FlyXCDB—A Resource for *Drosophila* Cell Surface and Secreted Proteins and Their Extracellular Domains

Jimin Pei[1], **Lisa N. Kinch**[1] **and Nick V. Grishin**[1,2]

*1 - **Howard Hughes Medical Institute,** University of Texas Southwestern Medical Center, Dallas, TX 75390, USA*
*2 - **Department of Biophysics and Department of Biochemistry,** University of Texas Southwestern Medical Center, Dallas, TX 75390, USA*

***Correspondence to Jimin Pei:** Howard Hughes Medical Institute, 6001 Forest Park Dr., Dallas, TX 75390, USA.*
jpei@chop.swmed.edu
https://doi.org/10.1016/j.jmb.2018.06.002
Edited by Dan Tawfik

Abstract

Genomes of metazoan organisms possess a large number of genes encoding cell surface and secreted (CSS) proteins that carry out crucial functions in cell adhesion and communication, signal transduction, extracellular matrix establishment, nutrient digestion and uptake, immunity, and developmental processes. We developed the FlyXCDB database (http://prodata.swmed.edu/FlyXCDB) that provides a comprehensive resource to investigate extracellular (XC) domains in CSS proteins of *Drosophila melanogaster*, the most studied insect model organism in various aspects of animal biology. More than 300 *Drosophila* XC domains were discovered in *Drosophila* CSS proteins encoded by over 2500 genes through analyses of computational predictions of signal peptide, transmembrane (TM) segment, and GPI-anchor signal sequence, profile-based sequence similarity searches, gene ontology, and literature. These domains were classified into six classes mainly based on their molecular functions, including protein–protein interactions (class P), signaling molecules (class S), binding of non-protein molecules or groups (class B), enzyme homologs (class E), enzyme regulation and inhibition (class R), and unknown molecular function (class U). Main cellular functions such as cell adhesion, cell signaling, and extracellular matrix composition were described for the most abundant domains in each functional class. We assigned cell membrane topology categories (E, secreted; S, type I/III single-pass TM; T, type II single-pass TM; M, multi-pass TM; and G, GPI-anchored) to the products of genes with XC domains and investigated their regulation by mechanisms such as alternative splicing and stop codon readthrough.

Introduction

A significant fraction of genes in metazoan genomes encode proteins that are located on the cell surface or are secreted to the extracellular space [1–5]. These cell surface and secreted (CSS) proteins carry out a plethora of functions such as the establishment of extracellular matrix (ECM), cell adhesion and communication, signal transduction, nutrient digestion and transportation, immunity, and development [6–10].

The majority of eukaryotic CSS proteins go through a process of co-translational translocation involving the signal recognition particle (SRP) [11, 12]. SRP recognizes a short hydrophobic segment (signal sequence) of the nascent chain-ribosome complex and targets it to the endoplasmic reticulum (ER) membrane, where the signal sequence is recognized by a translocon that translocates the newly synthesized polypeptide [13]. The signal sequence could be part of an N-terminal signal peptide that is subsequently cleaved off by a signal peptidase or could eventually become part of a transmembrane (TM) segment that spans the lipid bilayer [14]. Predictions of N-terminal signal peptides and/or TMs are thus often suggestive of proteins going through the secretory pathway [2, 15]. While most of the proteins going through the secretory pathway are secreted or located on the cell surface, some are retained and carry out their functions within the endomembrane system including ER and the Golgi apparatus [16].

The topology of integral membrane proteins in the secretory pathway falls into several classes [17, 18]. Type I TM proteins possess a single TM segment and

have their C-termini located in the cytosol. Their primary sequences possess an N-terminal signal peptide that harbors the signal sequence for SRP recognition and eventually gets cleaved off by signal peptidase. Type II and type III TM proteins have a single TM segment that serves as the signal sequence recognized by SRP, but is not processed by signal peptidase. Type II and type III TM proteins have their N-termini and C-termini located in the cytosol, respectively. Multi-pass TM proteins (type IV) have multiple TM segments, and their topologies are often more difficult to resolve by computational predictions. The orientations of TM segments in single or multi-pass TM proteins are largely determined by positive charges near their ends, and the cytosolic sides are often enriched with positively charged residues [19].

Another class of proteins going through the secretory pathway is post-translationally modified and anchored to the cell surface via a C-terminally attached glycophosphatidylinositol (GPI) moiety [20, 21]. These proteins possess both an N-terminal signal peptide and a C-terminal hydrophobic segment that is cleaved off and replaced by the GPI. Other common post-translational modifications of CSS proteins include various forms of glycosylation and proteolytic cleavages by proteases other than signal peptidase [22, 23]. These modifications help their sorting to the cell surface/extracellular space as well as their cellular functions. CSS proteins can be internalized through endocytosis pathways and then either get recycled back to the cell surface/extracellular space or get degraded inside the lysosome or by the proteasome [24–26].

A number of large-scale bioinformatic or experimental studies have been carried out to study the repertoire of secreted proteins (secretome) or cell surface proteins (surfaceome) using various techniques [1, 5, 27, 28]. CSS proteins often have distinct combinations of functional domains located in the extracellular space [29]. Aside from experimental evidence, localization of CSS proteins can often be predicted from distinct signal sequences and domain compositions present in their primary sequences.

Drosophila melanogaster (*Dmel*) is the most studied insect model organism due to numerous advantages such as easy genetic manipulation [30, 31]. The whole genome of *Dmel* was sequenced in 2000 [32]. A later comparative genomics study also sequenced the genomes of 11 *Drosophila* species related to *Dmel* [33]. Their genome sequences, transcripts, and data of a wide range of experimental studies such as gene expression are maintained in FlyBase [34]. In this study, we combined automatic computational predictions and manual analysis to investigate the repertoire of extracellular (XC) domains in *Dmel* CSS proteins, with the aid of the proteomes of 11 related *Drosophila* species in the FlyBase database. We defined and classified XC domains presented in these *Drosophila* CSS proteins and assigned cell membrane topology (CMT) categories to non-redundant protein isoforms of *Dmel* genes in the FlyXCDB database (http://prodata.swmed.edu/FlyXCDB).

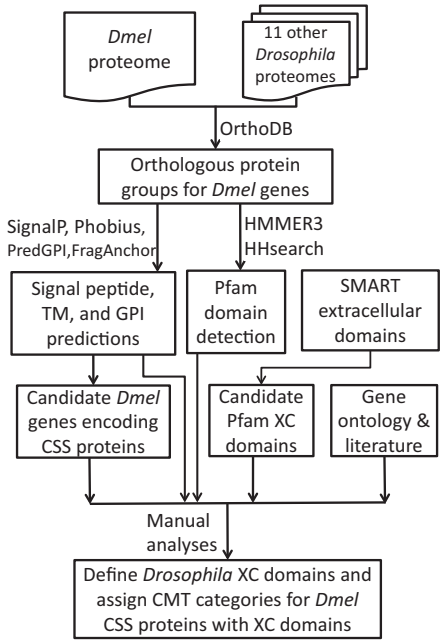
Results and Discussion

Detection and definition of XC domains in *Drosophila* proteins

We manually compiled a set of Pfam (version: 28) [35] domains and homologous domain groups that are present in *Drosophila* CSS proteins and could locate and function in the extracellular space. *Drosophila* XC domains were defined using previous knowledge of XC domains [36], predictions of signal peptide/TM segment/GPI signal sequence, profile-based domain detection, gene ontology, and literature (Fig. 1a, details described in Materials and Methods).

Pfam domains, corresponding to functional regions in proteins, are organized in two levels: family and clan [37]. A Pfam family is constructed by building a profile

(a)



(b) ***Drosophila* XC domain classes**

Class	#domains	Definition
P	118	XC domains likely involved in protein-protein interactions
S	27	XC domains mostly found in extracellular signaling molecules
B	51	XC domains that likely bind non-protein molecules and groups
E	66	XC domains that are enzyme homologs
R	43	XC enzyme regulatory and inhibitory domains
U	38	putative XC domains with unknown molecular function

Fig. 1. Defining XC domains in *Drosophila* CSS proteins. (a) A flowchart of the data sources and methods used to define *Drosophila* XC domains and assign CMT categories to *Dmel* CSS proteins with XC domains. (b) *Drosophila* XC domain classes, the numbers of domains, and their definitions.

hidden Markov model using a manually curated multiple alignment of representative sequences. A Pfam clan consists of a group of Pfam families that are evolutionarily related (homologous), and in many cases have similar molecular or cellular functions. Each *Drosophila* XC domain entry corresponds to either a single Pfam family or a group of Pfam families that often, but not always correspond to a collection of Pfam families in the same Pfam clan. For example, LRR(g) ((g) is applied to a *Drosophila* XC domain entry that has more than one Pfam families) is a single entry of *Drosophila* XC domain that corresponds to the Pfam LRR clan. Ig(g) is a single entry of *Drosophila* XC domain that includes a subset of Pfam families in the Pfam clan Ig that were detected in *Drosophila* proteins, such as I-set, Ig_2, ig, and V-set.

In some cases, domains in a Pfam clan are not considered collectively as a single entry of *Drosophila* XC domain, but are treated separately, if they have significant sequence dissimilarities (e.g., undetectable by HHsearch), different subcellular localizations, or are classified into different functional classes (described below). For example, a subset of the Pfam families in the diverse E-set clan with the immunoglobulin-like fold (IG-fold) were detected in the extracellular regions of *Drosophila* proteins and were classified into several *Drosophila* XC domain entries. Some of the Pfam families in the E-set clan, such as A2M_N, CBM39, FixG_C, HYR, Integrin_alpha2, PKD, REJ, and TIG, correspond to individual *Drosophila* XC domain entries. Pfam families Cadherin, Cadherin_2, Cadherin_3, and Cadherin_pro of the E-set clan are combined in a single entry of *Drosophila* XC domain called the Cadherin(g) as these Pfam domains are more closely related to each other than other E-set domains and are detected in the cadherin superfamily proteins [38]. *Drosophila* XC domain fn3(g) corresponds to Pfam families fn3, fn3_2, fn3_4, Pur_ac_phosph_N, Tissue_fac, Interfer-bind, EpoR_lig-bind, and IL6Ra-bind from the Pfam E-set clan as well as three Pfam domains DUF2369, DUF4959, and DUF4998 that have not been classified in a Pfam clan, but exhibit significant sequence similarity to other fn3-like domains.

In total, we identified and defined 342 *Drosophila* XC domains that could locate and function in the extracellular space of *Dmel*. These XC domains are detected in the products of 2,509 protein-coding genes out of the nearly 14,000 protein-coding genes of *Dmel* (FlyBase version FB2015_03). The number of XC domains and the genes containing them could change in future updates of FlyXCDB, given new knowledge about subcellular localizations of *Dmel* proteins and new definitions of Pfam domains.

A functional classification of *Drosophila* XC domains

Drosophila XC domains reside in CSS proteins with a plethora of cellular functions. A functional enrichment analysis of XC domain-containing genes by DAVID [39] revealed that the major categories of cellular functions (biological processes) of these genes include cell adhesion and recognition, immune and defense response, aminoglycan/polysaccharide/chitin metabolism, response to pheromone and chemical stimulus, gland development, and ECM/structure organization. Using a domain-and evolution-based strategy, these numerous biological processes can be described using a limited set of broad functional classes that were developed iteratively and concurrently with XC domain analysis and classification.

In FlyXCDB, *Drosophila* XC domains were manually classified into six broad functional classes (P, S, B, E, R, and U) mainly according to general molecular functions gathered from literature of collective homologous domains (Fig. 1b). About one third of the XC domains (118 out of 342) are in the class P, which includes domains likely involved in protein–protein interactions. Class S and class R domains are also involved in protein–protein interactions, but they are separated from class P domains considering their interacting targets. Class S includes 27 domains mostly occurring in XC signaling molecules, and they are involved in signal transduction by interacting with cell surface signaling receptors. Class R includes 42 enzyme regulatory and inhibitory domains that could interact with XC enzyme domains. Class E includes 66 XC domains of enzyme homologs in CSS proteins (including a few domains that have lost the catalytic activities). Class B includes 51 XC domains that likely interact with non-protein molecules and groups such as carbohydrates and lipids. Class U has 38 XC domains (including predicted ones) with unknown molecular function. It should be noted that some XC domain can possess multiple molecular functions, e.g., involved in both protein–protein interactions and binding non-protein molecules. We mainly relied on manual literature mining in assigning the most relevant functional class to any XC domain.

The number of genes associated with any XC domain type is not distributed evenly. Nearly a third of *Drosophila* XC domains (108) are present in only one gene, and more than half of the XC domains (182) are found in three or less genes. On the other hand, 10 XC domains have more than 50 *Dmel* genes associated with them: Trypsin(g) (259 genes), Ig(g) (132 genes), LIG(g) (119 genes), Chitin_bind_4 (118 genes), CBM_14_19(g) (110 genes), EGF(g) (88 genes), PBP(g) (77 genes), LRR(g) (72 genes), fn3(g) (58 genes), and PBP_GOBP (51 genes). Gene duplication and domain shuffling [40] could explain their high occurrence frequencies. Abundant XC domains in each functional class are shown in Fig. 2.

Some *Drosophila* XC domains tend to co-occur with other XC domains in the same gene products. We constructed a domain co-occurrence network [41] where vertices exist between any two XC domains that can co-occur in the same gene product (Fig. 3a). This

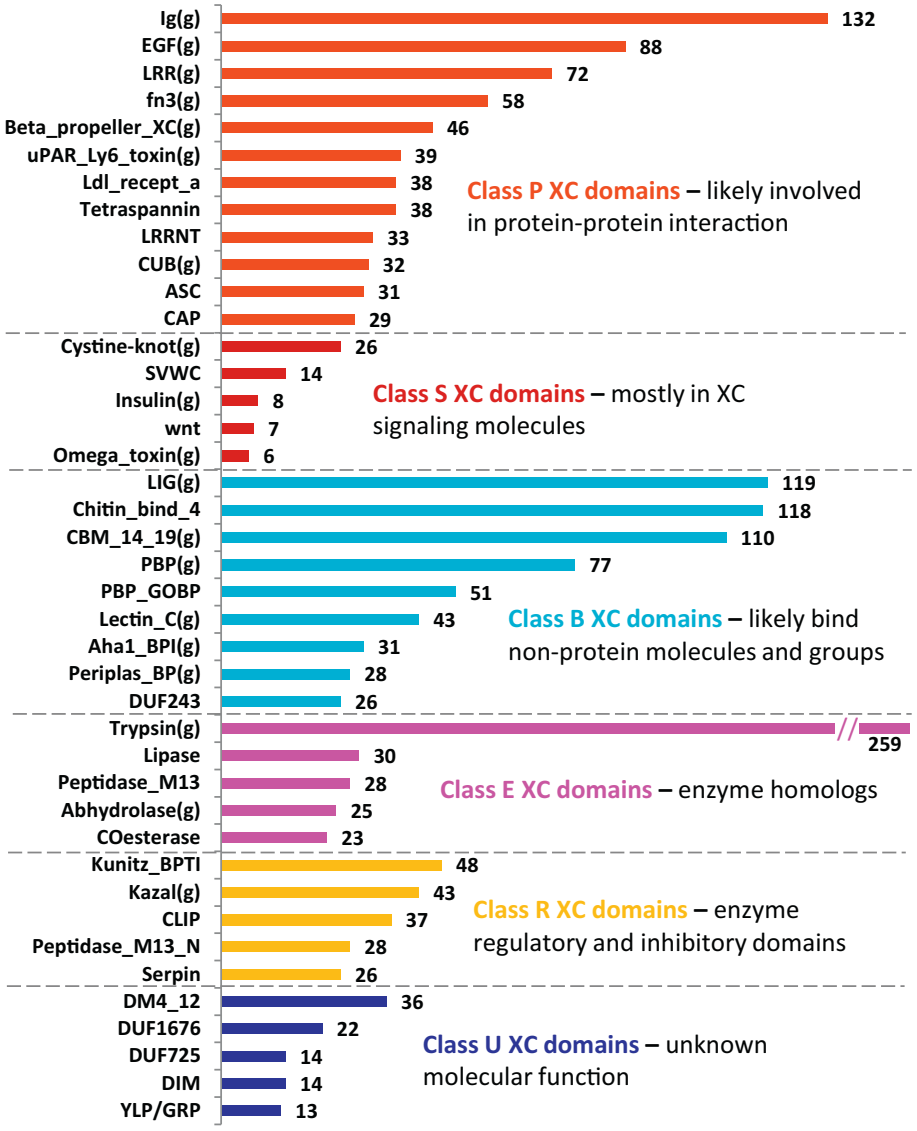


Fig. 2. Most abundant *Drosophila* XC domains. XC domains with the most number of genes are shown for each functional class. They include top five most abundant domains in each functional class as well as domains associated with more than 25 genes. Domains are colored as follows: class P, red; class S, dark red; class B, cyan; class E, magenta; class R, orange; class U, blue.

network consists of a major connected component with 137 XC domains, 8 connected components with three XC domains, 20 connected components with two XC domains, and 141 singletons corresponding to XC domains that do not co-occur with other XC domains. The majority of class S domains (mainly occurring in signaling molecules, 26 out of 27) and class U domains (unknown molecular function, 29 out of 38) do not co-occur with other XC domain types. On the other hand, 26 XC domains co-occur with 10 or more other XC domain types. More than half of them (15 out of 26) are class P domains that are likely involved in protein–protein interactions. Among them, EGF(g) co-occurs with the most number (61) of XC domain types. The co-occurrence network for genes containing EGF(g) is shown in Fig. 3b, which also contains those for five other XC domains that frequently co-occur with other XC domains types: Ldl_recept_a, Laminin_G(g), Ig(g), CUB(g), and TSP_1 (they co-occur with 36, 30, 27, 20, and 20 XC domain types, respectively).

Class P XC domains likely involved in protein–protein interactions

The designation of "P" for a protein-interacting XC domain is mostly based on literature finding of the domain's involvement in protein interactions or its

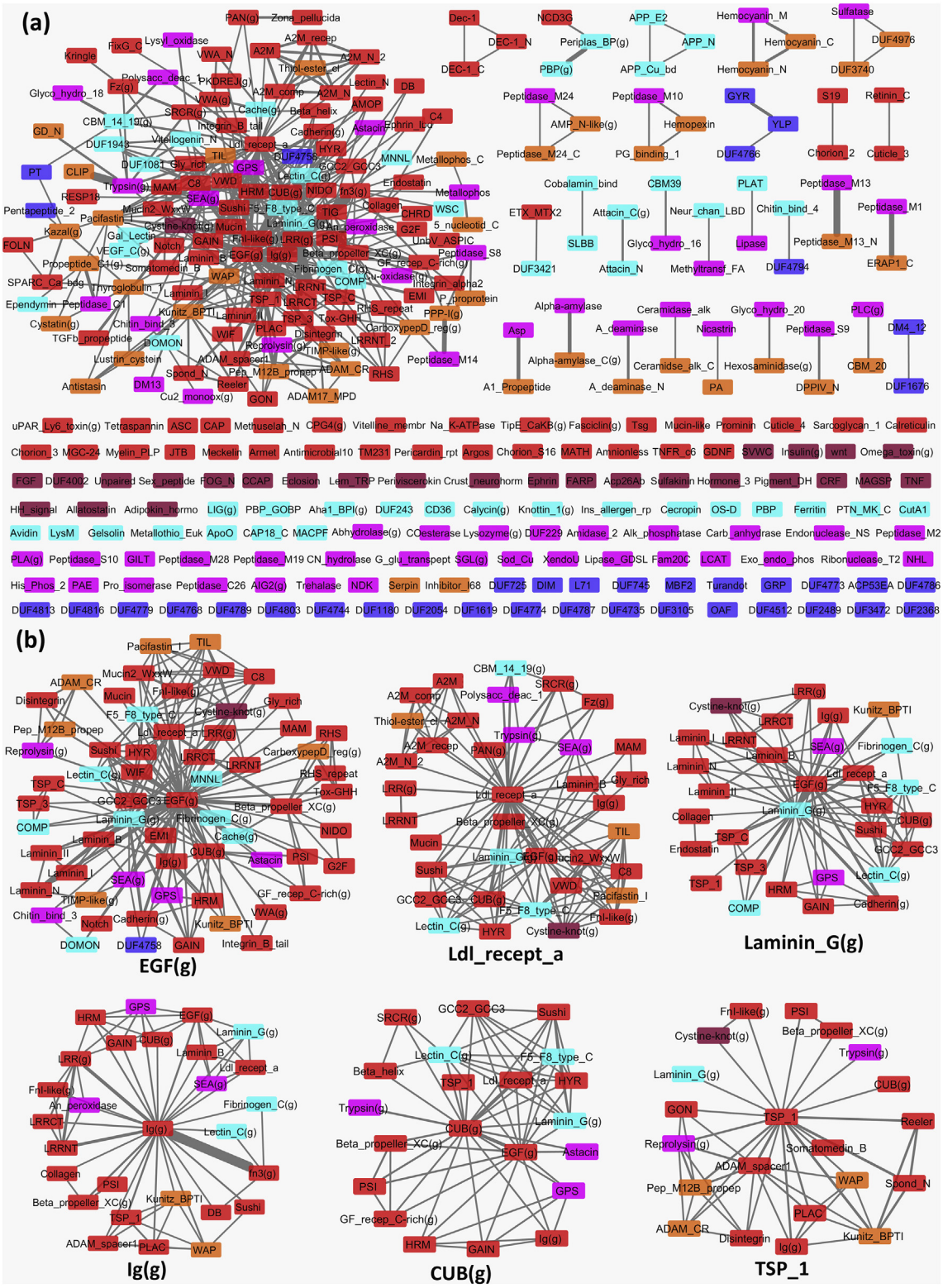


Fig. 3. Co-occurrence network of *Drosophila* XC domains. (a) The network for all genes containing XC domains. Domains are colored as follows: class P, red; class S, dark red; class B, cyan; class E, magenta; class R, orange; class U, blue. Two domains are connected if they co-occur in at least one gene. The thickness of the line is proportional to the number of co-occurring genes. This network consists of a large connected component shown in the upper left region, 28 connected components with two or three XC domains shown in the upper right region, and 137 singleton XC domains shown at the bottom. (b). From top left to bottom right, co-occurrence networks for genes with EGF(g), genes with Ldl_recept_a, genes with Laminin_G(g), genes with Ig(g), genes with CUB(g), and genes with TSP_1.

occurrence in major ECM proteins. The most abundant domains in this class (Table 1), such as Ig(g), fn3(g), LRR(g), and EGF(g), are frequently found in cell surface proteins and ECM components that are involved in cell adhesion, signaling, and cell–cell communications [42,43]. The 12 most abundant XC domains in this class, each of which is present in more than 25 *Dmel* genes, include Ig(g), EGF(g), LRR(g), fn3(g), Beta_propeller_XC(g), uPAR_Ly6_toxin(g), Ldl_recept_a, Tetraspanin, LRRNT, CUB(g), ASC, and CAP (Table 1).

