction of functional specificity determinants from protein sequences using log-likelihood ratios. Bioinformatics 22, 164–171.
Peisach, E., Wang, J., de los Santos, T., Reich, E., Ringe, D., 1999. Crystal structure of the proenzyme domain of plasminogen. Biochemistry 38, 11180–11188.
Perona, J.J., Hedstrom, L., Rutter, W.J., Fletterick, R.J., 1995. Structural origins of substrate discrimination in trypsin and chymotrypsin. Biochemistry 34, 1489–1499.
Petryk, A., et al., 2003. Shade is the *Drosophila* P450 enzyme that mediates the hydroxylation of ecdysone to the steroid insect molting hormone 20-hydroxyecdysone. Proc. Natl. Acad. Sci. U. S. A. 100, 13773–13778.
Pils, B., Schultz, J., 2004. Inactive enzyme-homologues find new function in regulatory processes. J. Mol. Biol. 340, 399–404.
Pugalenthi, G., Bhaduri, A., Sowdhamini, R., 2005. GenDiS: genomic distribution of protein structural domain superfamilies. Nucleic Acids Res. 33, D252–D255.
Rawlings, N.D., Barrett, A.J., 1994. Families of serine peptidases. Methods Enzymol. 244, 19–61.
Retief, J.D., 2000. Phylogenetic analysis using PHYLIP. Methods Mol. Biol. 132, 243–258.
Rose, T., LeMosy, E.K., Cantwell, A.M., Banerjee-Roy, D., Skeath, J.B., Di Cera, E., 2003. Three-dimensional models of proteases involved in patterning of the *Drosophila* embryo. Crucial role of predicted cation binding sites. J. Biol. Chem. 278, 11320–11330.
Ross, J., Jiang, H., Kanost, M.R., Wang, Y., 2003. Serine proteases and their homologs in the *Drosophila melanogaster* genome: an initial analysis of sequence conservation and phylogenetic relationships. Gene 304, 117–131.
Rost, B., 2002. Enzyme function less conserved than anticipated. J. Mol. Biol. 318, 595–608.
Roxstrom-Lindquist, K., Terenius, O., Faye, I., 2004. Parasite-specific immune response in adult *Drosophila melanogaster*: a genomic study. EMBO Rep. 5, 207–212.
Suetake, T., et al., 2000. Chitin-binding proteins in invertebrates and plants comprise a common chitin-binding structural motif. J. Biol. Chem. 275, 17929–17932.
Szafron, D., et al., 2004. Proteome analyst: custom predictions with explanations in a web-based tool for high-throughput proteome annotations. Nucleic Acids Res. 32, W365–W371.
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.
Tesch, L.D., Raghavendra, M.P., Bedsted-Faarvang, T., Gettins, P.G., Olson, S.T., 2005. Specificity and reactive loop length requirements for crmA inhibition of serine proteases. Protein Sci. 14, 533–542.
von Mering, C., et al., 2005. STRING: known and predicted protein–protein associations, integrated and transferred across organisms. Nucleic Acids Res. 33, D433–D437.
Wertheim, B., et al., 2005. Genome-wide gene expression in response to parasitoid attack in *Drosophila*. Genome Biol. 6, R94.
Workman, C., et al., 2002. A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biol. 3 research0048.
Yoshino-Yasuda, I., Kobayashi, K., Akiyama, M., Itoh, H., Tomomura, A., Saheki, T., 1998. Caldecrin is a novel-type serine protease expressed in pancreas, but its homologue, elastase IV, is an artifact during cloning derived from caldecrin gene. J. Biochem. (Tokyo) 123, 546–554.
Zdobnov, E.M., et al., 2002. Comparative genome and proteome analysis of Anopheles gambiae and *Drosophila melanogaster*. Science 298, 149–159.