Class P XC domains with an IG-fold. The general IG-fold is made up of seven core antiparallel strand elements forming two β-sheets and is one of the most populated folds in the structure databases (Fig. 4) [44, 45]. The classical immunoglobulin (Ig) domains, originally found in vertebrate immunoglobulin proteins, adopt such a fold and are included in the Pfam Ig clan. This clan contains a number of Pfam families such as I-set, C-set, ig, and Ig_2. In FlyXCDB, the classical Ig domain [Ig(g)] represents the largest class P XC domain in terms of the number of *Dmel* genes containing them (132 genes). Most members of Ig(g) have a disulfide bond between the second and sixth β-strands that bridges the two β-sheets. A conserved large hydrophobic residue from the third β-strand, usually a tryptophan, stacks against the disulfide bond in the core of the structure (Fig. 4A). Compared to other IG-fold domains such as fibronectin-type III domain [fn3(g)] (Fig. 4B) and Cadherin(g) (Fig. 4C), Ig(g) has two β-strands connecting the third and fifth core strand elements and pairing with them respectively (Fig. 4A). The main function of Ig(g) is cell adhesion, which is crucial in development and nervous system dynamics such as axon guidance, synaptogenesis, and synaptic plasticity. In addition, Ig(g) is frequently found in various cell surface receptors that interact with extracellular signaling molecules [46]. As a protein–protein interaction module, Ig(g) is also present in some ECM proteins. It is found in intracellular proteins as well, including several large muscle proteins [47].

A number of other *Drosophila* XC domains also possess the IG-fold. They generally lack the Ig(g) sequence signatures (the disulfide bond linking the second and sixth β-strands and the conserved aromatic residue in the third β-strand). A collection of IG-fold Pfam domains are combined in the Pfam E-set clan, such as fn3, Cadherin, and A2M. As described above, we split the E-set clan into different groups based on sequence similarities and their functions. The *Drosophila* XC domain fn3(g) consists of Pfam domain fn3 and a few other domains that are homologous based on HMMER or HHsearch searches. The common sequence features of fn3(g) are three conserved large hydrophobic (mostly aromatic) residues in the second β-strand (Trp), the third β-strand (Tyr), and the last β-strand (mostly Tyr). They lie in the core of the structure, with the first two aromatic residues forming a stacking interaction (Fig. 4b). *Dmel* has 58 genes containing extracellular fn3(g) domains. Most of them (44) also possess the Ig(g) domain. Both Ig(g) and fn3(g) often occur in tandem repeats in *Dmel* CSS proteins [47]. IG-fold XC domains in *Dmel* CSS proteins and their main cellular functions are described below.

Ig(g) and fn3(g) domains in cell adhesion molecules. The Dpr (defective proboscis extension response) gene family [48] is the largest Ig(g) domain-containing group in *Dmel* with 21 members (*dpr1–21*). They encode cell adhesion proteins with two Ig(g) domains. They are predicted to be type I TM (e.g., Dpr5), type II TM (e.g., Dpr9), GPI-anchored (e.g., Dpr12), or secreted (e.g., Dpr10) proteins. A recent large-scale study of interactions of *Dmel* extracellular proteins identified a group of Dpr-interacting proteins (DIPs) [49] with three Ig(g) domains. The DIP gene family includes 11 members: *DIP-α–ζ* and two closely related genes *CG31814* and *CG45781* [49]. Dpr/DIP pairings appear to be important for the patterns of synaptic connectivity [50]. The Dpr–DIP interaction network plays crucial roles in presynaptic terminal development, trophic factor responses, and neurotransmission [42].

Another gene family encoding Ig(g)-containing proteins is the Beat family with 14 members [51]. The founding member, Beaten path Ia (encoded by *beat-Ia*), is an anti-adhesion protein that regulates defasciculation at motor axon choice points [52]. Beat family proteins all possess two Ig(g) domains. Beat-Ia interacts with Side (Sidestep), which is important for target recognition in axon guidance [53]. Side belongs to a family of type I TM proteins with five Ig(g) domains and one fn3(g) domain (the other members being CG14372, CG42313, CG34113, CG34114, CG12950, CG12484, and CG34371). The interactions between other Beat family members and the Side family proteins were identified in a large-scale study of extracellular protein interactions [49].

The Dscam (Down syndrome cell-adhesion molecules) gene family has four members (*Dscam1–4*) in *Dmel* with multiple Ig(g) domains and multiple fn3(g) domains. *Dscam1* encodes axon guidance receptors that bind the adaptor protein Dock and activate the protein kinase Pak. *Dscam1* exhibits high molecular diversity via alternative splicing of four exon clusters that results in tens of thousands of isoforms [54, 55]. These isoforms help recognition between neurons by homophilic repulsion and provide the basis for self-avoidance in the development of nervous system [56]. *Dscam2–4* do not have the complex repertoire of isoforms as *Dscam1*, yet their products also play important roles as cell recognition molecules that regulate neural circuit assembly [57].

Other examples of Ig(g)-containing cell adhesion molecules in axon guidance include Ama (Amalgam), which is a secreted protein with three Ig(g) domains and

Table 1. Class P XC domains likely involved in protein–protein interactions

	XC domain	No. genes
1	Ig(g)	132
2	EGF(g)	88
3	LRR(g)	72
4	fn3(g)	58
5	Beta_propeller_XC(g)	46
6	uPAR_Ly6_toxin(g)	39
7	Ldl_recept_a	38
8	Tetraspannin	38
9	LRRNT	33
10	CUB(g)	32
11	ASC	31
12	CAP	29
13	TSP_1	18
14	Cadherin(g)	18
15	Zona_pellucida	18
16	Methuselah_N	16
17	FnI-like(g)	15
18	Sushi	15
19	Retinin_C	14
20	Fz(g)	13
21	HRM	12
22	LRRCT	10
23	PSI	10
24	Collagen	10
25	PAN(g)	9
26	GF_recep_C-rich(g)	9
27	TGFb_propeptide	8
28	VWA(g)	8
29	GAIN	7
30	Disintegrin	7
31	Cuticle_3	7
32	Reeler	7
33	SRCR(g)	7
34	Laminin_N	6
35	VWD	6
36	CPG4(g)	6
37	A2M_N	6
38	A2M_recep	6
39	A2M_N_2	6
40	EMI	6
41	A2M	6
42	A2M_comp	6
43	Vitelline_membr	6
44	Na_K-ATPase	6
45	Somatomedin_B	5
46	MAM	5
47	Integrin_alpha2	5
48	TipE_CaKB(g)	5
49	C8	5
50	Spond_N	5
51	ADAM_spacer1	5
52	Laminin_B	4
53	PKDREJ(g)	4
54	Mucin	4
55	Fasciclin(g)	4
56	WIF	4
57	Tsg	3
58	Mucin-like	3
59	NCD3G	3
60	NIDO	3
61	Prominin	3
62	Cuticle_4	3
63	SPARC_Ca_bdg	3
64	VWA_N	3
65	TIG	3
66	Chorion_2	3
67	PLAC	3
68	Mucin2_WxxW	2
69	Sarcoglycan_1	2
70	Calreticulin	2
71	RHS	2
72	FOLN	2
73	Integrin_B_tail	2
74	Antimicrobial10	2
75	Tox-GHH	2
76	C4	2
77	HYR	2
78	Laminin_II	2
79	Chorion_3	2
80	RHS_repeat	2
81	Kringle	2
82	GCC2_GCC3	2
83	ETX_MTX2	1
84	GON	1
85	Gly_rich	1
86	MGC-24	1
87	G2F	1
88	UnbV_ASPIC	1
89	Myelin_PLP	1
90	JTB	1
91	Laminin_I	1
92	Meckelin	1
93	FixG_C	1
94	Armet	1
95	TM231	1
96	Pericardin_rpt	1
97	Argos	1
98	RESP18	1
99	Lectin_N	1
100	DB	1
101	TSP_C	1
102	TSP_3	1
103	Chorion_S16	1
104	MATH	1
105	LRRNT_2	1
106	S19	1
107	Amnionless	1
108	CHRD	1
109	Ephrin_lbd	1
110	DEC-1_N	1
111	DEC-1_C	1
112	Endostatin	1
113	TNFR_c6	1
114	AMOP	1
115	Dec-1	1
116	Notch	1
117	Beta_helix	1
118	GDNF	1

interacts with Nrt (Neurotactin) [58]. Fas3 (Fasciclin 3), a type I TM protein with three Ig(g) domains, mediates synaptic target recognition through homophilic interaction [59]. Fas2 (Fasciclin 2), a membrane-bound protein with Ig(g) and fn3(g) domains, contributes to neuronal recognition [60]. Two genes *rst* and *kirre* encode highly similar type I TM proteins (60% sequence identity) with five Ig(g) domains. They mediate heterophilic cell adhesion with another paralogous pair of cell surface

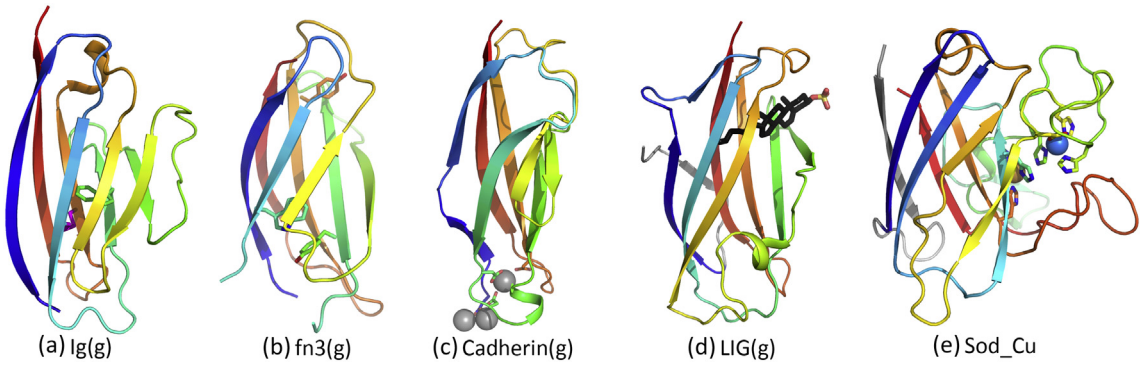


Fig. 4. IG-fold XC domains of different classes. Structures of IG-fold domains are shown for class P domains, Ig(g) (5eo9, B42–B143), fn3(g) (2ibh, A579–A672) and Cadherin(g) (3ubh, A751–A850); class B domain, LIG(g) (5kwy, C20–C149); and class E domain, Sod_Cu (1oal, A1–A151). These structures are rainbow-colored from the first core β-strand to the last core β-strand. Sidechains of conserved residues in Ig(g) and fn3(g) in the structure core are shown. The disulfide bond in Ig(g) is colored magenta. Ligand in LIG(g) is shown in sticks. Calcium (gray balls) and calcium-binding residues are shown for Cadherin(g). Zinc (blue ball) and copper (orange ball) and their binding residues are shown for Sod_Cu.

proteins Sns and Hbs in regulation of myoblast fusion and nephrocyte diaphragm assembly [61, 62].

Ig(g) and fn3(g) domains in cell surface receptors. Ig (g) and fn3(g) domains are found in a variety of cell surface receptors in signal transduction pathways. These receptors can be broadly catalogued into three types: receptor tyrosine kinases (RTKs), receptor tyrosine phosphatases (RTPs), and receptors not linked to cytoplasmic enzyme domains.

Four RTKs—Otk, Btl, Htl, and Pvr—possess multiple Ig(g) domains, but no fn3(g) domains in their extracellular regions. Otk interacts with Wnt4 and inhibits the canonical Wnt signaling in embryonic patterning [63]. Otk also plays important roles in establishing layer-specific neuronal connectivity [64]. Otk has a paralog, Otk2, that is also a type I TM protein with Ig(g) domains, but lacks the tyrosine kinase domain. Otk and Otk2 are co-receptors of Wnt2, and they interact with Frizzled receptors and function in development of the genital tract [65]. Btl (Breathless) and Htl (Heartless) are two RTKs for FGF ligands, while Pvr is the RTK for PDGF/VEGF ligands.

Another five RTK homologs—Sev, Eph, Tor, InR, and Wsck—have XC fn3(g) domains, but no Ig(g) domains. Sev (Sevenless) and Eph are involved in juxtacrine signaling, as their ligands are membrane-bound proteins (Boss and Ephrin, respectively) from neighboring cells. Sev has a large extracellular region (over 2000-amino-acid residues) including several fn3 (g) domains as well as regions of beta-propeller domains of the YWTD type [66]. Eph has two fn3(g) domains in the juxtamembrane region, an N-terminal Ephrin_bind domain responsible for ligand binding, and two EGF-like domains (Ephrin_rec_like) [67]. Tor (Torso) possesses only fn3(g) domains in its extracellular region. The insulin-like peptide receptor InR (Insulin-like receptor) has a Furin-like cysteine domain, fn3(g) domains, and two regions of LRR repeats of the Recept_L_domain type. Wsck is an RTK homolog with a catalytically inactive kinase domain [68], as well as two fn3(g) domains and a WSC domain in the extracellular region. The function of Wsck is yet to be determined.

Eight homologs of membrane-bound RTPs are found in the *Dmel* genome [69]. Three of them—Lar, Ptp69D, and Ptp99A—possess both Ig(g) and fn3(g) domains in their extracellular regions. They have two phosphatase domains in their cytosolic regions, only one of which is catalytically active. On the other hand, Ptp10D, Ptp4E, and Ptp52F possess fn3(g) domains and a single intracellular phosphatase domain that is catalytically active. These six fn3(g)-containing RTPs (Lar, Ptp69D, Ptp99A, Ptp10D, Ptp4E, and Ptp52F) play critical roles in the development of *Dmel* nervous system, motor neuron axon guidance, and axon target recognition [70,71]. Each of the two closely related RTP paralogous pairs Ptp69D/Ptp99A and Ptp10D/Ptp4E could have redundant roles [72]. The ligands of RTPs from vertebrates and invertebrates are largely unknown [73]. Ligand candidates of some RTPs have been proposed to be their interaction partners such as the cell adhesion protein Contactin and heparan sulfate/chondroitin sulfate in ECM [73]. Syndecan, a heparan sulfate proteoglycan (HSPG), is a ligand of Lar [74]. The type I TM protein Sas (Stranded at second), another fn3(g)-containing protein, is a ligand of Ptp10D [75, 76].

Several cell surface signaling receptors with Ig(g) and/or fn3(g) domains do not have protein kinase or phosphatase domains in their cytoplasmic regions. Two major pathways regulating *Dmel* axon guidance (Slit-Robo and Netrin-Frazzled) [77] involve cell surface receptors with Ig(g) and fn3(g) domains. Three Robo receptors (Robo1–3) in *Dmel* regulate axon guidance as well as organ development [78]. They all have five extracellular Ig(g) domains and three fn3(g) domains. The gene *fra* (*frazzled*) encodes a cell surface receptor for the Netrin ligands (NetA and NetB) [79]. It has

multiple roles in axon and dendrite guidance, development, and morphogenesis. Fra has four Ig(g) domains and six fn3(g) domains. Another Netrin receptor is Unc-5, which elicits repulsive signals in motor axon guidance [80]. Unc-5 possesses two Ig(g) domains and two TSP_1 domains in the extracellular region, and ZU5 and Death domains in the cytoplasmic region.

Dome (Domeless), with one Ig(g) domain and seven fn3(g) domains, is the cell surface receptor for three cytokine ligands Upd1–3. They signal through the JAK–STAT pathway and play important roles in immunity and development [81, 82]. While Dome does not have a cytoplasmic kinase domain, it induces activation of the intracellular JAK kinase Hopscotch upon ligand binding. Et (Eye transformer) is a protein remotely related to Dome (25% identity of BLAST alignment) and serves as a negative regulator of the JAK–STAT pathway by interacting with Dome in a dominant negative fashion [83]. Dome and Et are possibly paralogs generated by a gene duplication event, as they are reciprocal best hits of BLAST searches in the *Dmel* proteome. Ig(g) domains are also found in two adhesion G protein-coupled receptors (GPCRs) (CG15744 and CG44153) that are orphan receptors with unknown functions and ligands.

Ig(g) and fn3(g) in other CSS proteins. Besides their abundance in cell adhesion molecules and signaling receptors, Ig(g) and fn3(g) domains are found in other CSS proteins in signaling pathways such as signaling ligands, co-receptors, and antagonists. For example, Ig(g) is present in the EGF(g)-containing signaling protein Vn (Vein) and two semaphorin proteins (Sema2a and Sema2b) that act as signaling ligands for plexins. The fn3(g) domain is found in Sas mentioned above [75, 76]. Two Ig(g) and fn3(g)-containing proteins, Ihog (Interference hedgehog) and Boi (Brother of ihog), are paralogs and serve as co-receptors of the Hedgehog receptor Patched. Interestingly, they also interact with the EGF ligand Vn, providing a possible link between the EGF and Hedgehog signaling pathways [49]. Ig(g) domains in Ihog, Boi, and Vn are responsible for their interactions. ImpL2 (Imaginal morphogenesis protein-Late 2), a protein with two Ig(g) domains, binds insulin-like peptides and is an antagonist of the insulin signaling pathway [84, 85].

Ig(g) are also present in some ECM proteins. It is found in the proteoglycan Trol (the *Dmel* ortholog of vertebrate Perlacan) with a modular domain content that also includes EGF(g), SEA(g), Ldl_recept_a, Laminin_G(g), and Laminin_B domains [86]. Two related ECM proteins Hig (Hikaru genki) and Hasp (Hig-anchoring scaffold protein) with Ig(g) and Sushi domains interact with each other and play distinct roles in synaptogenesis [87]. Hig and Hasp are likely paralogs arisen from gene duplication, as they are reciprocal best BLAST hits against the *Dmel* proteome. Compared to Hig, Hasp is longer with more Sushi repeats and also has an N-terminal WAP domain. Ppn (Papilin) is an ECM protein with an Ig(g) domain and several domains of ECM characteristics such as TSP_1, ADAM_spacer1, and PLAC [88]. Nolo is another possible ECM proteins with one Ig(g) domain, multiple TSP_1 domains, and two PLAC domains. The multi-domain protein Pxn (Peroxidasin) has Ig(g), LRR(g), VWC, and an enzyme domain An_peroxidase. It plays a role in ECM organization [89]. In addition, two Ig(g) domains are found in CG6867 together with the Collagen domain and the OLF domain that are often present in ECM components [90, 91].

Class P IG-fold domains in cadherins. Proteins of the cadherin superfamily [38,92] (called cadherins below) possess the Cadherin(g) domain that has an IG-fold and can bind calcium (Fig. 4C). Cadherin(g) domains are involved in homophilic or heterophilic interactions among cadherins [93]. *Dmel* has 17 cadherins that are membrane-bound cell adhesion and signaling molecules with diverse domain contents and functions. Three classical cadherins (CadN, CadN2, and Shg) have the Cadherin_C domain in the cytoplasmic region to mediate interactions with cytoskeleton. The juxtacrine signaling between two cadherins—the receptor Ft (Fat) and its ligand Ds (Dachsous)—plays crucial roles in the establishment of planar cell polarity [94] as well as in the hippo pathway that regulates cell differentiation [95]. Stan (Starry night, also called Flamingo) is another cadherin that functions in planar cell polarity through its interaction with the Frizzled receptors [96]. The cell adhesion functions of other cadherins, such as CadN and Shg, are important for axonogenesis and axon guidance [97, 98].

Sixteen of the 17 *Dmel* cadherins are type I single-pass TM proteins, while Stan is a GPCR with seven TM segments. Most of *Dmel* cadherins are large proteins with tandem repeats of cadherin domains. For example, Ds possesses more than 30 repeats of Cadherin(g) domains in its extracellular region. Eight *Dmel* cadherins (Ft, Shg, CadN, CadN2, Stan, Kon, Cals, and Kug) contain the Laminin_G(g) domain, and seven of them (except Cals) also have EGF(g) domains. A divergent SEA domain is present in between the cadherin repeats and the EGF repeats in six such cadherins (Ft, Shg, Cadn, Cadn2, Kug, and Stan) [99]. Another divergent SEA domain, found in the juxtamembrane regions of human cadherins CDH23, PCDH15, and CDHR2 [99], is also present the juxtamembrane regions of six *Dmel* cadherins (Cad74A, Cad86C, Cad87A, Cad88C, Cad89D, and Cad99D). Cad96Ca is a RET-like RTK [100] with a single Cadherin(g) domain in its extracellular region.

Class P IG-fold domains in thioester-containing proteins. Six genes encoding thioester-containing proteins (TEPs) are present in *Dmel* (*Tep1–5* and *Mcr*). They are homologs of vertebrate alpha 2-macroglobin and complement components C3 and C4 [101]. Alpha 2-macroglobin acts as a broad range protease inhibitor. Like C3 and C4, arthropod TEPs are involved in immunity response as they can bind

pathogens and target them to phagocytosis. *Dmel* Tep1, Tep2, and Tep4 are predicted to be secreted proteins, while Tep3 and Mcr are predicted to be GPI-anchored and type I TM proteins, respectively. Mcr and Tep5 have substitutions at the cysteine position in the thioester motif, which is intact in Tep1–4. Without signal peptide predictions and with undetectable expression levels [102], *Dmel Tep5* could be a pseudogene [103]. Mcr was shown recently to be a key component of septate junction [104].

TEP homologs in vertebrates and invertebrates have a similar modular domain structure in their extracellular regions, including eight copies of IG-fold domains [105]. Regions of the eight IG-fold domains in TEP homologs are covered partially by the Pfam domains A2M_N, A2M_N_2, A2M, and A2M_recept. The sixth IG-fold domain harbors a bait region that is the target of proteolytic cleavage. Conformational changes after the cleavage collapse the alpha 2-macroglobin around the protease active site and block the protease's access to other targets. A divergent CUB domain is located in between the seventh and eighth IG-fold domains, and an alpha–alpha helical barrel domain of the 6_Hairpin fold is inserted in the CUB domain [105]. This helical domain contains the intrachain beta-cysteinyl-gamma-glutamyl thioester (in the CGEQ consensus motif) that becomes active upon conformational changes. The exposed thioester bond is liable to nucleophilic attack by amines. The alpha–alpha helical barrel domain corresponds to the Pfam domains Thio-ester_cl (containing the thioester motif) and A2M_comp.The C-terminal IG-fold domain (A2M_recept) interacts with cell surface receptors that can target TEP proteins for endocytosis.

Class P IG-fold domains in integrins and dystrophin-associated complex. Integrin complex and dystrophin-associated glycoprotein complex are cell surface receptors for ECM. They provide a signaling linkage between ECM and cytoskeleton. IG-fold domains have been identified in proteins in these complexes.

An integrin complex is made up of an α subunit and a β subunit, both of which are type I TM proteins [106]. In *Dmel*, five genes (*if*, *ItgaS4*, *ItgaPS5*, *mew*, and *scb*) and two genes (*Itgbn* and *mys*) encode integrin α subunits and β subunits, respectively [107]. The extracellular region of a *Dmel* integrin α subunit consists of an N-terminal 7-bladed beta-propeller domain of the FG-GAP type and three C-terminal IG-fold domains. These domains are involved in the interaction with the integrin β subunit [108]. None of *Dmel* integrin α subunits possess the VWA-like domain inserted in the beta-propeller, as observed in some vertebrate integrin α subunits. The three IG-fold domains (called thigh, calf1, and calf2) are represented as a single Pfam family (Integrin_alpha2). The integrin β subunit has a different modular domain structure compared to the α subunit. Its extracellular region has a PSI domain, an IG-fold domain, a VWA domain inserted inside the IG-fold domain, four EGF repeats, and a juxtamembrane tail domain (Pfam: Integrin_B_tail). The PSI domain, IG-fold domain, and the VWA domain are integrated into one Pfam domain named Integrin_beta.

Dystroglycan and sarcoglycans are cell surface proteins in the dystrophin glycoprotein complex [109]. *Dmel* Dystroglycan (encoded by *Dg*) has two cadherin-like IG-fold domains and two SEA domains [99]. The Pfam family DAG1 includes the juxtamembrane SEA domain, the TM, and cytoplasmic region, while the IG-fold cadherin-like domains in Dg (CADG) have not been incorporated in the Pfam database. *Dmel* sarcoglycan α-subunit (Scgα) similarly has the CADG domain and the SEA domain [99], and they are together included in the Pfam Sarcoglycan_2 entry. Dg and Scgα are type I TM proteins possibly related by gene duplication. The other two sarcoglycan subunits (Scgβ and Scgδ) are type II TM proteins with the Pfam domain Sarcoglycan_1, and they may not be evolutionarily related to Dg and Scgα.

Class P IG-fold domains in zona pellucida proteins. The Pfam Zona_pellucida (ZP) domain is found in a variety of extracellular proteins in metazoans [110]. It is about 260 residues long and contains two IG-fold domains [111]. Proteins with ZP domains have diverse functional roles ranging from structural components of ECM to cell surface receptors. *Dmel* has 18 genes encoding ZP domain-containing proteins. Most of them are predicted to be membrane-anchored, while two are predicted to be secreted. *Dmel* ZP genes are expressed in various epithelial tissues during embryogenesis [112]. Eight of them are required for reorganization of embryonic epidermal cells during morphogenesis, and they are likely involved in homo- and/or heterotypic interactions [113]. Two ZP genes (*miniature* and *dusky*) are required for apical membrane reorganization during wing epidermis differentiation [114]. Three ZP genes (*papillote*, *piopio*, and *dumpy*) are required for cell adhesion to apical ECM and microtubule organization. The protein encoded by *papillote* contains only the N-terminal IG-fold domain, unlike other *Dmel* ZP domain-containing proteins with two IG-fold domains. The gene *qsm* (*quasimodo*) encodes a GPI-anchored protein. It is a clock-controlled gene involved in light input to circadian clock [115]. ZP domain co-occurs with several N-terminal PAN(g) domains in proteins encoded by six *Dmel* genes. One of them is *nompA*, which encodes an apical ECM protein involved in mechanotransduction [116].

Other class P IG-fold domains—PKDREJ(g), Reeler, Spond_N, MATH, TIG, FixG_C, HYR, Na_K-ATPase, and WIF. The vertebrate multi-pass TM protein PKD1 has a large extracellular region that contains tandem PKD domains with the IG-fold (in the Pfam E-set clan) [117]. In addition, it possesses a region of about 400-amino-acid residues called the REJ (receptor for egg jelly) domain that is also present in the sperm receptor for egg jelly [118]. The region of REJ domain

is predicted to contain several PKD-like IG-fold domains. PKD and REJ domains are incorporated in the *Drosophila* XC domain named PKDREJ(g). *Dmel* has three multi-pass TM proteins (CG30048, CG42685, and PRY) with a domain structure similar to vertebrate PKD1, with PKDREJ(g), GPS, and an intracellular PLAT domain. Another protein CG7565 is also predicted to contain PKDREJ(g) as HMMER found weak hits to REJ, PKD, and other E-set IG-fold domains such as Y_Y_Y and Big_3_4. Functions of these proteins have not been experimentally characterized.

Two IG-fold Pfam domains, Reeler and Spond_N, are present in the N-terminus of *Dmel* protein Fat-spondin. The Reeler domain was originally discovered in vertebrate proteins Reelin and F-spondin. Crystal structures of Reeler domains revealed an IG-fold that could interact with other proteins or heparins [119, 120]. *Dmel* has seven genes encoding Reeler-containing proteins. Four of them (Fat-spondin, CG14515, CG17739, and CG30046) are F-spondin proteins with Reeler, Spond_N, TSP_1, and Kunitz_BPTI domains. Reeler co-occurs with a DOMON domain in the multi-pass TM protein ferric chelate reductase 1 (encoded by *CG8399*). Reeler is also present on its own in two secreted proteins (CG14515 and L(2)34Fc). Spond_N domain is structurally similar to the calcium-binding C2 domain and could interact with membrane and the ECM receptor integrin [121, 122]. Spond_N is present in Mspo (M-spondin) and the four F-spondin proteins in *Dmel*.

The IG-fold domain MATH (Meprin And TRAF-Homology) is found in both extracellular proteins (Meprins) and intracellular proteins (TRAFs) [123]. Meprins are extracellular metallopeptidases in vertebrates [124]. They have a modular domain structure with the metallopeptidase domain, a MAM domain, a MATH domain, and the membrane-proximal EGF domain [124]. TRAFs [tumor necrosis factor (TNF) receptor-associated factors] regulate cytokine signaling by interacting with TNF receptors through their MATH domains [125]. *Dmel* has five intracellular proteins with the MATH domain, such as Traf4 and Traf6. It also has a divergent XC MATH domain (detected by HHsearch and not by HMMER) found in the secreted or GPI-anchored product of *CG13247*.

Three additional IG-fold domains in the Pfam E-set clan (TIG, FixG_C, and HYR) are also identified in the *Dmel* proteome. The TIG domain is found in two plexin molecules (PlexA and PlexB) that serve as cell surface receptors for semaphorins [126]. TIG is also found in the type I TM protein Mesh, which has a modular extracellular region including VWD, AMOP, Sushi, and NIDO domains. Mesh, highly expressed in the digestive system, forms a complex with Ssk (Snakeskin, a four-pass TM protein) and is required for septate junction formation in *Dmel* midgut [127]. CG43394, a predicted type I TM protein, has a region predicted to be the IG-fold domain FixG_C. HYR (Hyalin Repeat) domain is another IG-fold domain [128] present in the products of two *Dmel* genes (*uif* and *frac*).

The sodium/potassium transporting ATPase beta chains, which are type II TM proteins, contain an IG-fold domain in the extracellular region (included in the Pfam family Na_K-ATPase) [129]. *Dmel* has six genes (*nrv1–3*, *CG5250*, *CG33310*, and *CG11703*) encoding these proteins. The WIF domain, found in Wnt antagonists (described below in section Class P XC domains that serve as antagonists of signaling molecules), also adopts the IG-fold.

IG-fold domains in other classes of XC domains. A number of IG-fold domains were identified in other classes of *Drosophila* XC domains, making the IG-fold the most versatile fold in the extracellular space. They include six class B domains (LIG(g), Neur_chan_LBD, CBM_39, DOMON, PBP, and MNNL), three class E domains (LIPO_10, Lysyl_oxidase, and Sod_Cu), and five class R domains (Alpha-amylase_C(g), CBM_20, Hemocyanin_C, CarboxypepD_reg(g), and Ceramidse_alk_C).

The Pfam LIG clan contains IG-fold domains that mainly bind lipids (Fig. 4D). Genes with two Pfam domains (DUF1091 and E1_DerP2_DerF2) in this clan are expanded in the *Dmel* genome. Proteins with the E1_DerP2_DerF2 domain are encoded by eight NPC2 genes (*Npc2a–h*). They are involved in sterol homeostasis and steroid biosynthesis [130]. In contrast to a single copy of NPC2 in other organisms such as yeast, worms, and mammals, the expansion of NPC2 genes in *Dmel* has allowed acquisition of new functions other than sterol transport. The secreted Npc2a, Npc2e, and Npc2f were found to bind a variety of bacterial cell wall components such as lipopolysaccharide (LPS), peptidoglycan, and lipoteichoic acid, suggesting their roles in immunity [131]. Npc2b was identified as a seminal fluid protein in male accessory gland and may have functions in reproduction [132]. DUF1091 domain is remotely related to E1_DerP2_DerF2 with the same cysteine pattern. DUF1091-containing genes are mainly found in *Diptera* insects and are greatly expanded in the *Dmel* genome (111 genes). Together with eight NPC2 genes, they make the LIG(g) domain the most abundant class B XC domain. The majority of DUF1091-containing genes have not been experimentally studied. A subset of these genes such as *CheA29a* and *CheB42a* were found to be expressed in chemosensory sensilla in the front legs of mature males and could be involved in male-specific chemical senses and pheromone response [133,134].

An IG-fold domain (Neur_chan_LBD) [135] was found in subunits of ligand-gated ion channels encoded by 23 *Dmel* genes. They are pentamer-forming multi-pass TM proteins [136]. The ligands for these proteins are neurotransmitters such as glycine, nicotinic acetylcholine, and gamma-aminobutyric acid (GABA).

CBM39 is an IG-fold domain found in several GNBPs (Gram-negative bacteria binding proteins) in *Dmel*. They play important roles as pattern recognition molecules that recognize and bind peptidoglycan and

β-1,3-glucan from bacterial and fungal cell walls [137, 138]. GNBP1, GNBP2, and GNBP3 share the same domain structure that contains a CBM39 domain and an inactivated Glyco_hydro_16 domain with a jelly roll fold. CBM39 is also found in two genes (*GNBP-like3* and *CG12780*) encoding secreted proteins without other XC domains.

The Pfam PBP (phosphatidylethanolamine-binding protein) domain with the IG fold [139] is detected in three predicted secreted proteins (A5, CG17917, and CG17919) and six predicted intracellular proteins (without predicted signal peptide or TM segment). *CG17917* and *CG17919* are neighboring genes with different expression patterns. *CG17919* has a broad tissue expression, while *CG17917* has a testis-specific expression. The gene *a5* (*antennal protein 5*) is expressed in a subset of olfactory hairs in antenna and could encode an odorant-binding protein [140]. Another extracellular lipid-binding domain with IG-fold is MNNL, found in the N-termini of Notch ligands Dl (Delta) and Ser (Serrate) [141]. MNNL domain is a C2-like domain that can bind phospholipids in a calcium-dependent manner [141].

The DOMON (dopamine β-monooxygenase N-terminal) domain [142] belongs to the Pfam clan CBD9-like, which also includes four other domains CBM9_1, CBM9_2, CDH-cyt, and Clucodextran_C. These domains are involved in heme and carbohydrate binding [142]. *Dmel* has eight genes encoding proteins that contain the DOMON domain. In all of them the DOMON domain co-occurs with an enzyme domain, including Cytochrom_B561, Cu2_monooxygen, DM13, and LPMO_10. These enzymes possess oxidoreductase activities, consistent with the heme-binding activity of the DOMON domain.

Three IG-fold domains [LIPO_10, Lysyl_oxidase, and Sod_Cu (Fig. 4E)], possess enzyme activities and were classified as class E domains. They are all copper-binding oxidoreductases (described below in section Class E XC domains that are oxidoreductase homologs (EC 1.-.-.-)).

Five IG-fold XC domains were classified as enzyme regulatory domains (class R) as they co-occur with enzyme domains. Two of them [CBM_20 and Alpha-amylase_C(g)] are derived from the Pfam GHD clan. The CBM_20 domain lies N-terminally to the GDPD (glycerophosphoryl diester phosphodiesterase) domain [143] in proteins encoded by five *Dmel* genes. Two of these genes (*CG9394* and *CG11619*) encode predicted secreted proteins, while two genes (*CG2818* and *CG3942*) encode predicted intracellular proteins (lacking predicted signal peptide and TM segment). Another gene (*CG18135*) encodes isoforms with and without a predicted signal peptide, suggesting varied subcellular localizations. Alpha-amylase_C(g) is associated with the Alpha-amylase domain in proteins encoded by 15 *Dmel* genes, 14 of which are predicted to be CSS proteins. Three other class R IG-fold domains are Hemocyanin_C, CarboxypepD_reg(g), and Ceramidse_alk_C. Functions of these domains are largely unknown. These domains could contribute to the structural stability of the associated enzyme domains, as observed for Ceramidse_alk_C [144]. As IG-fold domains, they may have additional roles in protein–protein interactions or protein–membrane interactions.

Class P XC EGF-like domains. The classical EGF domain is a small domain with six conserved cysteines forming three disulfide bonds [145]. It is composed of two subdomains, EGF-N and EGF-C, each of which has two β-hairpins. The N-terminal EGF-N subdomain has two disulfide bonds. The first one connects the N-terminal half of the first β-hairpin to the N-terminal half of the second β-hairpin (called NN-connection, colored magenta in Fig. 5), and the second disulfide bond connects the N-terminal half of the first β-hairpin to the C-terminal half of the second β-hairpin (called the NC-connection, colored blue in Fig. 5). The sequential order of the four cysteines are “abab”, where the a's form the NN-connection and the b's forms the NC-connection. The C-terminal subdomain has only one disulfide bond of the NC-connection type (Fig. 5). There is typically one residue separating the last cysteine of EGF-N and the first cysteine of EGF-C. The classical EGF domains thus have a sequential disulfide pattern of “ababcc”. They are represented in a number of Pfam domains such as EGF, EGF_CA, hEGF, cEGF, FXa_inhibition, EGF_2, EGF_3, and EB. Classical EGF domains can occur as a single copy and in tandem repeats.

Classical EGF domains were originally identified in signaling molecules such as epidermal growth factors. *Dmel* signaling molecules with classical EGF domains include Gurken, Spitz, and Vein that interact with the cell surface receptor Egfr (Epidermal growth factor receptor). Gurken and Spitz (encoded by *Grk* and *spi*) are TGF-α-like proteins [146] synthesized as type I TM precursors and are processed by proteolysis to form the soluble ligands with an EGF domain [147]. Vein is a secreted EGF ligand that also possesses an Ig(g) domain. While these signaling molecules for Egfr have only one EGF domain, tandem repeats of classical EGF domains are frequently found in CSS proteins, such as Notch and its signaling ligands Delta and Serrate. The cell surface receptor Notch, for example, has more than 30 EGF repeats. They contribute to the function of Notch in multiple ways, such as mediating its dimerization, protecting it from proteolysis, serving as a spacer, and interacting with Delta and Serrate [148]. The Nimrod gene cluster located on chromosome 2 has 10 genes encoding proteins with EGF domains, including both type I TM proteins and secreted proteins [149]. Like two other EGF-containing proteins Eater and Draper, NimC1 has been shown to be a phagocytosis receptor on hemocytes [150–152]. These proteins are important for removing microorganisms for host defense and apoptotic cells in developmental processes.

A number of other types of EGF-like domains exist, showing differences in their number of subdomains or β-hairpins and their disulfide bond connection patterns. Laminin_EGF, often occurring in tandem repeats, has one additional EGF-C subdomain compared to the classical EGF domain (Fig. 5) [153]. It has eight cysteines with the sequential order of "ababccdd". Laminin_EGF is found in various CSS proteins such as laminins, the GPCR cadherin Stan, and two signaling molecules (NetA and NetB). The DSL domain [154], found in Notch ligands Delta and Serrate, has three subdomains with NC-, NN- and NC-connections respectively (Fig. 5) and a sequential cysteine pattern of "aabbcc". Regions with Laminin_EGF and DSL domains often exhibit significant sequence similarity to classical EGF domains according to HMMER or HHsearch searches. Classical EGF domains, Laminin_EGF, and DSL are together classified as the *Drosophila* XC domain EGF(g).

A number of cysteine-rich domains possess structures consisting of β-hairpins similar to classical EGF domains, but showing variations in the number of β-hairpins and disulfide connectivity patterns (Fig. 5). They all possess at least one NC-connection (colored magenta in Fig. 5) between two neighboring β-hairpins commonly found in the EGF-N subdomain and the EGF-C subdomain. These EGF-like domains are found in numerous CSS proteins including signaling molecules and their receptors, cell adhesion molecules, and ECM proteins. These domains include Ldl_recept_a, Sushi, EMI, FOLN, GF_recep_C-rich(g), TNFR_c6, Ephrin_rec_like, NCD3G, TIL, VEGF_C, ADAM_CR, ADAM17_MPD, Argos, and Kringle (Fig. 5). Some of these EGF-like domains are described below.

Ldl_recept_a. Ldl_recept_a is a small cysteine-rich domain originally identified as a repeated module for ligand binding in low-density lipoprotein (LDL) receptors [155]. LDL receptors and LDL-receptor-related proteins (LRPs) have Ldl_recept_a repeats, EGF(g) repeats, and beta-propeller domain(s) of the Ldl_recept_b type. Structural studies suggested that Ldl_recept_a repeats interact with the LDL ligand in a neutral pH environment, and such an interaction is replaced by the intramolecular interaction with the beta-propeller domain at endosomal pH to discharge the ligand [156]. *Dmel* has seven genes encoding such proteins. Ldl_recept_a domain has four β-hairpins and three disulfide bonds with a connectivity of "abacbc" (Fig. 5). The third β-hairpin is widened to accommodate binding of a calcium ion (Fig. 5a). Ldl_recept_a is also present in a variety of other CSS proteins with modular domain contents. In total, 38 *Dmel* genes were found to encode Ldl_recept_a-containing proteins. Ldl_recept_a co-occurs with the second largest number of XC domains such as LRR(g), Beta_propeller_XC (g), CUB(g), EGF(g), and Trypsin(g) (Fig. 3b).

Ldl_recept_a domain co-occurs with CUB(g) domain in nine *Dmel* gene products, all of which are predicted type I TM proteins. These proteins likely function as regulators of other cell surface proteins such as signaling receptors and ion channels. The large product of *uif* (*uninflatable*) additionally contains EGF(g), Lectin_C(g), Laminin_G(g), F5_F8_type_C, and HYR domains. Uif is a negative regulator of the Notch signaling pathway [157]. The binding and trafficking of Uif with Notch through endosomes is important for asymmetric endosome motility and Notch-dependent cell fate assignation [158]. The other eight genes with Ldl_recept_a and CUB(g) domains appear to have no other XC domains. One of them, *Culd*, regulates endocytic trafficking of rhodopsin and TRPL channel and is required for the survival of photoreceptor cells [159]. Another gene, *Neto* (*Neuropilin and tolloid-like*) is required for the clustering of ionotropic glutamate receptors (iGluRs) at neuromuscular junctions and thus plays important roles in postsynaptic densities and synapse functionality [160].

Ldl_recept_a is found in Trol, a large basement membrane (BM) protein (over 4,000aa) that is an ortholog of vertebrate HSPG2 (Perlecan). Five Trypsin-type proteases (Ndl, Corin, Tequila, Modsp, and CG1632) have the Ldl_recept_a domain, four of them (except Modsp) also possess the SRCR(g) domain. Ldl_recept_a domain is present in four *Dmel* proteins with the Polysacc_deac_1 enzyme domain. A single copy of Ldl_recept_a domain is detected in the signaling molecule Jeb (Jelly belly). It interacts with the RTK Alk (anaplastic lymphoma kinase), which also has a Ldl_recept_a domain in its extracellular region together with other XC domains such as MAM, EGF(g), and Gly_rich. In addition, Ldl_recept_a co-occurs with LRR(g) repeats in two GPCRs—Lcr3 and CG34411.

Sushi. Sushi domain was identified in a number of vertebrate cell adhesion molecules and complement components [161]. It has four conserved cysteines in three β-hairpins. The last two β-hairpins have the NC connection characteristic of an EGF domain (Fig. 5). *Dmel* has 15 Sushi-containing genes with diverse domain contents. It is present in four genes (*Sr-CI*, *Sr-CII*, *Sr-CIII*, and *Sr-CIV*) encoding class C scavenger receptors [162] that also possess a MAM domain and a Somatomedin_B domain (except *Sr-CIII*). Sushi domain also co-occurs with a number of other domains such as Ldl_recept_a, EGF(g), Ig(g), and F5_F8_type_C in a variety of cell surface receptors and adhesion molecules.

EMI. The EMI domain is often found at the N-termini of CSS proteins [163]. Structures of two EMI domains from two human proteins, fibrillin and transforming growth factor beta-induced protein (TGFBIp), have been solved [163,164]. EMI domain has four conserved cysteines in three β-hairpins (Fig. 5). It has an NN-connection between the first and second β-hairpins and an NC-connection between the second and third hairpins. In particular, it has a "CC" sequence motif in the second β-hairpin. *Dmel* has six genes with EMI domains, including four Nimrod family genes

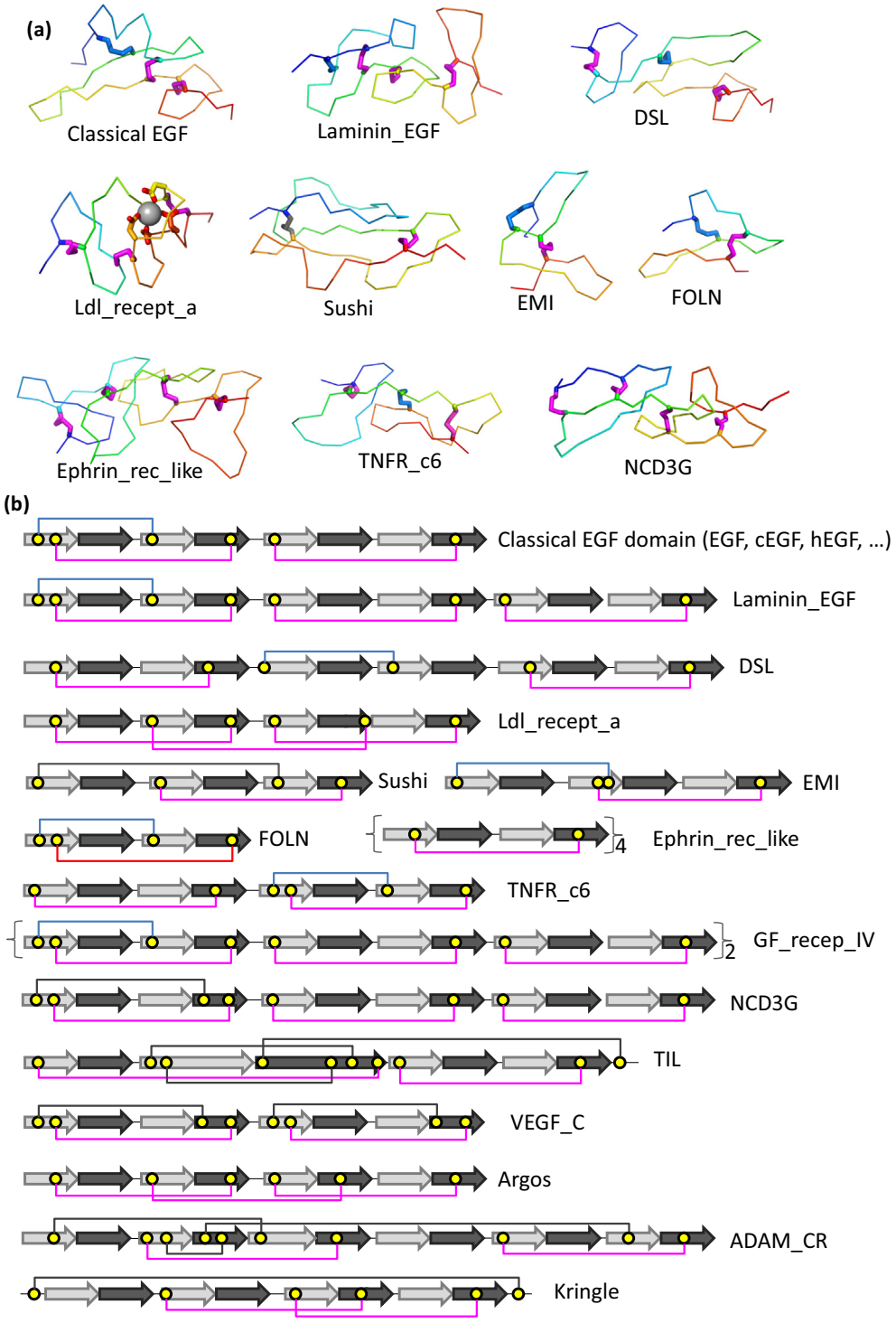


Fig. 5. Structures and cysteine patterns of EGF-like domains. (a) Structures of representative EGF-like domains. Backbone C-α trace of each structure is rainbow-colored from N-terminus (blue) to C-terminus (red). Disulfide bonds are in sticks. NN-connection, NC-connection, and other disulfide bonds are colored blue, magenta, and gray, respectively. The structures and domain ranges are: classical EGF domain, 3c9a, C50–C97; Laminin_EGF, 4plo, A404–A455; DSL, 4xbm, A178–A225; Ldl_recept_a, 2fcw, B86–124; EMI, 2m74, A56–A83; Sushi, 1ly2, A67–A129; FOLN, 2p6a, D137–D162; Ephrin_rec_like, 4m4p, A261–A327; NCD3G, 2e4a, A508–AA566; TNFR_c6, 1sg1, A119–A161. Calcium (gray ball) and calcium-binding residues in Ldl_recept_a are also shown. (b) Cartoon representations of EGF-like domains. Each β-hairpin is represented by a lighter gray arrow followed by a darker gray arrow. Cysteine positions are in yellow circles. NN-connection, NC-connection, and other disulfide bonds are shown as blue, magenta, and gray brackets connecting the cysteines, respectively.

(*NimA*, *NimB3*, *NimC3*, and *NimC4*), *dprp*, and *slow*. They all have the EMI as the N-terminal cysteine-rich domain followed by classical EGF domains.

FOLN. The FOLN domain has only two β-hairpins corresponding to the classical EGF-N subdomain (Fig. 5). This domain is stabilized by its interaction with the neighboring Kazal(g) domain [165]. *Dmel* has two genes with the FOLN domain. The gene *Fs* (*Follistatin*) is a regulator of TGF-β signaling during development [166]. The other gene (*SPARC*) encodes a BM protein that also contains the Kazal(g) domain and the SPARC_Ca_bdg domain [167].

EGF-like domains in signaling receptors and other CSS proteins. Several EGF-like domains are found in various cell surface receptors, including GF_recep_C-rich(g) in the EGF receptor Egfr and the insulin receptor InR, TNFR_c6 in the TNF receptors Wengen and Grindelwald, Ephrin_rec_like (previously named GCC2_GCC3) in the Ephrin RTK Eph, and NCD3G in three GPCRs (Fig. 5).

Class P domain LRR(g). Leucine-rich repeat is a common structural motif found in both extracellular and intracellular proteins. Each repeat unit has 20- to 30-amino-acid residues and contains the consensus motif LxxLxLxxN/CxL. The LRR repeats form an α/β-horseshoe structure and are often involved in protein–protein interactions [168]. The XC domain LRR(g) is frequently found in the extracellular region of cell surface receptors. One of them is Tl (Toll), which controls the immune response to pathogens such as Gram-positive bacteria and fungi [169]. Unlike vertebrate Toll-like receptors, *Dmel* Tl does not act as a pattern recognition receptor, but signals through the binding of the cytokine Spatzle [170] (Fig. 6a). The intracellular region of Tl contains a TIR (Toll/IL-1R receptor) domain, which is also found in several other Toll-like receptors. An intracellular complex of signaling adaptors assembles around the TIR domain after receptor activation, ultimately leading to to activation of gene expression of immune effectors [170]. Tl also plays a plethora of roles in development, such as dorsal-ventral polarity [171] and larval hemocyte formation and differentiation [172]. Besides Tl, eight Toll-like receptors (18w, MstProx, Tehao, Tollo, Toll-4, Toll-6, Toll-7, and Toll-9) with extracellular LRR repeats and an intracellular TIR domain exist in the *Dmel* proteome [173]. Some of them, such as 18w, Tehao, and Toll-9, could be involved in immune response against bacteria [174–176]. Some are highly expressed (18w, Toll-6, Toll7, and Tollo) during embryogenesis and metamorphosis, suggesting their roles in development [177].

LRR(g) domains are found to associate with five GPCRs. Two of them, Lgr1 and Rk, bind heterodimer Cystine-knot hormones Gpa2/Gpb5 and Burs/Pburs, respectively. Lgr3 is the receptor for Ilp8 (Insulin-like peptide 8) in controlling growth and body size [178]. The ligands of two additional LRR(g)-containing GPCRs, Lgr4 and CG15744, have not been experimentally revealed.

LRR(g) includes the Pfam domain Recep_L_ domain, a divergent LRR domain found in the L1 and L2 regions of InR and Egfr [179]. These two RTKs also share a Furin-like cysteine-rich domain in between L1 and L2. InR possesses four fn3(g) domains, while Egfr has two GF_recep_IV domains after L2. Two other genes, *Sdr* and *CG10702*, encode proteins with Recep_L_domain, Furin-like, and fn3(g) domains similar to InR. Sdr (Secreted decoy of InR) is a secreted protein that binds insulin-like peptides and acts as an antagonist of insulin/IGF signaling to restrict body growth [84]. CG10702 is a type I TM protein with an InR-like ecto-domain composition, but lacks the cytoplasmic tyrosine kinase domain. It could also be a negative regulator of insulin/IGF signaling given its similarity to Sdr and InR.

Some LRR(g)-containing proteins are involved in cell adhesion. Con (Connectin), a GPI-anchored protein, mediates homophilic cell–cell adhesion in the *Dmel* neuromuscular system and plays a role in axon guidance [180, 181]. Caps (Capricious) and Trn (Tartan) are two closely related type I TM proteins that are involved in cell interactions in various tissues [182–184]. Caps, Trn, and Fili (Fish-lips) could regulate cell survival and apoptosis via their cell adhesion properties [185, 186]. Two other LRR(g)-containing cell surface proteins, Haf (Hattifattener) and Conv (Convoluted), are regulators of neuromuscular axon targeting [187, 188]. Conv, a predicted GPI-anchored protein, is essential for tracheal tube morphogenesis and apical matrix organization [189]. Chaoptin is a photoreceptor cell-specific adhesion protein required for photoreceptor cell morphogenesis [190].

LRR(g)-containing proteins can serve as signaling molecules and their antagonists. One example is the *sli* (*slit*) gene, which encodes a signaling molecule that interacts with the Robo receptors. The Sli-Robo signaling is important for axon guidance in the nervous system as well as cell migration in the development of other tissues [78]. *Lrt* is expressed in tendon and promotes tendon-muscle targeting via its interaction with Robo receptors [191]. Lrt could be a signaling molecule itself or act as a modulator of Sli-Robo signaling. Both Sli and Lrt use their LRR repeats for Robo interactions. Six kekkon genes (*kek1–6*) exist in *Dmel*. They encode type I TM proteins with one extracellular Ig(g) domain and LRR repeats [192]. Kek1 is an antagonist of EGF signaling via its interaction with Egfr [193]. Kek5 modulates BMP signaling and is an antagonist of the BMP ligand Gbb (Glass bottom boat) [194]. The extracellular LRR(g)-containing protein Ltl (Larval translucida) is another antagonist of BMP signaling and regulates wing growth and vein patterning [195]. Wdp (Windpipe) is a type I TM protein that is a negative feedback regulator of the JAK/STAT signaling pathway in maintenance of intestinal

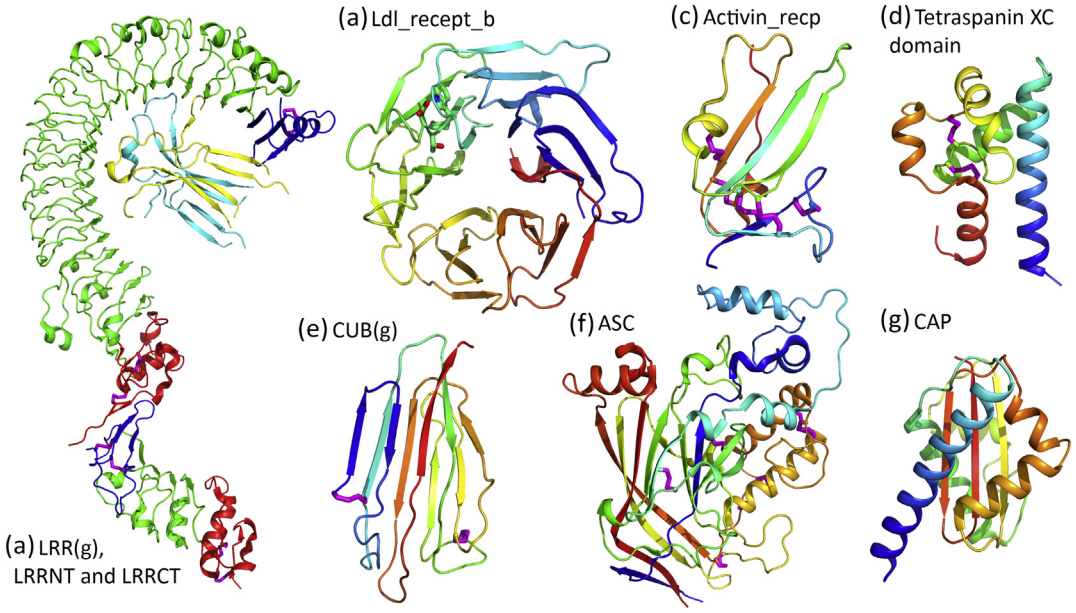


Fig. 6. Structures of representative class P domains. (a) *Dmel* Toll receptor extracellular region (4lxr) with LRR(g) (green), LRRNT (blue), and LRRCT (red) domains. The Spatzle ligands are shown in yellow and cyan. (b) A Ldl_recept_b beta-propeller domain (3sov, A20-A281) with the sidechains of the YWTD motif in the third blade shown. (c) An Activin_recp domain (2h62, C34–C117) belonging to the uPAR_Ly6_toxin(g) XC domain. (d) A tetraspanin XC domain (1g8q, A113–A202). (e) A CUB(g) domain (3kq4, B932–B1044). (f) the XC domain (ASC) of an EnaC channel (2qts, A72–A423). (g) a CAP domain (4ifa, A212–A344). Disulfide bonds are shown in magenta.

homeostasis. Windpipe interacts with the Dome receptor and promotes its internalization and lysosomal degradation [196].

The N-terminal region of LRR repeats often contains a β-hairpin capping motif where the first β-strand is antiparallel to the other β-strands in the LRR repeats [197]. This region corresponds to the LRRNT domain in Pfam. While in most cases, two disulfide bonds are present in the LRRNT domain, either one of them could be missing in some structures (e.g., pdb: 3t6q and pdb: 4mn8). Similarly, the C-terminal region of LRR repeats frequently contains a capping motif called LRRCT in Pfam with two conserved disulfide bonds [197]. Both LRRNT and LRRCT domains are present in *Dmel* Tl (Fig. 6A, colored blue and red, respectively). The high sequence diversity of LRRNT and LNNCT coupled with their short lengths prevented detection in some LRR-containing proteins at the default HMMER e-value or HHsearch probability score cutoffs.

Class P XC beta-propeller domains. More than 50 genes encode proteins with XC beta-propeller domains. Most of them (46) have the class P Beta_propeller_XC(g) domain. Members of this XC domain have been shown to engage in protein interactions, such as the Ldl_recept_b domain in LDL receptors [156] and the FG-GAP domain in the integrin α subunits [108]. A few genes possess XC beta-propeller domains classified as enzyme homologs (class E) or enzyme regulators (class R).

MRJP. The largest gene family containing an XC beta-propeller domain is the yellow family [198] with 14 genes. All except one encode secreted proteins (*yellow-b* encodes a predicted GPI-anchored protein). They possess the Pfam MRJP domain (a six-bladed beta-propeller) that has been found in a family of major royal jelly proteins in honey bee [199]. The *y* (*yellow*) gene is necessary for production of black melanin and regulates cuticle pigmentation during development [200]. This gene is also required in normal male courtship behavior and mating success [201].

Ldl_recept_b. The Ldl_recept_b domain corresponds to the YWTD repeats found in six-bladed beta-propeller domains (Fig. 6b) in the LDL receptor family proteins [66]. Nine *Dmel* genes encode Ldl_recept_b-containing proteins with large modular extracellular regions. Among them, Ndg (Nidogen) is a secreted protein and a prominent component of the BM [202]. The other eight genes encode type I TM proteins that also include EGF(g) repeats and Ldl_recept_a domains (except the gene *cue*). LpR1 and LpR2 are lipophorin receptors involved in the uptake of neutral lipids from circulating apolipoproteins [203]. The *yl* (*yolkless*) gene encodes a vitellogenin receptor [204]. Arr (Arrow) is a type I TM protein that functions as a co-receptor for Wingless in the canonical Wnt signaling pathway [205]. Mgl (Megalin) regulates

cuticle pigmentation by promoting endocytosis of the Yellow protein [206].

FG-GAP/VCBS. The FG-GAP seven-bladed beta-propeller domains [207] are present in the five integrin α subunits of *Dmel*. The regions with Pfam hits to FG-GAP domains often have significant HMMER scores to VCBS, another Pfam beta-propeller domain that has been found in bacteria *Vibrio*, *Colwellia*, *Bradyrhizobium*, and *Shewanella* (hence the name). Three additional FG-GAP/VCBS-containing genes (*CG6184*, *CG3618,* and *CG7739*) without known function exist in *Dmel*. They encode type I or type II TM proteins.

Sema. The Sema domain with seven-bladed beta-propeller fold is found in the semaphorin family proteins consisting of several signaling molecules and their receptors [208]. In *Dmel*, Sema2a and Sema2b are two secreted signaling molecules, while Sema1a, Sema1b, Sema5c, PlexA, and PlexB are type I TM proteins. Sema1a, PlexA, and PlexB serve as receptors for Sema2a and Sema2b [209,210]. Sema1a can also act as a signaling molecule involved in juxtamembrane signaling via the interaction with PlexA [211].

EPTP. Three genes, *clos*, *fs(1)M3*, and *fs(1)N*, encode minor proteins important for maintaining the integrity of *Dmel* eggshell vitelline membrane [212, 213]. Both Clos and Fs(1)M3 contain the EPTP domain predicted to have the beta-propeller fold [214]. Fs(1)N could be their remote homolog as weak HHsearch hits to EPTP domains (probability score: 86.3) were detected for it.

Kelch. The Kelch-type beta-propeller domains were mostly discovered in intracellular proteins as protein–protein interaction modules [215]. *Dmel* genes *dsd* and *CG7466* encode two extracellular proteins that contain XC Kelch repeats together with CUB(g), EGF(g), and PSI domains. The functions of these genes are unknown.

OLF. OLF (Olfactomedin-like) domain has a beta-propeller fold with six blades. In *Dmel*, it is found in a single gene *CG6867* with two Ig(g) domains and the Collagen domain. It could be a component of ECM [90, 91].

Other Dmel genes with extracellular beta-propeller domains. Divergent beta-propeller domains were also found by HHsearch (not by HMMER) in the receptor Sevenless [66] and two teneurin proteins (Ten-m and Ten-a). Teneurins are type II TM proteins that function as synaptic organization proteins [216]. Two beta-propeller domains (SGL(g) and NHL) are classified in the enzyme class (class E) (described below). Another two beta-propeller domains (DPPIV_N and Hemopexin) are classified in the class R as enzyme regulatory and inhibitory domains.

Class P domain uPAR_Ly6_toxin(g). We identified 39 *Dmel* genes encoding proteins with the uPAR_Ly6_toxin(g) domain. This group of cysteine-rich domains include Pfam families in the uPAR_Ly6_toxin clan (Activin_recp, UPAR_LY6, UPAR_LY6_2, Toxin_TOLIP (Toxin_1), BAMBI, and PLA2_inh), as well as several other remotely related Pfam domains (such as QVR, DUF753, Ly-6_related, and DUF4723) that can be found by HHsearch with similar cysteine patterns. These domains are in various CSS proteins including receptors and snake toxins. The prototype domain was first identified in the receptor for mammalian urokinase-type plasminogen activator (uPAR) and Ly-6 molecules that are lymphocyte differentiation antigens [217]. The Activin_recp domain (Fig. 6c) in this group is in five TGF-β receptor serine/threonine kinases (Put, Sax, Tkv, Wit, and Babo) that are type I TM proteins that interact with TGF-β ligands. The other uPAR_Ly6_toxin(g)-containing genes mostly encode proteins with predicted GPI anchors. *Boudin*, *crooked*, *coiled*, and *crimpled* are required for the assembly of septate junction [218, 219]. The gene *qvr* (*quiver*, formerly called *sleepless*) encodes a protein with the Pfam QVR domain. *Qvr* promotes sleep by up-regulating the activity of the Shaker potassium channels and antagonizing nicotinic acetylcholine receptors [220, 221]. Retroactive, another protein with the QVR domain, is required for cuticle organization and epithelial tube growth [222, 223].

Tetraspanin, ASC, and other class P XC domains in multi-pass TM proteins. Quite a few class P XC domains are present in multi-pass TM proteins. Tetraspanin and ASC are the two most abundant ones in terms of the number of genes containing them. Tetraspanins are a large family of proteins characterized by four TM segments. They act as molecular facilitators that associate with cell surface signaling complexes [224, 225]. Tetraspanins have a small XC domain between the third and fourth TM segments. It has four conserved cysteines forming two disulfide bonds [226] (Fig. 6d). Some tetraspanins have two additional cysteines possibly forming a third disulfide bond. Tetraspanin XC domain was proposed to be involved in protein–protein interactions [227]. *Dmel* has 38 genes encoding tetraspanin proteins. Most of them are poorly studied. Lbm (Late bloomer) is the first characterized *Dmel* tetraspanin that facilitates synapse formation [228]. A later analysis of 35 *Dmel* tetraspanin genes revealed that they mainly fall into three tissue expression patterns: the nervous system, the gut, and low or high overall expression [229]. *Tsp2A* is highly expressed in midgut and hindgut and was shown to regulate septate junction formation [230]. Tsp42Ej primarily resides in the lysosome and has been shown to promote rhodopsin degradation [231]. Tsp3A, Tsp86D, and Tsp26D promote Notch signaling by regulating the trafficking of the ADAM protease Kuz [232].

ASC domain (Fig. 6f) is found in the degenerin (DEG)/epithial Na+ channel (ENaC) gene family products [233], also known as the pickpocket genes

in *Dmel* [234]. A total of 31 such genes encode double-TM proteins that form homotrimers or heterotrimers and serve as acid-sensing sodium channels for mechanical nociception. Besides Tetraspanin and ASC, class P XC domains TipE_CaKB(g), Prominin, Myelin_PLP, Meckelin, and TM231 are also associated with multi-pass TM proteins.

Several class P domains (Methuselah_N, Fz(g), HRM, GAIN, and NCD3G) are associated with GPCRs. The Methuselah_N domain is present in the N-terminal extracellular region of Mth (Methuselah), a GPCR involved in aging and stress response [235, 236]. *Dmel* has 15 additional methuselah-like genes (*mthl1–15*) with this domain. While this domain appears to be arthropod-specific, the methuselah-type GPCRs have a broader phylogenetic distribution including protosotomes, invertebrate deuterostomes, and cnidarians [237].

The Pfam Fz (Frizzled) domain is a cysteine-rich domain often found in GPCRs and RTKs [238, 239]. In *Dmel*, these proteins include four Frizzled receptors (Fz1–4) in Wnt signaling, Smo (Smoothened) in Hedgehog signaling, and two RTKs (Nrk and Ror). Fz domain is also present in type II TM protein Corin, a trypsin-type protease with a modular extracellular region that includes a SRCR(g) domain and a Ldl_recept_a domain. CG1632 has a similar domain structure as Corin, except that it has an additional juxtamembrane SEA(g) domain. Another type II TM protein CG6739 has one Fz domain and several Ldl_recept_a domains. The *Drosophila* XC domain Fz(g) also include Pfam domains Glypican and Mid1. Two glypicans, Dally and Dally-like, contain a cysteine-rich domain (Pfam: Glypican) remotely related to Fz domain [239]. Glypicans are GPI-anchored heparin sulfate proteoglycans that serve as ECM receptors, play critical roles in Wnt and Hedgehog signaling pathways [240]. Another domain distantly related to Pfam Fz domain is found in Mid1, a calcium channel subunit [239, 241].

Class P domain CUB(g) and other jelly roll-fold XC domains. CUB (*C*omplement C1r/C1s, *U*egf, *B*mp1) domain [242] is found in a functionally diverse set of proteins and can associate with a variety of other XC domains, such as Ldl_recept_a, EGF(g) and Trypsin (g) (Fig. 3b). With a jelly roll fold (Fig. 6e), CUB(g) domain could be involved in protein–protein interactions. For example, the CUB(g) domains in vertebrate neuropilins are responsible for semaphorin binding [243]. The Pfam family CRF-BP (corticotropin-releasing factor binding protein) [244] is distantly related to the Pfam CUB domain, and its region corresponds to two tandem CUB domains as suggested by the HHsearch results. *Dmel* has 32 genes encoding proteins with CUB(g) (CUB or CRF-BP) domains.

Like the IG-fold, the jelly roll fold is a versatile β-sandwich fold identified in a number of other class P XC domains (ASC, TGFb_propeptide, Laminin_N, MAM, ADAM_Spacer1, Laminin_B, Calreticulin, TSP_C, and Ephrin_lbd), in one class S domain (TNF), three class B domains (Laminin_G(g), F5_F8_type_C, and PLAT), two class E domains [Glyco_hydro_16 and Cu2_monoox(g)], and one class R domain (P_proprotein) [45].

Class P domain CAP. CAP [*C*ysteine-rich secretory proteins (CRISPs), *A*ntigen 5 (Ag5), and *P*athogenesis-related 1 (Pr1) proteins] domain (Fig. 6g) is found in a large family of mostly secreted proteins involved in reproduction and immunity [245]. CAP domains are functionally diverse, and they are involved in protein–protein interactions [246] and protein-lipid interactions [247]. *Dmel* has an expanded set of CAP-containing genes (29 detected) that fall into two main groups [248] with varying numbers of cysteines. They also show various tissue expression patterns. A subset of them, such as *scpr-A*, *scpr-B*, *scpr-C,* and *antr*, have enriched expression levels in testis or male accessory gland, indicating possible roles in male reproduction [248]. *Ag5r* and *Ag5r2* are preferentially expressed in salivary gland and midgut, respectively.

Class P XC domains in ECM proteins. Dmel ECM consists of the apical (external) part on the body surface (exoskeleton of cuticle) and the basal (internal) part inside the body [202]. *Dmel* cuticle is mainly made up of chitin polysaccharide fibrils and a variety of structural proteins such as chitin-binding proteins (described below in class B domains). The internal ECM is mainly the BM that separates cells of internal organs and epidermis from hemolymph. *Dmel* BMs have similar components to that of vertebrates, including proteins such as laminins, collagen type IV, nidogen, and proteoglycans.

Class P XC domains in BM proteins. A laminin complex is a heterotrimer consisting of three chains (α, β, and γ). *Dmel* has two laminin complexes that differ in the α chain [LanA and Wb (Wing blister)] while sharing the β chain (LanB1) and the γ chain (LanB2). All *Dmel* laminin chains contain Laminin_N domain at the N-terminus, EGF repeats of the Laminin_EGF type, and coiled coil domains (Laminin_I and Laminin_II in LanA and Wb) involved in heterotrimerization. Laminin_B domain is detected in the two α chains and the γ chain. The two α chains additionally have several Laminin_G(g) domains at their C-termini. Laminin_N, Laminin_B, and Laminin_G(g) domains all adopt the jelly roll fold [249–251]. Laminin_N domain is involved in polymerization of laminins [250]. Laminin_N domain is also detected in the two netrin proteins (NetA and NetB), which are ECM proteins and signaling ligands for the receptor Frazzled. Netrins are probably derived from laminins as they also contain the Laminin_EGF repeats.

Another major component of BM is collagen type IV, which is an ancient type of collagen [252]. *Dmel* has two genes (*Cg25C* and *vkg*) encoding these collagens, which are secreted proteins with collagen repeats (Pfam: Collagen) and two C4 domains at their C-termini. The Collagen domain is also present in several other secreted or membrane-anchored proteins in *Dmel*. Six of them (CG14089, CG31268, CG31437, CG42342, CG5225, and Elal) do not have other detectable Pfam domains. Among them, CG42342 is a type II TM protein, while others are secreted. Collagen domain co-occurs with Ig(g) domain and a beta-propeller domain (OLF) in the type II TM protein CG6867. It also co-occurs with Endostatin and Laminin_G(g) domains in the secreted protein Mp (Multiplexin). Mp is the ortholog of vertebrate collagen types XV and XVIII. It plays a role in motor axon pathfinding [253] and normal muscle function [254].

Perlecan is an HSPG conserved in bilaterians as well as some basal metazoan lineages such as Cnidarians and Placozoa [255]. *Dmel* perlecan, encoded by the *trol* gene, is a secreted protein with various XC domains such as EGF(g), Ig(g), SEA(g), Laminin_G(g), Laminin_B, and Ldl_recept_a. Three other HSPGs in *Dmel* are membrane anchored by GPI (Dally and Dally-like) or by a TM segment (Syndecan) [256]. They serve as ECM receptors. Another core BM component is Nidogen, which also contains several XC domains such as EGF(g), Beta_propeller_XC(g) (Ldl_recept_b type), NIDO, and G2F.

Class P XC domains in Dmel eggshell proteins. The multi-layered *Dmel* eggshell is a specialized ECM architecture [257]. Three layers of the eggshell mainly contain proteins—the oocyte proximal vitelline membrane (VM), the inner chorionic layer, and the outer endochorion [258]. The VM layer has four major structural proteins (encoded by *Vm26Aa*, *Vm26Ab*, *Vm32E*, and *Vm34Ca*) [259, 260] with the Pfam Vitelline_membr domain that has three conserved cysteines. Two other genes (*Vm26Ac* and *Vml*) also encode proteins with this domain and are mainly expressed in the ovary. A minor protein encoded by *psd* (*palisade*) is essential for vitelline membrane assembly [261]. It is predicted to be largely disordered and does not have known Pfam domains. Three genes, *clos*, *fs(1)M3*, and *fs(1)N*, are also minor proteins maintaining the integrity of vitelline membrane [212, 213]. As described above, they are predicted to contain the EPTP domain of the beta-propeller fold.

Several genes have restricted expression in the ovary and encode chorion-specific proteins with known Pfam domains. Two proteins Cp36 and Cp38 (previously named s36 and s38), encoded by two neighboring X chromosome genes, are abundant structural components of chorion [262]. They are homologs with the Pfam Chorion_3 domain. Three chorion proteins (Cp15, Cp18, and Cp19) [263] possess the low complexity Pfam domain Chorion_2. One of them (Cp19) also has a C-terminal low complexity Pfam domain S19. Another chorion protein Cp16 (previously named s16) [264] contains the Pfam domain Chorion_S16. *Cp15*, *Cp16*, *Cp18*, and *Cp19* are neighboring genes on chromosome 3. The *dec-1* (*defective chorion 1*) gene [265] encodes several isoforms of chorion-specific proteins with an N-terminal DEC-1_N domain, several Dec-1 repeats in the middle region, and the C-terminal DEC-1_C domain. These domains all contain low complexity regions. Several minor proteins were identified in a study of fractionated eggshell matrices using mass spectrometry [266]. Some of them contain known Pfam domains, such as Muc2B with the CBM_14 domain. Three genes (*Cp7Fa*, *Cp7Fb*, and *Cp7Fc*) without known Pfam domains reside in a chorion gene cluster that also includes *Cp36* and *Cp38* [267]. In another large-scale study of genes expressed in eggshell [268], several predicted secreted proteins (e.g., CG13113, CG13299, CG14187, and CG13998 without known Pfam domains) could be chorion proteins.

Class P XC domains that serve as antagonists of signaling molecules. Several class P domains are found in extracellular signaling molecule antagonists. The gene *aos* (*argos*) encodes an antagonist of EGF signaling [43]. Argos (with the Pfam domain Argos) functions through direct binding of EGF ligands [269]. The Argos domain adopts an EGF-like fold (Fig. 5). Tsg and CHRD are two domains in proteins that act as TGF-β binding proteins and antagonists. Tsg, a cysteine-rich domain, is found in the products of three *Dmel* genes (*tsg*, *srw*, and *cv*). Four CHRD domains and four VWC domains are present in the product of *sog* (*short gastrulation*). *Dmel* Tsg and Sog form a complex with the TGF-β signaling heterodimer made up of Dpp and Scw. They facilitate the transport of Dpp/Scw in the ECM and play crucial roles in dorsal ventral patterning in *Dmel* [270]. The WIF domain with the IG-fold [271] is found in antagonists of the Wnt signaling. *Dmel shf* (*shifted*) encodes a secreted protein with a WIF domain and multiple EGF(g) domains. This gene is a positive regulator of Hedgehog signaling [272]. The WIF domain is also identified in three *Dmel* RTKs (Drl, Drl-2, and Dnt).

Class S XC domains mainly found in extracellular signaling molecules

Class S domains, mainly found in XC signaling molecules, are involved in interactions with cell surface receptors to transduce signals across the cell membrane. *Dmel* possesses genes that encode a variety of secreted and membrane-bound signaling molecules, including peptide hormones, neuropeptides, cytokines, and growth factors. They are involved in endocrine, paracrine, autocrine, and juxtacrine signaling events. Peptide hormones are synthesized in glands, secreted into the circulatory system, and are mainly involved

in long-range actions in the endocrine system. *Dmel* neuropeptides [273] act as neuromodulators in the central and peripheral nervous system. Some signaling peptides can function as both neuropeptides and peptide hormones released into circulation. *Dmel* cytokines mediate local and systemic immune responses. Signaling growth factors belong to a broad category of protein ligands that mostly have short range growth effects on the cells that produce them (autocrine system) or on nearby cells (juxtacrine and paracrine systems). They include morphogens such as Wingless and Hedgehog that form a spatial gradient to help determine cell fates during developmental processes. Membrane-anchored signaling molecules, such as Delta, Serrate, Ephrin, and Boss, are involved in juxtacrine signaling. Some signaling molecules contain domains with versatile functions, such as EGF(g), Ldl_recept_a, and Cadherin(g) described above as class P domains, while most of the proteins with these domains do not serve as signaling molecules. On the other hand, a set of XC domains classified in class S are exclusively or mostly found in extracellular signaling molecules (bold domains in Table 2).

Seven Pfam domains, namely, TGF_beta, Spaetzle, Cys_knot, DAN, PDGF, Noggin, and IL17, possess the Cystine-knot fold and are grouped in the XC domain Cystine-knot(g) [274]. Most of these domains share six conserved cysteines forming three disulfide bonds, one of which crosses an intramolecular loop formed by the other two [275]. They exhibit great variation in terms of sequence diversity, the number and location of additional intramolecular disulfide bonds, and the ability to form intermolecular disulfide bonds. *Dmel* Cystine-knot(g)-containing proteins play diverse roles as growth factors (with TGF_beta, PDGF, and Noggin domains), cytokines (with the Spaetzle domain), and peptide hormones (with the Cys_knot, DAN, and IL17 domains). They act on various types of cell surface receptors such as receptor kinases and GPCRs (Table 2). Besides Cystine-knot(g), eight class S domains are present in multiple *Dmel* genes, such as Insulin, SVWC, and wnt. On the other hand, 18 class S domains are present in only one gene. *Dmel* also possesses a variety of neuropeptides and peptide hormones not included in the Pfam database.

Class S domains in signaling molecules that are growth factors and mainly function in paracrine systems. Signaling molecules that are growth factors in paracrine systems include proteins with these class S domains—Cystine-knot(g) (TGF_beta, PDGF, and Noggin), FGF, wnt, HH_signal, and FOG_N. They also include proteins with class P domains such as EGF(g) (in Grk, Spi, Vn, and Krn), Sema (in Sema2a and Sema2b), LRR(g) (in Sli) and Ldl_recept_a (in Jeb) (Table 2).

Cystine-knot(g) growth factor domains—TGF_beta, PDGF, and Noggin. Among the seven Cystine-knot Pfam families found in *Dmel* proteins, three (TGF_beta, PDGF, and Noggin) are in growth factors mainly involved in paracrine signaling.

Seven genes (*dpp*, *scw*, *gbb*, *mav*, *Actbeta*, *daw*, and *myo*) encoding TGF-β signaling ligands are found in the *Dmel* genome [276, 277]. They share the same domain architecture with an N-terminal propeptide domain (Pfam: TGFb_propeptide) and the C-terminal signaling domain (Pfam: TGF_beta). The propeptide domain is cleaved intracellularly, but remains bound to the signaling domain after secretion to the extracellular space. Removal of the propeptide domain activates the TGF-β signaling ligands, which can exist as homodimers or heterodimers such as Dpp/Scw. The TGF-β receptor is a ligand-induced transient heteromeric complex of two receptor serine/threonine kinases called a type I receptor and a type II receptor (both are type I TM proteins). Ligand binding activates the receptor kinase and leads to phosphorylation of intracellular Smad proteins, which translocate to nucleus and regulate transcriptional responses of a variety of target genes. Two branches of TGF-β signaling exist in *Dmel*. The BMP branch includes signaling ligands Dpp, Scw, Gbb, and Mav, type I receptors Sax and Tkv, type II receptors Wit and Punt, and the phosphorylation target Mad. The activin branch includes signaling ligands Actbeta, Daw, and Myo, type I receptor Babo, either Wit or Punt as type II receptor, and the phosphorylation target Smox [276]. The five receptor proteins (Sax, Tkv, Wit, Punt, and Babo) all possess the Activin_recp domain in the extracellular space responsible for binding TGF-β ligands.

Three PDGF domain-containing genes (*Pvf1–3*) are found in the *Dmel* genome. They share the same domain architecture: N-terminal PDGF domain and C-terminal cysteine-rich domain with CxCxC motif that can bind heparin. The PDGF/VEGF receptor is Pvr, an RTK with extracellular Ig(g) domains. The PDGFs and their receptor play essential roles in oogenesis, salivary gland guidance and migration as well as hemocyte migration during embryonic development [278–280].

The Cystine-knot domain Noggin is found in the product of *trk* (*trunk*). It signals through the receptor Torso (encoded by *tor*), which is responsible for gene activation in the anterior and posterior ends of the embryo through the Ras pathway. Torso is an RTK with three XC fn3(g) domains. It is also the receptor for another Cystine-knot ligand, the peptide hormone Ptth (prothoracicotropic hormone) [281]. Ptth possesses the IL17 domain of the Cystine-knot fold and plays regulatory roles in developmental timing and body size [282].

FGF. Three genes *bnl*, *pyr*, and *ths* encode proteins with the FGF (fibroblast growth factor) domain [283]. These genes are involved in developmental processes such as branching morphogenesis of organs and mesodermal migration. Two FGF receptors (encoded by *htl* and *btl*) are RTKs with extracellular Ig(g)

domains. The FGF domain adopts the β-Trefoil fold and binds FGF receptors and HSPGs [284].

wnt. Seven *Dmel* proteins possess the wnt domain: Wg (Wingless), Wnt2, Wnt4, Wnt5, Wnt6, Wnt10, and WntD. The most studied member, Wg, is a morphogen involved in short-range and long-range paracrine signaling in embryonic segment polarity and tissue patterning of wings and other body structures [285]. The Wg ligand binds the Frizzled receptors such as Fz2 and the co-receptor Arr (Arrow) and signals through the classical Wnt pathway [285]. The movement of Wg and its graded distribution is regulated by its interactions with the ECM and endocytosis [285]. Other wnt domain-containing proteins also play key roles in tissue development. Wnt2 is involved in muscle cell development and pigment cell origin [286, 287]. Wnt4 regulates cell motility in ovary [288], dorsoventral specificity of retinal projections [289], and synaptic target specificity [290]. Wnt5 signals through atypical RTK Drl (Derailed) in axon guidance during central nervous system development [291]. Wnt6 likely plays a role in maxillary palp formation [292]. *Wnt10* is expressed in the embryonic mesoderm, central nervous system, and gut , but the details of its function remain to be elucidated [293]. WntD (Wnt inhibitor of Dorsal) acts as a feedback inhibitor of Dorsal (a nuclear factor-κB homolog) and plays crucial roles in both embryonic development and host defense [294, 295]. WntD binds the Frizzled receptor Fz4 and is also considered an immunosuppressor cytokine based on its role in immunity [296].

HH_signal. Hedgehog (encoded by *hh*) is a morphogen contributing to segment polarity determination during embryonic development [297]. Hedgehog signaling is also crucial for stem cell maintenance and neuronal cell migration [298]. The secreted Hedgehog ligand binds its membrane receptor Patched (encoded by *ptc*) to release the inhibition of the GPCR family protein Smoothened and activate downstream signal transduction events. The Hedgehog ligand is generated from its precursor protein after translocation to ER and signal peptide removal. Hedgehog is further modified N-terminally with palmitoylation and autoprocessed to remove the C-terminal Hint domain and covalently attach a cholesterol group. The C-terminal Hint domain is then degraded via the ER-associated degradation (ERAD) pathway [299]. The Hint domain is thus not secreted to the extracellular space and not classified as an XC domain despite its co-occurrence with the N-terminal HH_signal domain. HH_signal domain binds zinc and is remotely related to metalloproteases in the Pfam Peptidase_MD clan [300].

FOG_N. Dmel fog (*folded gastrulation*) encodes a secreted signaling molecule that binds the methuselah-like GPCR Mthl1 and plays an important role in epithelial morphogenetic shape changes during gastrulation [301, 302]. It also regulates motor axon guidance and glial organization in the nervous system [303]. FOG_N domain corresponds to the N-terminal region of Fog (730aa) with two conserved cysteines. It exhibits weak similarity to an EGF subdomain and the cytokine Gbp1, which binds another methuselah-like GPCR Mthl10 (described below). The C-terminal region of Fog is predicted to be largely disordered.

Class S domains in membrane-anchored signaling molecules in juxtacrine systems. Several juxtamembrane signaling systems between membrane-bound ligands and their cell surface receptors exist in *Dmel* (Table 2). Examples of juxtamembrane signaling ligands include Delta and Serrate (with EGF and MNNL domains) that bind the receptor Notch [304] as well as the GPCR-linked signaling ligand Boss that binds the receptor Sevenless [305]. Two membrane-bound cadherins (Ds and Fat) are involved in juxtamembrane signaling in the hippo signaling pathway [306]. The Pfam Ephrin domain is found in the TM protein Ephrin in *Dmel* [307], which signals through the RTK Eph and plays crucial roles in axonal path finding during embryonic central nervous system development.

Class S domains in signaling molecules functioning as cytokines. Five class S domains (Unpaired, TNF, Diedel, SVWC, and Spaetzle from the Cystine-knot(g) XC domain group) are mainly found in cytokines functioning in *Dmel* immunity (Table 2).

Unpaired. Three unpaired genes (*upd1–3*) encode cytokines that activate the JAK/STAT signaling pathway in *Dmel* development and immune response [308, 309]. These signaling ligands bind the cell surface receptor Dome with Ig(g) and fn3(g) domains [310].

TNF. The gene *egr* encodes the only TNF-domain containing protein (Eiger) in *Dmel* [311]. Eiger precursor, a type II TM protein, is processed to release the XC domainsas a soluble factor. Eiger binds two TNF receptors Wengen (encoded by *wgn*) [312] and Grindelwald (encoded by *grnd*) [313]. Wengen has a typical XC TNF receptor domain (Pfam: TNFR_c6) (an EGF-like domain shown in Fig. 5), while Grindelwald has a divergent copy of this domain (not detected by HMMER or HHsearch with significant scores). Like the mammalian TNF family proteins, Eiger is capable of inducing cell death via the JNK pathway [311]. Secreted by the fat body cells as an adipokine and metabolic hormone, Eiger also plays a critical role in mediating nutrient response by acting upon insulin-producing cells [314].

Diedel. The Pfam Diedel domain (formerly DUF4002) is found in proteins encoded by three genes (*Diedel*, *Diedel1*, and *Diedel2*). Diedel is a small protein adopting a ferredoxin-like fold with 10 conserved cysteines [315]. Diedel is upregulated after septic injury and may act as a negative regulator of the JAK/STAT signaling pathway [316]. Diedel and its viral homolog also suppress the IMD (immune deficiency) pathway of *Dmel* [317]. As an immune response protein, Diedel could function as a cytokine. However, the receptor of

Table 2. *Drosophila* signaling molecules (Sig.Mol.), their XC domains, and their receptors

Sig.Mol.	XC domains[a]	Receptors[b]
Paracrine signaling		
Grk,Spi,Vn,Krn	EGF	Egfr
NetA,NetB	EGF,NTR	Fra,Unc5
Sli	EGF,LRR	Robo1–3
Jeb	Ldl_recept_a	Alk
Dpp,Gbb,Scw,Mav	**[TGF_beta]**	Tkv,Sax,Punt,Wit
Actbeta,Daw,Myo	**[TGF_beta]**	Babo,Punt,Wit
Pvf1–3	**[PDGF]**	Pvr
Trk	**[Noggin]**	Tor
Ths,Pyr,Bnl	**FGF**	Htl,Btl
Wg,Wnt2,4,5,6,10,D	**wnt**	**Fz,Fz2–4**,Arr,Drl
Hh	**HH_signal**	Ptc
Fog	**FOG_N**	**Mthl1**
Sema2a,Sema2b	Sema,PSI	Sema1a,PlexA,PlexB
Juxtacrine signaling		
Dl,Ser	EGF,MNNL	N
Boss	–	Sev
Ephrin	**Ephrin**	Eph
Ds	Cadherin	Ft
Sema1a,Sema1b	Sema,PSI	PlexA
Sas	VWC, fn3	Ptp10D
Cytokine signaling		
Upd1–3	**Unpaired**	Dome
Spz,Spz3–6,NT1,CG17672	**[Spaetzle]**	Tl
Egr	**TNF**	Wgn,Grnd
Vago	**SVWC**	–
Diedel,Diedel1–2	**Diedel**	–
WntD	**wnt**	**Fz4**
Gbp1	–	Egfr,**Mthl10**
Male-to-female signaling		
SP,Dup99B	**Sex_peptide**	**SPR**
Sfp93F,five others[c]	**Omega_toxin**	–
Acp26aa	**MAGSP**	–
Acp26ab	**Acp26ab**	–
Peptide hormones and neuropeptides		
Ilp1–7	**Insulin**	InR
Ilp8	–	**Lgr3**
Burs,Pburs	**[DAN]**	**Rk**
Gpa2,Gpb5	**[Cys_knot]**	**Lgr1**
Ptth	**[IL17]**	Tor
Akh	**Adipokin_hormo**	**AkhR**
Mip	–	**SPR**
Crz	–	**CrzR**
Amn	–	–
AstA	**Allostatin**	**AstA-R1,AstA-R2**
AstC,AstCC	–	**AstC-R1,AstC-R2**
Capa	**Periviscerokin**	**CapaR**
Hug	–	**PK2-R1,PK2-R2**
ETH	–	**ETHR**
CCAP	**CCAP**	**CCAP-R**
CCHa1	–	**CCHa1-R**
CCHa2	–	**CCHa2-R**
CNMa	–	**CNMaR**
Dh31	–	**Dh31-R**
Dh44	**CRF**	**Dh44-R1,Dh44-R2**
Eh	**Eclosion**	CG10738
FMRFa	**FARP**	**FMRFaR**
Dsk	**Sulfakinin**	**CCKLR-17D1**
Ms	–	**MsR1,MsR2**
ITP	**Crust_neurohorm**	–
Lk	–	**Lkr**
NPF	**Hormone_3**	**NPFR**
SNPF	–	**SNPF-R**
Rya	–	**Rya-R**
Nplp1	–	Gyc76C
Nplp2	–	–
Nplp3	Retinin_C	–
Nplp4	–	–
Orkokinin	–	–
Pdf	**Pigment_DH**	**Pdfr**
Proc	–	**Proc-R**
SIFa	–	**SIFaR**
Tk	**Lem_TRP**	**TkR99D**
Natalisin	–	**TkR86C**
Trissin	–	**TrissinR**

[a] Domain types—bold: class S; in brackets: Cystine-knot(g).
[b] Receptor types—bold: GPCR; underlined: receptor kinase; double underlined: receptor guanylyl cyclase.
[c] CG43061,CG42870,CG34034,CG42869,CG43618.

Diedel proteins remains to be identified. WntD is proposed to be another cytokine that negatively regulates IMD and TOLL pathways [296].

Spaetzle. The Pfam Spaetzle domain adopts the Cystine-knot fold and is included in the *Drosophila* XC domain Cystine-knot(g). Extracellular Spaetzle domain-containing proteins act as cytokines or neurotrophins. We identified seven such proteins in *Dmel*, including six previously described members—Spz, Spz3, Spz4, Spz5, Spz6, and NT1 (Neurotrophin 1, previously named Spz2) [318], as well as a new and divergent member CG17672. These proteins likely bind Toll-like receptors and play critical roles in embryonic development, neuronal survival, and immune response. Activation of Spz requires proteolytic processing by the trypsin protease Easter, which is part of a cascade of trypsin cleavage events.

SVWC. The SVWC (single-domain von Willebrand factor type C) domain is remotely related to the VWC domain. SVWC is exclusively found in single-domain extracellular proteins in contrast to VWC that is often found in multi-domain proteins. SVWC has two fewer conserved cysteines than VWC and appears to be invertebrate-specific [319]. SVWC-containing genes are expanded in the *Dmel* genome with 14 copies. Some of these genes have more restricted tissue expression patterns than others. For example, *Vago* and *CG2444* are mainly expressed in fat body, and *CG34460* is mainly expressed in testis. Most of SVWC-containing genes have not been experimentally studied. The Vago protein contributes to innate immune response by controlling viral load in the fat body after infection with *Drosophila* C virus [320]. The Vago ortholog in the mosquito *Culex* also stimulates antiviral response to West Nile virus infection through the activation of the JAK/STAT pathway [321]. Vago could thus be a cytokine functionally similar to vertebrate interferons [321]. Whether the SVWC domain of Vago

acts through direct binding of immune receptors remains to be studied.

Gbp1. Gbp1 (Growth-blocking peptide 1) was identified as a potent cytokine involved in immune response induced in infectious and noninfectious conditions [322]. Gbp1 has been found in a number of insect orders and is characterized by a sequence segment (C-x(2)-G-x(4,6)-G-x(1,2)-C-[KR]) with two conserved cysteines [323]. The sequence signature is also present in several other secreted proteins, such as Gbp2 (gene neighbor of Gbp1) and Gbp3 [322,324]. The segment exhibits weak sequence similarity to the C-terminal subdomain of the EGF module [323], and Gbp1 has been shown to interact with and signal through Egfr [325]. A recent study also identified the GPCR Mthl10 as another receptor for Gbp1 [326]. While *Dmel* Gbp proteins have not been incorporated in the Pfam database, HHsearch results (using *Dmel* Gbp1 as query) suggest weak similarity to the Pfam domain Secapin (HHsearch probability score: 95.7) found in honey bee secreted peptides, the Pfam domain FOG_N (HHsearch probability score: 92.6) in the signaling molecule Fog, and the Pfam domain GBP_PSP (HHsearch probability score: 50.3) [constructed from GBP peptides from *Pseudaletia separata*, paralytic peptides from *Manduca sexta, Heliothis virescens*, and *Spodoptera exigua* and plasmatocyte-spreading peptide (PSP1)]. Several EGF domains such as Tme5_EGF_like and EB were also among the weak hits.

Dmel peptide hormones and neuropeptides. A number of genes in *Dmel* encode neuropeptides and peptide hormones [327]. Known Pfam domains in them are classified as class S XC domains. However, not all of these peptide products have been included in the Pfam database. Some of them bear similar sequence motifs and could be evolutionarily related (Fig. 7). These *Dmel* peptide hormones and neuropeptides, with or without class S XC domains (Table 2), are described below.

Insulin-like peptides. Eight insulin-like peptide genes (*Ilp1–8*) are present in the *Dmel* genome. *Ilp1–7* encode typical insulin-like precursor proteins [328] with the Pfam Insulin domain detectable by HMMER. Ilp8 maintains the cysteine patterns of insulin proteins [329]. However, the diversity of its sequence prevents its Insulin domain from being detected by HMMER or HHsearch. Typical insulin-like peptides signal through the insulin-like receptor (InR, an RTK) and play crucial roles in metabolism, growth, development, reproduction, aging, and stress response [328]. The divergent Ilp8 signals through Lgr3 (a GPCR with LRR(g) and Ldl_recept_a domains) in coordination of growth and developmental timing [330].

Cystine-knot hormones—Burs/Pburs, Gpa2/Gpb5, and Ptth. Burs (Bursicon) and Pburs (Partner of Bursicon) are two evolutionarily related hormones with the Pfam DAN domain in the Cystine-knot clan. They form a complex that functions in wing tanning by binding the GPCR Rk (Rickets) with extracellular LRR repeats [331].

The Pfam Cys_knot domain in the Cystine-knot clan are found in Gpa2 (glycoprotein hormone alpha 2) and Gpb5 (glycoprotein hormone beta 5). They form a hormone complex that activates the GPCR Lgr1 [332, 333]. Divergent copies of Cys_knot domains were also identified in three multi-domain proteins Hml, Ccn, and Sli, where they might not function as hormones.

Ptth regulates the production of ecdysone, a steroid hormone that stimulates molting and metamorphosis [282]. Ptth possesses the Pfam IL17 domain, which also belongs to the Cystine-knot clan of signaling domains.

Akh, Mip, and Crz. Akh (Adiptokinetic hormone) is secreted by the corpora cardiaca and regulates carbohydrate and lipid homeostasis [334]. Five myoinhibitory peptides (Mip-1–5) [335] are derived from the protein precursor encoded by the *Mip* gene (also called *AstB*) and are characterized by two conserved tryptophans (Fig. 7). Although Mip peptides are not currently in the Pfam database, weak HHsearch hits to Akh (Pfam: Adipokin_hormo) and mammalian thyrotropin-releasing hormone (Pfam: TRH) [336] suggest possible homology. Mip and the Akh peptides share the φxxxWa|KR motif in the precursor sequence (Fig. 7), where "a" denotes the amide converted from a glycine after the two C-terminal positively charged residues are removed by a carboxypeptidase. *Crz* (*Corazonin*) encodes a cardioactive peptide hormone that is not included in the Pfam database. As previously noticed [337], it also bears weak sequence similarity to Akh (HHsearch probability score 78.8). Crz and Akh peptides share the [YF]SxxW motif (Fig. 7), and both start with a glutamine residue right after the signal peptide cleave site. Similar conservation is also observed in the mammalian GnRH (Gonadotropin-releasing hormone). These hormones as well as their receptors could have predated the bilaterian common ancestors [337].

Amnesiac. The *amn* (*amnesiac*) gene encodes predicted neuropeptides that regulate behaviors such as olfactory memory and sleep [338, 339]. The predicted Amn peptide sequences exhibit weak similarity to vertebrate PACAP (pituitary adenylate cyclase-activating peptide) and GHRH (growth hormone-releasing hormone) [340].

AstA. The *Dmel AstA* (*Allostatin A*) gene encodes a protein precursor that is processed to generate four neuropeptides, all with the C-terminal [YF]xFGLa motif (Fig. 7). This motif corresponds to the Pfam domain of Allatostatin. *AstA* modulates metabolism and feeding of *Dmel* by regulating the signaling of adipokinetic hormone and insulin-like peptides [341].

AstC and AstCC. Dmel AstC (Allostatin C) and its homolog AstCC [342] do not have detectable Pfam

domains. The AstC and AstCC peptides have a conserved disulfide bond (Fig. 7).

Capa, Hugin, and ETH. The *Capa* (*Capability*) gene encodes three neuropeptides of two distinct types [343]. The N-terminal two peptides are PVK-like peptides with the AFPRVa motif (Pfam: Periviscerokin) showing the most similarity to periviscerokinin (PVK) peptides [344]. The C-terminal peptide is a pyrokinin-like peptide with the FxPRLa motif, which has been found in a number of insect hormones such as myostimulatory pyrokinin, pheromone biosynthesis activating neuropeptide (PBAN), diapause-inducing hormone (DH), and melanization and reddish coloration hormone (MRCH) [343]. Another gene *Hugin* also encodes two peptides [345]. One of them (Hug-PK) matches the pyrokinin-like peptide consensus, while the other peptide (Hug-γ) found Pfam PBAN domain with a weak HHsearch probability score (78.6). *Dmel* ecdysis-triggering hormone is encoded by the *ETH* gene. Two peptides ETH-1 and ETH-2 are derived from the gene product, and both can induce premature eclosion. Like Hugin, ETH does not contain detectable Pfam domains by HMMER search. The HHsearch result of ETH showed a weak hit to the Pfam PBAN domain (probability score: 46.8) with the pyrokinin-like peptide motif (FxPRLa). All peptides derived from *Capa*, *Hugin*, and *ETH* share the conserved PR[LI]a motif at the C-terminal ends (Fig. 7), suggesting a common evolutionary origin.

CCAP. CCAP (Crustacean cardioactive peptide) is conserved across the arthropod lineage. CCAP is involved in control of ecdysis behavior and regulation of molting. In addition, it has a variety of functions in different arthropod organisms [343]. CCAP precursor is included in the Pfam database as the CCAP domain. It is processed to a short peptide (PFCNAFTGCa) with two conserved cysteines forming a disulfide bond [346].

CCHamide. CCHamide (named after two conserved cysteines and a conserved histidine) peptides also possess a disulfide bond. Two *Dmel* genes (*CCHa1* and *CCHa2*) encode CCHamide peptides with two conserved cysteines and two conserved histidines (Fig. 7). CCHa2 peptide was recently shown to be an orexigenic brain-gut peptide in *Dmel* affecting feeding and development [347].

CNMa. A novel neuropeptide gene, *CNMa*, was recently discovered [348]. Its two isoforms encode two slightly different peptides (the PD isoform has two additional N-terminal residues compared to the PB isoform; see Fig. 7). Two conserved cysteines are present in these peptides and other insect homologs [348].

Dh31 and Dh44. Two diuretic peptide hormones Dh31 and Dh44 stimulate fluid secretion in Malpighian

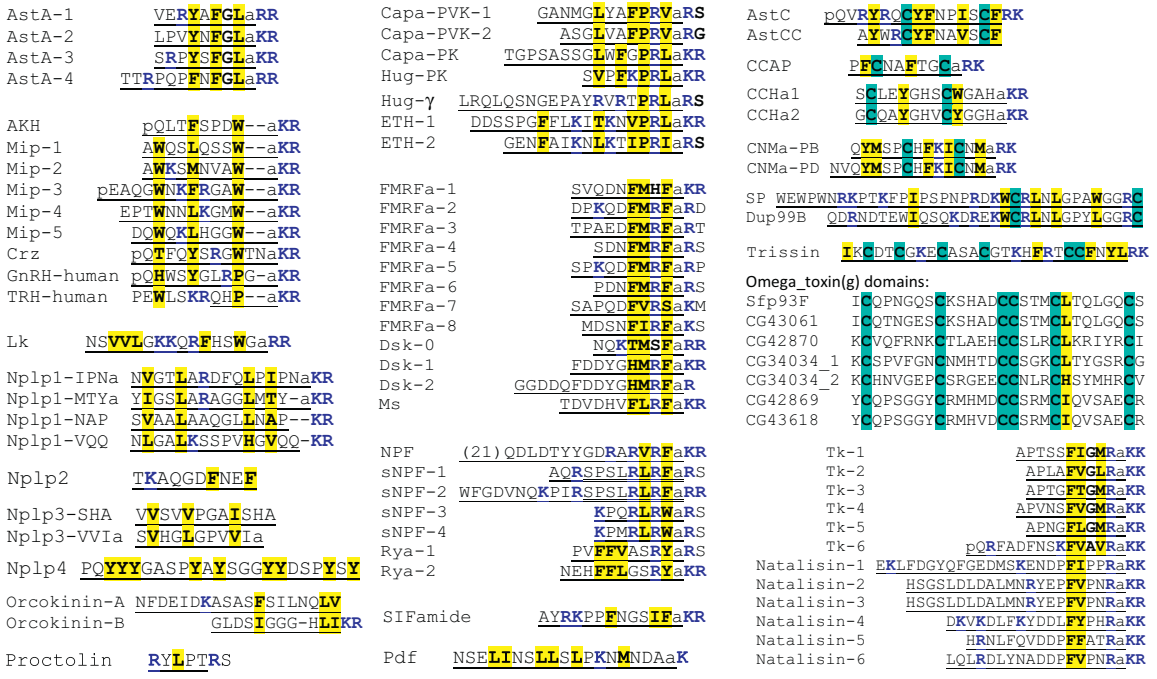


Fig. 7. Sequence patterns in select *Dmel* peptide hormones and neuropeptides. The mature peptides are shown as underlined sequence regions. N-terminal glutamine or glutamate cyclization and C-terminal glycine amidation are shown as small letters "p" and "a", respectively. Positive charged residues in peptides or after the cleavage sites are shown in blue and bold letters. Conserved hydrophobic positions are in yellow background. Peptides with cysteines are in the upper right region, with cysteines shown in cyan background. The alignment of Omega_toxin(g) domains are also shown.

tubules of *Dmel*. Dh44 is homologous to vertebrate corticotropin-releasing factor (CRF) and contains the Pfam CRF domain. While Dh31 has no detectable Pfam domains, it exhibits some similarity to vertebrate calcitonin [349]. A C-terminal GRRRR motif of Pfam Calcitonin domain (Calc_CGRP_IAPP) is indeed found with a weak probability score (49.7) by HHsearch.

Eclosion. The Eclosion hormone (Pfam domain: Eclosion) functioning in the ecdysis cascade is encoded by the *Eh* gene [350]. It is a small secreted protein (53 aa) with six conserved cysteines. The receptor for Eclosion is a membrane-bound guanylyl cyclase (CG10738) [351], unlike GPCRs for most other peptide hormones and neuropeptides (Table 2).

FMRFa, Drosulfakinin, and Myosuppressin. The *Dmel FMRFa* gene encodes eight FMRFamide peptides of the FARP family (Pfam: FARP, FMRFamide related peptide family). These peptides are homologous to those encoded by the *Dmel* sulfakinin gene *Dsk* (*Drosulfakinin*) (Pfam: Sulfakinin) with the HMRFa motif and those encoded by the *Ms* (*Myosuppressin*) gene with the FLRFa motif (Fig. 7). Ms has a weak HHsearch hit to Pfam ELH (egg-laying hormone) domain that contains a similar motif.

ITP. Dmel ITP (*Ion transport peptide*) gene has three splice variants and encodes ITP, ITPL1, and ITPL2 peptides with more than 70 residues [352]. ITP is homologous to crustacean hyperglycemic hormone (CHH) and MOLT-inhibiting hormone (MIH). They all belong to the Pfam family Crust_neurohorm (crustacean neurohormone). CHH and MIH structures consist of mainly α-helices with three conserved disulfide bonds [353, 354]. *ITP* regulates ionic and fluid homeostasis by stimulating chloride transport and inhibiting acid release in Malpighian tubules [352]. *ITP* is also expressed in some clock neurons and could play a role in clock output pathways [355].

Leucokinin. Dmel Lk (Leucokinin) peptide (NSVVLGKKQRFHSWGa), encoded by the *Lk* gene, regulates food intake and water balance [356]. HHsearch using the precursor sequence of Lk found a weak hit (probability score: 65.5) to the Pfam family Kinin, which has only eight positions including the FxSWG motif.

NPF, sNPF, and Rya. NPF (Neuropeptide F), encoded by the *Dmel NPF* gene, is a relatively long peptide with 36-amino-acid residues. It has a C-terminal motif of RVRFa and bears similarity to the vertebrate NPY (Neuropeptide Y) family peptides with the RxRYa motif (Pfam family: Hormone_3). A related set of shorter peptides are encoded by the *sNPF* (*short neuropeptide F*) gene with the RLR[FW]a motif (Fig. 7). Recently, two RYamide peptides were discovered sharing the FFxxxRYamide motif [357]. In *Dmel*, they are encoded by the *Rya* gene, and their receptor is Rya-R [358]. These two peptides have the sequence motif of FFxxxRYamide, which is also similar to vertebrate NPY peptides. It was proposed that NPF, sNPF, and Rya peptides are evolutionarily related [357], as they share the conserved arginine and an aromatic residue before the C-terminal amide (Fig. 7).

Nplp1–4. Four neuropeptide-like protein precursor genes (*Nplp1–4*) were found to encode a diverse set of peptides (Fig. 7), most of which had unclear function at the time of their discovery [359, 360]. One Nplp1 peptide (Nplp1-VQQ) partners with the receptor guanylate cyclase Gyc76C to regulate immune response in stress conditions [361]. *Nplp2* gene has two isoforms, one of which is generated by stop codon readthrough. *Nplp3* gene product has two peptides (VVSVVPGAISHA and SVHGLGPVVIa) derived from a Retinin_C domain. Nplp4 peptide is tyrosine-rich (PQYYYGASPYAYSGGYYDSPYSY). Receptors for peptides encoded by *Nplp2*, *Nplp3*, and *Nplp4* are yet to be discovered.

Orcokinin. The *Orcokinin* gene has two splice variants that encode two different peptides (Orcokinin-A: NFDEIDKASASFSILNQLV and Orcokinin-B: GLDSIGGGHLI) [362]. Orcokinin-A is found in central nervous system, while Orcokinin-B mainly occurs in the midgut enteroendocrine cells [362], suggesting that they have different functions. Neither has detectable Pfam domains.

Pdf. Dmel Pdf (*Pigment-dispersing factor*) gene encodes a small peptide (NSELINSLLSLPKNMNDAa) that belongs to the family of pigment-dispersing hormones (Pfam: Pigment_DH). The Pdf peptide plays an important role as a signaling molecule in circadian behavior control [363].

Proctolin. Another neuropeptide that has not been incorporated into the Pfam database is Proctolin, the first sequenced insect neuropeptide. It is encoded by the *Proc* gene and processed from a precursor of 140 amino acids. However, the mature neuropeptide of Proctolin is only five residues long (RYLPT) after signal peptide removal and further proteolytic processing. Proctolin peptide has a plethora of myostimulatory effects and also functions as a cotransmitter in glutamatergic motoneurons [343].

SIFa. The myotropic peptide SIFamide was first discovered in the fleshfly *Neobellieria bullata* [364]. It has since been identified in various insects. The *Dmel* SIFamide (AYRKPPFNGSIFa) is encoded by the *SIFa* gene. It signals through the GPCR SIFaR and promotes sleep in *Dmel* [365].

Tachykinin and Natalisin. Six tachykinin-related peptides (Tk-1–6) are encoded by the *Dmel Tk* gene. They all have the Fφ[GA]φRa motif (Fig. 7), which is incorporated in the Pfam database as the Lem_TRP domain. *Tk* is expressed in males to control their higher level of aggressive behavior compared to females [366]. *Natalisin* is recently identified to encode several neuropeptides with a sequence motif (FφPxRa) similar to that of tachykinin peptides [367] (Fig. 7). Natalisin peptides affect mating behaviors of both males and females and regulate sexual activity and fecundity in insects [367]. Natalisin and tachykinin peptides bind

two closely related GPCRs—TkR86C and TkR99D, respectively.

Trissin. Trissin encodes a recently identified neuropeptide (IKCDTCGKECASACGTKHFRTCCFNYL), which has 27 amino acids and three disulfide bonds [368]. Its receptor is the GPCR protein TrissinR. The function of the Trissin peptide is yet to be determined.

Class S domains in (putative) signaling molecules passed from male to female. Sex_peptide. Dmel Sex Peptide (encoded by *SP*) is secreted in the male seminal fluid and induces responses such as sexual receptivity reduction and egg production upon transfer to mated female [369]. Sex Peptide is enriched with tryptophan residues and contains hydroxyproline (Hyp) residues and a disulfide bond [370]. The Pfam Sex_peptide domain is found in both Sex Peptide and its paralog Dup99B (Fig. 7).

Omega_toxin(g). Six predicted secreted proteins contain one (Sfp93F, CG42869, CG42870, CG43061, and CG43618) or two (CG34034) small cysteine-rich domains (Fig. 7) related to domains from the Pfam Omega_toxin clan such as Toxin_19, Toxin_21, Toxin_18, Omega-toxin, and Conotoxin. These domains with six conserved cysteines are predicted to have a knottin-like fold with 1–4, 2–5, and 3–6 disulfide pairing of cysteines [371]. Omega_toxin(g)-containing genes are expressed in the male accessory gland, and their products are secreted seminal fluid proteins. The six genes are located in two gene clusters on chromosome 3R (*Sfp93F*/*CG42869*/*CG42870*/*CG34034* and *CG43061*/*CG43618*). Omega_toxin-fold domains have been found in venom proteins of various invertebrates such as spider insectidal peptides (Pfam domains Toxin_21 and Omega-toxin) and conotoxins (Pfam domains Conotoxin and Toxin_18), where they mainly act as ion channel blockers. The male accessory gland products are known to confer sperm competition and exert a variety of effects on female behavior, reproductivity, and life span. The cost of mating with exposure to these seminal fluid products increases female death rate [372]. The functions of these *Dmel* genes with Omega_toxin(g) domains are unknown. Given the toxic effects of Omega_toxin(g) domains in venoms of other invertebrates, they could contribute to sperm competition or have negative effects on female fitness after mating.

MAGSP and Acp26Ab. Two *Dmel* proteins Acp26aa and Acp26ab are encoded by neighboring genes [373]. They are secreted in male accessory gland and passed to females during copulation. They also enter the circulation system (hemolymph) after being transferred inside the female body. They promote egg production and modulate hatch time, like the *Dmel* Sex Peptide. Acp26aa bears weak similarity to the egg-laying hormone of Aplysia [373]. Whether they act as signaling molecules remains to be studied. The XC domains of Acp26aa and Acp26ab are MAGSP and Acp26Ab in the Pfam database, respectively.

Class B XC domains likely involved in binding non-protein molecules and groups

A total of 51 *Drosophila* XC domains are included in this functional class based on literature analysis that indicates their ability to bind non-protein molecules and groups. They bind various ligandsincluding carbohydrates and lipids (Table 3). LIG(g), Chitin_bind_4, and CBM_14_19(g) are the most populated class B domains with more than 100 genes associated with each of them. LIG(g) domain with the IG-fold has been described above.

Class B chitin-binding domains. The arthropod exoskeleton has a multi-layered structure (epicuticle, procuticle, epithelium, and BM). The cuticle (epicuticle + procuticle), the non-cellular material produced by epithelium cells, is the main part of the exoskeleton and is a special type of ECM due to its composition [374]. The major component of cuticle is chitin, a long-chain polymer of the sugar derivative N-acetylglucosamine. A number of residing proteins are integral structural components of cuticle as well. Most of these proteins belong to a limited number of groups based on their evolutionary origins [375,376]. Three abundant class B domains (Chitin_bind_4, CBM_14_19(g), and DUF243) are in cuticle proteins that bind chitin.

Chitin_bind_4. The largest cuticle protein group, CPR, has the R&R consensus sequence initially recognized and defined by Rebers and Riddiford in 1988 [377]. The consensus motif region corresponds to the Pfam Chitin_bind_4 domain. In *Dmel,* this domain is second most populated class B domains found in proteins encoded by 118 genes. The majorities of these proteins are named cuticle proteins (such as Cpr5C) or larval cuticle proteins (such as Lcp1). Crys (Crystallin) is a component of the laminated corneal lens in the eye [378]. It also contributes to the formation of the peritrophic matrix, a chitinous layer lining the midgut [379].

CBM_14_19(g). The second major superfamily of chitin-binding domains [CBM_14_19(g)] corresponds to Pfam clan CBM_14_19, which has two Pfam families CBM_14 and CBM_19. A total of 110 *Dmel* genes were found to encode secreted proteins with these domains, the majority of which have better HMMER or HHsearch scores to the CBM_14 domain than to the CBM_19 domain. These small domains (about 60- to 70-amino-acid residues) are mainly made up of β-strands and possess six conserved cysteines [380] (Fig. 8a). The proteins containing them include members of the CPAP (Cuticular Proteins Analogous to Peritrophins)

Table 3. Class B XC domains that likely bind non-protein molecules and groups

	XC domain	No. gene	Ligand
1	LIG(g)	119	Lipid, pheromone
2	Chitin_bind_4	118	Carbohydrate (chitin)
3	CBM_14_19(g)	110	Carbohydrate (chitin)
4	PBP(g)	77	Neurotransmitter, iron
5	PBP_GOBP	51	Pheromone, odorant
6	Lectin_C(g)	43	Carbohydrate
7	Aha1_BPI(g)	31	Juvernile hormone
8	Periplas_BP(g)	28	Glutamate,small molecules
9	DUF243	26	Carbohydrate (chitin)
10	Neur_chan_LBD	23	Neurotransmitters
11	Laminin_G(g)	23	Carbohydrate (heparan,HSPG)
12	Fibrinogen_C(g)	18	Carbohydrate (heparan,HSPG)
13	CD36	14	Lipid
14	Calycin(g)	11	Lipid
15	Knottin_1(g)	9	Lipid (pathogen membrane)
16	Attacin_C(g)	8	Lipid (pathogen membrane)
17	Ins_allergen_rp	7	Lipid
18	F5_F8_type_C	7	Carbohydrate (heparan, HSPG)
19	DOMON	6	Carbohydrate,heme
20	Cache(g)	6	Small molecules
21	CBM39	6	Lipid (pathogen membrane)
22	Cecropin	5	Lipid (pathogen membrane)
23	OS-D	4	Pheromone,odorant
24	Attacin_N	4	Lipid (pathogen membrane)
25	VEGF_C(g)	3	Carbohydrate (heparan,HSPG)
26	Vitellogenin_N	3	Lipid
27	PBP	3	Lipid
28	DUF3421	3	Carbohydrate
29	DUF1943	3	Lipid
30	DUF1081	3	Lipid
31	MNNL	2	Lipid
32	Ferritin	2	Iron
33	Gal_Lectin	2	Carbohydrate
34	PTN_MK_C	2	Carbohydrate (heparan,HSPG)
35	APP_N	1	Carbohydrate (heparan,HSPG)
36	CutA1	1	Copper
37	Avidin	1	Biotin
38	APP_Cu_bd	1	Copper
39	APP_E2	1	Carbohydrate (heparan,HSPG)
40	Cobalamin_bind	1	Cobalamin
41	PLAT	1	Lipid
42	SLBB	1	Cobalamin
43	Gelsolin	1	LPS
44	Metallothio_Euk	1	Metal
45	ApoO	1	Lipid
46	COMP	1	Hydrophobic compounds
47	WSC	1	Carbohydrate
48	Ependymin	1	Lipid
49	CAP18_C	1	LPS
50	MACPF	1	Lipid
51	LysM	1	Peptidoglycan

family of cuticle proteins [376, 381]. They are also present in two other arthropod chitin-containing apical ECM structures—the peritrophic matrix lining the digestive tract and the apical matrix of trachea. Chitin-binding activity was experimentally shown for peritrophin proteins with this domain from the larvae of *Lucilia cuprina* [382] and the malaria vector *Anopheles gambiae* [383].

The obstructor multigene family [384] of *Dmel* with CBM_14 domains has 10 members (*obst-A,B,E,F,G,J, H,I*, *Gasp*, and *Peritrophin-A*). Most of these genes encode proteins with three CBM_14 domains (the one exception is Obst-J with four such domains). Some of these genes such as *obst-A*, *Gasp* [385], and *Peritrophin-A* are highly expressed in larval trachea. *Gasp* and *obst-A* control airway tube diameter and integrity during larval development [386]. Obst-A organizes ECM assembly at the apical cell surface [387]. CBM_14 domain is found in seven chitinases (Cht3–8 and Cht12) containing the Pfam domain Glyco_hydro_18 with chitinase activity. It also co-occurs with two other XC enzyme domains [Polysacc_deac_1 and Trypsin(g)] and the Ldl_recept_a domain in a few secreted proteins. Nine CBM_14-containing genes are named mucins (*Muc11A*, *Muc18B*, *Muc26B*, *Muc68D*, *Muc68E*, and *Muc96D*) or mucin-related (*Mur18B*, *Mur89F*, and *Mur2B*). Their proteins have mucin-like low complexity regions enriched with proline, threonine, and serine residues [388]. Some of them (e.g., *Muc18B*, *Muc26B*, *Muc68D*, *Muc68E*, and *Muc96D*) are highly expressed in the digestive system suggesting their roles as peritrophins [388], while others have enriched expressions in other tissues such as Malpighian tubules (*Muc11A* and *Mur18B*) and ovary (*Muc2B*) (data from FlyAtlas [389]). Many CBM_14-containing genes (e.g., about two thirds of them with names beginning with CG) have not been experimentally characterized. They could contribute to the integrity of apical ECM via the predicted chitin-binding activity, or could evolve new functions such as antimicrobial activity observed in the Tachycitin protein from horseshoe crab [380].

DUF243. The Tweedle family cuticle proteins [390] are encoded by 26 genes in *Dmel.* Mutations of *TwdlD* caused *Dmel* body shape change at the larval and pupal stages [390]. Tweedle proteins are characterized by the Pfam DUF243 domain. This domain was predicted to consist of mainly β-strands and could bind chitin [390]. Indeed, a Tweedle family protein from *Bombyx mori* was experimentally shown to bind chitin [391].

Other class B domains that bind carbohydrates. Besides chitin-binding domains, a number of carbohydrate-binding domains are found in *Dmel* CSS proteins (Table 3). One of the popular XC domains is Lectin_C (g) (C-type lectin) domain, which is found in the products of 43 genes. While typical C-type lectins rely on calcium for carbohydrate binding (Fig. 8B), some of them have evolved to bind other ligands such as lipids and proteins [392]. They are important for innate immunity and cell–cell interactions. *Dmel* C-type lectins were found to interact with bacteria [393] and enhance cell encapsulation in immune response [394]. For the majority of the *Dmel* proteins with the Lectin_C(g) domain, no other known Pfam domains were detected. Some of them have an N-terminal region predicted to be mostly α-helical. One protein (Lectin-24A) has an N-

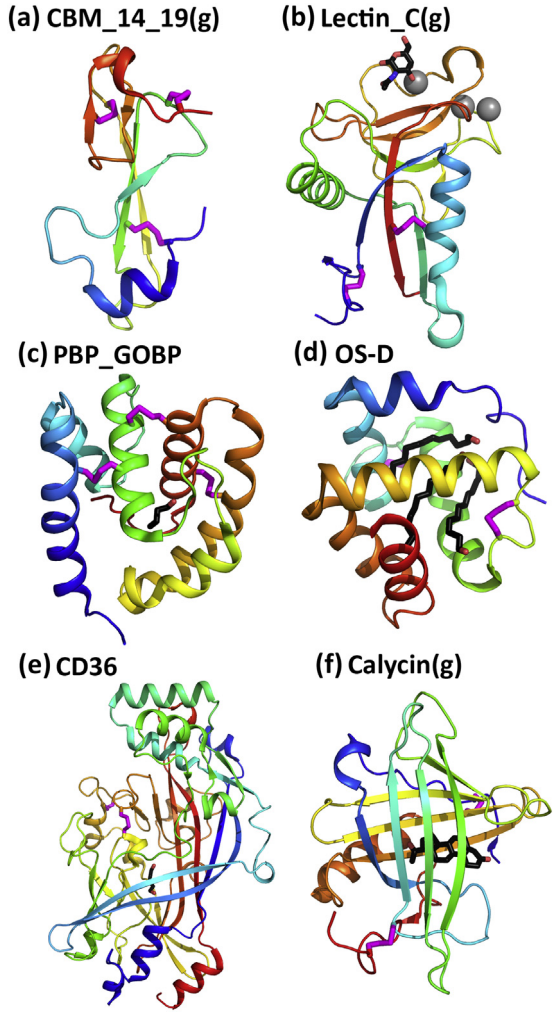


Fig. 8. Structures of representative class B XC domains binding non-protein molecules and groups. XC domains are in rainbow from N-terminus (blue) to C-terminus (red), with disulfides in magenta sticks and ligands in black sticks. Carbohydrate-binding domain folds are illustrated for (a) CBM_14_19(g) (4z4a, A33–A104) and (b) Lectin_C(g) (1wmz, A1–A140, with ligand *N*-acetyl-D-galactosamine). Odorant and pheromone-binding domain folds are illustrated for (c) PBP_GOBP (1ooh, A1–A124, with ligand butanol) and (d) OS-D (1n8v, A3–A103, with ligand bromo-dodecanol). Lipid-binding domain folds are illustrated for (E) CD34 (4f7b, A37–A429, with ligand PEG) and (F) Calycin(g) (2hzq, A3–A168, apolipoprotein D (ApoD) in complex with progesterone).

terminal Lectin_N domain (predicted coiled coils). Lectin_C(g) domain co-occurs with F5_F8_type_C (another carbohydrate-binding domain) and Sushi in three type I TM proteins (Fw, CG9095, and Uif). It also co-occurs with Ig(g) and fn3(g) domain in the GPI-anchored cell adhesion molecule Contactin (encoded by the *Cont* gene) [395]. The *bark* (*bark beetle*) gene encodes a scavenger receptor with the Pfam UL45 domain (a divergent C-type lectin domain), Beta_helix repeats, SRCR(g) domains, and a CUB(g) domain. It plays an essential role in septate junction maturation [396].

Several XC domains are capable of binding ECM carbohydrate molecules or groups such as heparin and HSPGs. Among them, Laminin_G(g) domain, co-occurring with a number of XC domains (Fig. 3b), is detected in a variety of CSS proteins (encoded by 23 genes) including ECM proteins (e.g., laminins, collagens, and neurexins) and some cadherins. Other heparin-binding class B domains include Fibrinogen_C(g), F5_F8_type_C, VEGF_C(g), PTN_MK_C, APP_N, and APP_E2.

Class B XC domains in receptors and ion channels. The Pfam PBP clan includes more than 20 Pfam domains found in CSS proteins from various organisms, such as bacterial periplasmic binding proteins. *Dmel* has 77 genes encoding proteins with domains from this clan, which are included in the XC domain PBP(g). The majority of them (74) are multi-pass TM proteins that function as ligand-gated ion channels. These proteins possess Pfam domains Lig_chan-Glu_bd and SBP_bac_3 of the PBP clan and the TM domain Lig_chan. iGluRs mediate communication between neurons at synapses, and their ligands include neurotransmitters such as glutamate, AMPA (α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid), NMDA (*N*-methyl-D-aspartate), and kainate [397]. A subfamily of these proteins (the Ionotropic Receptors, IRs) are expanded in *Dmel* with more than 40 members. They play important roles as olfactory receptors that detect environmental chemical signals [398]. The other three genes (*Tsf1–3*) encode secreted or GPI-anchored proteins with the Transferrin domain of the Pfam PBP clan. They are iron-binding proteins induced during the immune response. Sequestration of iron is an important mechanism in fighting bacterial infection of *Dmel* [399].

Another Pfam clan containing bacterial periplasmic binding domains is Periplas_BP. *Dmel* has 28 genes encoding proteins with the ANF_receptor domain [400] that belongs to the Periplas_BP clan and is mostly associated with cell surface receptors. Sixteen of them co-occur with the PBP(g) domain in iGluRs. Five of them are present in GPCRs (MGluR, GABA-B-R1, GABA-B-R2, GABA-B-R3, and Mtt). MGluR (metabotropic glutamate receptor) binds the neurotransmitter glutamate, and Mtt (Mangetout) is its close homolog that has lost the binding activity [401]. GABA-B-R1, GABA-B-R2, and GABA-B-R3 are metabotropic GABA (B) receptors [402]. Another six genes encode receptor guanylyl cyclases with the ANF_receptor domain. Two of them bind neuropeptides (the eclosion hormone and NPLP1-VQQ), and the rest are orphan receptors [403].

Class B XC domains involved in odorant and pheromone binding. Chemoreception, the mechanism of

sensing environmental chemical signals, is crucial for the survival of *Dmel*. It has a number of chemosensory organs including two olfactory organs (antenna and maxillary palp), taste pegs, and sensilla at various locations such as labellum, legs, and wings [404]. *Dmel* has two major families of odorant binding and chemosensory proteins, both of which adopt small helical structures [405] (Fig. 8c and d). OBPs (odorant-binding proteins) constitute a large family encoded by more than 50 genes. They possess the Pfam PBP_GOBP domain [406] (Fig. 8c) that binds hydrophobic odorants. All except one are predicted secreted proteins (Obp50b is predicted to be a GPI-anchored protein). They carry the odorants and transport them to odorant receptors in the dendrites olfactory receptor neurons [407]. The chemosensory protein family (CSP) have four members of secreted proteins with the Pfam OS-D domain (Fig. 8d). Two proteins with this domain are also involved in viral response and metamorphosis [408]. Besides OBPs and CSPs, some of the DUF1091 family proteins with the LIG(g) domain could also contribute to male-specific sensing of pheromones, as described above. In addition, the gene *a5* (*antennal protein 5*) with the Pfam PBP domain (not to be confused with the Pfam PBP clan), expressed in a subset of olfactory hairs in antenna, could be an odorant-binding protein.

Class B XC domains that bind lipids. A number of class B XC domains (Table 3) are found to bind lipids, which include a variety of endogenous and exogenous molecules that are fatty acids or their derivatives.

Dmel has 14 genes encoding the functionally diverse CD36 proteins with a cysteine-rich XC domain (Fig. 8e) in between two TM segments [409]. As scavenger receptors, CD36 proteins are involved in recognition and internalization of oxidized lipid particles, apoptotic cells, and certain microbial pathogens, thus contributing to inflammatory responses, innate immunity, and metabolism [410]. Two CD36 genes, *crq* (*croquemort*) and *dsb* (*debris buster*), play important roles in phagosome maturation for dendrite clearance [411]. *Dmel ninaD* is highly expressed in midgut and is responsible for carotenoid intake [412]. *Dmel Snmp1* (*Sensory neuron membrane protein 1*) is involved in pheromone detection on dendrites of specialized neurons in antenna [413].

The Lipocalin family of fatty acid binding proteins (Fig. 8f) belongs to the large Pfam Calycin clan that contains more than 30 Pfam domains [414]. *Dmel* has 11 genes encoding predicted secreted or GPI-anchored proteins with such domains. Among them, GLaz and NLaz are homologs of vertebrate apolipoprotein D. They have been linked to stress resistance and life span [415–417].

Three genes (*Rfabg*, *cv-d*, and *Apoltp*) encode proteins homologous to vertebrate vitellogenin, a lipid-binding egg yolk protein. These proteins contain the Vitellogenin_N and DUF1943 domains for lipid binding. The Vitellogenin_N domain region has a β-barrel domain and a helical repeat domain, while DUF1943 corresponds to a β-meander domain [418]. These proteins also contain other domains such as DUF1081, VWD, and C8. Rfabg and Apoltp are apolipoproteins of the ApoB family. They are lipid carriers in hemolymph and transport lipids between tissues [419]. The gene *cv-d* encodes a vitellogenin-like lipoprotein that plays an important role in BMP signaling [420].

The Pfam ApoO domain, found in vertebrate apolipoprotein O [421], is present in a single *Dmel* gene *CG5903*. The protein encoded by *CG5903* could bind lipids like its vertebrate homologs.

The Ins_allergen_rp domain is found in seven *Dmel* genes. This domain was originally found in a cockroach protein that is a human allergen [422]. Structure of this domain revealed a helical fold that binds lipids and other hydrophobic molecules [423]. The cellular functions of *Dmel* genes with this domain are largely unknown. One of them, *jtb* (*jetboil*), was identified as a thermal nociception gene [424]. Several genes have enriched expression in certain tissues (*CG4409* in the nervous system, *CG3906* and *CG9021* in the digestive system, and *CG13905* and *CG14963* in the Malpighian tubules), suggesting that they are functionally diverse.

Several class B XC domains can interact with lipid components from membrane or cell wall of pathogens. These domains function as pattern recognition receptors (CBM39) [425] and antimicrobial peptides such as Attacin_N, Attacin_C(g), Knottin_1, and Cecropin [426].

Class E XC domains of enzyme homologs

A total of 66 domains were classified as XC enzyme homolog domains found in *Dmel* CSS proteins based on literature analysis (Table 4). The majority of them (48) are hydrolases (EC 3.-.-.-) that act upon peptide bonds (EC 3.4.-.-, 20 domains), ester bonds (EC 3.1.-.-, 18 domains), sugars (EC 3.2.-.-, 6 domains), and amide bonds other than peptide bonds (EC 3.5.-.-, 4 domains). Compared to hydrolases, the other enzyme classes are less represented in *Drosophila* XC domains, with nine oxidoreductases (EC 1.-.-.-), six transferases (EC 2.-.-.-), two lyases (EC 4.-.-.-), and one isomerase (EC 5.-.-.-). No domains with ligase activity (EC 6.-.-.-) were found to be extracellular.

Class E XC domains that are peptidase (EC 3.4.-.-) homologs. Trypsin(g). Trypsin(g) is discovered in the largest number of *Dmel* genes (259) among all XC domains. Trypsin(g)-containing proteins carry out a plethora of functions. The Jonah [427] members of Trypsin(g)-containing genes such as *Jon99Ci* and *Jon44E*, as well as many members from a trypsin gene cluster [428] such as *αTry* and *βTry*, are highly expressed in the digestive system, suggesting their role

Table 4. Class E XC domains of enzyme homologs

	Class E domain	No. gene[a]	EC[b]
1	Trypsin(g)	259(67)	3.4.21
2	Lipase	30(8)	3.1.1
3	Peptidase_M13	28(10)	3.4.24
4	Abhydrolase(g)	25(0)	3.1.1
5	COesterase	23(13)	3.1.1
6	Peptidase_M1	20(5)	3.4.11
7	Peptidase_M14	20(2)	3.4.17
8	Glyco_hydro_18	17(8)	3.2.1
9	Lysozyme(g)	17(8)	3.2.1
10	DUF229	16(3)	3.1
11	Alpha-amylase	14(1)	3.2.1
12	Amidase_2	13(8)	3.5.1
13	Astacin	13(0)	3.4.24
14	Asp	13(0)	3.4.23
15	Alk_phosphatase	13(1)	3.1.3
16	SEA(g)	11(10)	3.4.21
17	Peptidase_C1	10(2)	3.4.22
18	GPS	10(6)	3.4
19	Carb_anhydrase	10(4)	4.2.1
20	Endonuclease_NS	9(1)	3.1
21	Reprolysin(g)	9(1)	3.4.24
22	Metallophos	8(1)	3.1.3 3.1.4
23	An_peroxidase	8(0)	1.11.1
24	PLA2(g)	7(0)	3.1.1
25	Hemocyanin_M	7(7)	~~1.14.18~~
26	Lipo_10	6(1)	1.14.99
27	Peptidase_M2	6(3)	3.4.15
28	Polysacc_deac_1	6(2)	3.1.1
29	A_deaminase	6(1)	3.5.4.4
30	PLC(g)	5(1)	3.1.4
31	Peptidase_S9	5(2)	3.4
32	Peptidase_S10	5(0)	3.4.16
33	GILT	5(4)	1.8
34	Peptidase_M28	4(1)	3.4
35	Peptidase_M19	4(1)	3.4.13
36	CN_hydrolase	4(0)	3.5
37	Cu-oxidase(g)	4(0)	1.1
38	G_glu_transpept	4(0)	2.3.3.2
39	Glyco_hydro_16	3(3)	~~3.2.1~~
40	SGL(g)	3(0)	3.1.1
41	Glyco_hydro_20	3(0)	3.2.1
42	Cu2_monoox(g)	2(0)	1.14.17
43	DM13	3(?)	1.
44	Sod_Cu	3(2)	1.15.1.1
45	Peptidase_S8	3(0)	3.4.21
46	XendoU	2(0)	3.1
47	Peptidase_M10	2(0)	3.4.24
48	Lipase_GDSL	2(0)	3.1.1
49	Lysyl_oxidase	2(0)	1.4.3.13
50	Fam20C	2(0)	2.7.11.1
51	LCAT	2(0)	2.3.1.43
52	Nicastrin	1(1)	~~3.4~~
53	Exo_endo_phos	1(0)	3.1
54	Ribonuclease_T2	1(0)	3.1.27.1
55	NHL	1(0)	4.3.2.5
56	His_Phos_2	1(0)	3.1.3
57	Peptidase_M24	1(0)	3.4
58	Ceramidase_alk	1(0)	3.5.1.23
59	Pro_isomerase	1(0)	5.2.1.8
60	Peptidase_C26	1(0)	3.4.19.9
61	Methyltransf_FA	1(?)	2.1.1
62	AIG2(g)	1(0)	2.3.2.4
63	NDK	1(0)	2.7.4.6
64	PAE	1(0)	3.1.1
65	Trehalase	1(0)	3.2.1.28
66	Sulfatase	1(0)	3.1.6

in breakdown of dietary proteins. Several Trypsin(g) genes, such as *ndl* (*nudel*), *gd* (*gastrulation-defective*), *snk* (*snake*), and *ea* (*easter*), are involved in a protease cascade crucial in the early developmental process [429]. Tequila, a human neurotrypsin ortholog, regulates long term memory [430]. Still, many Trypsin(g)-containing genes do not have known functions. Some of them could have tissue-specific roles as suggested by their expression patterns. For example, *sphinx1*, *sphinx2*, *aqrs*, and *CG17424* have restricted expression in the male accessory gland.

The majority of Trypsin(g) genes are predicted to be secreted proteins and do not contain other domains. A subset of Trypsin(g) domains are associated with the class R CLIP regulatory domain. A small number of genes, such as *gd*, contains an N-terminal domain of unknown function (GD, a class R domain). Several genes possess class P domains such as Ldl_recept_a, CUB(g), and SRCR(g). *Dmel* Ndl, the first member of a protease cascade that leads to the maturation of the Toll-like receptor ligand, is a type II TM protein with Ldl_recept_a and SRCR(g) domains, an active Trypsin (g) domain, and an inactive Trypsin(g) domain (Pfam: DUF1986). Sixty-seven of the 259 Trypsin-containing genes, such as *Aqrs*, *intr*, and *spheroide*, could encode inactive enzymes as suggested by the lacking of one or more catalytic triad residues.

Peptidase_MA clan metalloprotease domains. Six XC domains (Peptidase_M13, Peptidase_M1, Astacin, Reprolysin, Peptidase_M2, and Petidase_M10) are from the Pfam Peptidase_MA clan with the zincin-like fold. These families are characterized by the HEXXH sequence motif, where the glutamate is the catalytic residue and the two histidines bind zinc.

Eighteen of the 28 Peptidase_M13 domains in *Dmel* are predicted to be catalytically active. They include several Neprilysin proteins that are important for cleavage of extracellular signaling proteins [431]. The 10 non-peptidase members of Peptidase_M13 domains include several divergent domains that can only be detected by HHsearch, and not by HMMER. One of them is encoded by the gene *goe* (*gone early*) that has been shown to attenuate Egfr signaling to control the number of the primordial germ cells [432].

Twenty *Dmel* genes encode proteins with Peptidase_M1 domain, five of which are predicted to be catalytically inactive. Little is known about their substrates and cellular functions.

All 13 Astacin-containing genes are predicted to encode catalytically active enzymes based on conservation of zinc-binding and catalytic residues. Eleven of them encode single-domain proteins with differing expression patterns. Some of them (e.g., *CG7631*

Notes to Table 4:
[a] The number of inactive enzymes shown in parentheses.
[b] EC numbers crossed-out if none of the XC domains in *Dmel* CSS proteins are catalytically active.

and *CG15255*) could encode digestive enzymes with high expression levels in midgut, while *Semp1* has restricted expression and functions in male accessory gland. Two Astacin-containing genes (*tld* and *tok*) possess similar modular domain architectures with several CUB(g) and EGF(g) domains. Their products are important for cleavage of Dpp antagonist Sog and the prodomains of TGF-β ligands [433].

Nine Reprilysin-containing genes encode multidomain proteins of the ADAM (A Disintegrin and Metalloprotease) family. Among them, only *mmd* appears to encode a catalytically inactive enzyme domain.

Peptidase_M2 domain has petpidyl dipeptidase activity. Three of the six genes with Peptidase_M2 domain encode active enzymes, including *Ance* (*Angiotensin converting enzyme*) that is important for *Dmel* development and reproduction [434].

Petidase_M10 domain is found in two matrix metalloproteases (Mmp1 and Mmp2) that cleave proteins in the ECM. They both have the PG_binding_1 domain and four Hemopexin domains (adopting a four-bladed beta-propeller fold).

Other metalloprotease homolog domains. Five other XC domains are metallopeptidase homologs—Peptidase_M14, Peptidase_M28, and Nicastrin from the Peptidase_MH clan, Peptidase_M19, and Peptidase_M24.

The majority of Peptidase_M14 members (18 of 20) are predicted to be catalytically active based on preservation of zinc-binding and catalytic residues. Some of them (e.g., CG17633 and CG12374) could be the orthologs of mammalian pancreatic carboxypeptidases based on sequence similarity and their enriched expression in digestive systems. The *svr* (*silver*) gene encodes the orthologous protein of mammalian carboxypeptidase D [435], which primarily resides in the trans-Golgi network and contributes to the maturation of neuropeptide and hormones.

Four Peptidase_M28 members include QC and IsoQC, which are glutaminyl-peptide cyclotransferases (also called glutaminyl cyclases, EC 2.3.2.5) [436]. These enzymes generate N-terminal pyroglutamic acid on peptide or protein substrates to help stabilize them against N-terminal degradation.

The type I TM protein Nicastrin is a subunit of the γ-secretase complex that also includes the aspartic TM protease Presenilin, Aph-1, and Pen-2 [437]. γ-Secretase catalyzes intramembrane proteolysis of TM proteins such as the amyloid precursor and Notch. The catalytically inactive metallopeptidase domain of Nicastrin could play a role in protein–protein interaction. The vertebrate protein Nicalin, a paralog of Nicastrin, forms a protein complex with Nomo and Tmem147 (a paralog of Aph-1). Based on sequence similarity searches, orthologs of Nicalin, Nomo, and Tmem147 were also discovered in *Dmel*, encoded by *CG4972*, *CG1371*, and *CG8675*, respectively. Unlike the γ-secretase complex that can locate on the cell surface, the Nicalin-Nomo-Tmem147 complex mainly resides in the ER [438].

The gene *CG6225* encodes a protein with Peptidase_M24 domain. It is likely an ortholog of human membrane-bound Xaa-Pro aminopeptidase 2.

Autoproteolysis domains—SEA(g) and GPS. Two XC domains, SEA(g) and GPS, possess autoproteolytic activity.

The SEA(g) domain is present in a number of proteins such as mucins, Notch, dystroglycan, α-sarcoglycan, and receptor phosphatase IA-2 [99]. Catalytically active SEA(g) domains use the serine residue in the GSφφφ motif (φ: a hydrophobic residue) for autoproteolysis [99]. All but one SEA(g) domains in *Dmel* have lost the autocleavage activity, such as Notch and Scgα. Notch is still proteolytically processed by furin-like proteases in a loop region at the same location of the autocleavage motif in other SEA(g) domains. Notch SEA(g) domain is represented as two Pfam domains (NOD and NODP) separated at the proteolysis site.

The Pfam GPS domain corresponds to a segment in a globular domain (GAIN) with mainly β-strands [439]. GPS domain was found in GPCRs and PKD (polycystic kidney disease) proteins. The autocleavage motif (Hφ[TS]) in the GPS domain is preserved in the protein products of four out of the 10 GPS-containing *Dmel* genes.

Other XC peptidase domains. Other XC peptidase domains include one aspartyl protease domain (Asp), two cysteine protease domains (Peptidase_C1 and Peptidase_C26), and three serine protease domains (Peptidase_S9 and Peptidase_S10 with the α/β hydrolase fold, and Peptidase_S8).

One of the 13 Asp-containing genes is *Pgcl* (*Pepsinogen C-like*). It exhibits high expression levels in the digestive system, suggesting a role in nutrient breakdown. It is also functionally important in the development process that controls planar cell polarity [440]. The CathD protein, an ortholog of mammalian cathepsin D, is located in both extracellular space and lysosome. Another Asp-containing gene (*CG31928*) encodes a minor protein located in chorion [266].

Ten genes encode proteins with the Peptidase_C1 domain (papain-like peptidase). Like their mammalian orthologs, some of them could locate at the lysosome or get secreted. Two of the Peptidase_C1-containing proteins are catalytically inactive due to mutations of active site residues. One of them is encoded by the *Swim* (*Secreted Wg-interacting molecule*) gene, which unlike other Peptidase_C1-containing genes, also possesses the Somatomedin_B domain.

Class E XC domains that are esterase (EC 3.1.-.-) homologs. Class E XC domains with α/β hydrolase fold. Four XC enzyme domains (Lipase, Abhydrolase (g), COesterase, and PAE) adopt α/β hydrolase fold and perform the hydrolysis of carboxylic ester bonds

(EC number: 3.1.1). Lipase, Abhydrolase(g), COesterase are among the most abundant XC enzyme domains found in 30, 25, and 23 genes, respectively. Catalytically active members perform hydrolysis on various substrates such as triglyceride, acetylcholine, and juvenile hormone. These functionally diverse domains are also present in some intracellular proteins. Lipase is the second most abundant class E XC domain in *Dmel*. Eight out of the 30 Lipase-containing genes could encode catalytically inactive proteins due to alterations in the catalytic triad residues. Among them are three major yolk proteins (Yp1-3) that have functions of binding lipids [441, 442]. For Abhydrolase(g) domain, all 25 members possess the catalytic triad residues, suggesting they are all active enzymes. About half of the COesterase-containing genes (13 out of 23) are predicted to encode catalytically inactive proteins. They include four neuroligin genes (*Nlg1–4*), which encode type I TM proteins that function as cell adhesion molecules in central nervous system. Neuroligins interact with neurexins in the connection between neurons and play crucial roles in synaptic transmission [443]. Inactive COesterase domains are also present in several other known cell adhesion molecules, such as Neurotactin (a type II TM protein), Gliotactin (a type I TM protein), and Glutactin (a secreted protein). PAE domain is present in the product of a single gene *Notum* that cleaves GPI anchors. Other XC domains with the α/β hydrolase fold include two serine peptidase domains (Peptidase S9 and Peptidase S10), as well as the LCAT (Lecithin:cholesterol acyltransferase) domain.

PLA2(g), Polysacc_deac_1, SGL(g), and Lipase_GDSL. Four other XC domains (PLA2(g), Polysacc_deac_1, SGL(g), and Lipase_GDSL) catalyze the hydrolysis of carboxylic ester bonds (EC number: 3.1.1).

The XC domain PLA2(g) contains three Pfam domains (Phospholip_A2_1, Phospholip_A2_2, and PLA2G12) with the phospholipase A2 activity that releases fatty acids from the second carbon group of glycerol. Seven PLA2(g)-containing genes encode secreted proteins predicted to be catalytically active based on conservation of active site residues (one histidine and two acidic residues).

Dmel Polysacc_deac_1-containing genes encode proteins that are chitin deacetylases [444] or the catalytically inactive homologs. These proteins adopt the TIM-barrel fold, and catalytically active members are metalloenzymes that bind one zinc using two histidines and one aspartic acid [445]. Two of the six Polysacc_deac_1-containing genes (*verm* (*vermiform*) and *Cda4*) are predicted to encode catalytically inactive proteins based on the missing of these metal-binding residues. The genes *serp* (*serpentine*) and *verm*, encoding two secreted ECM proteins with Ldl_recept_a and CBM_14 domains, are important for maintaining the size of tracheal tube and epicuticle structure in the developmental process [446].

SGL(g) is an enzyme domain with a beta-propeller fold. Several SGL(g)-containing proteins are present in the *Dmel* proteome. One of them, Smp-30, does not have a predicted signal peptide and should locate intracellularly like its human ortholog. Another one, Regucalcin, has two isoforms, one of which possesses a predicted signal peptide consistent with its extracellular localization. The metal binding residues of both *Dmel* Smp-30 and Regucalcin are conserved compared to the structure of human SMP-30/REGUCALCIN [447], suggesting that *Dmel* SGL-containing proteins are active glucolactonases. Two additional genes encode SGL-like proteins that also find hits to the Pfam domain Str_synth (strictosidine synthase). However, they lack the catalytic glutamate residue of strictosidine synthase. Consistent with previous computational studies of SGL and Str_synth domains [448], these proteins could instead perform a reaction of SGL based on the conservation of metal-binding residues.

Two genes (*CG7365* and *CG11029*) encode proteins with the Lipase_GDSL domain showing broad substrate specificity [449], which could be digestive enzymes as they have enriched expression in hindgut.

XC phosphoesterases. Eight enzyme domains (Alk_phosphatase, Endonuclease_NS, Metallophos, PLC(g), XendoU, Exo_endo_phos, Ribonuclease_T2, and His_Phos_2) are hydrolases of phosphoesters, such as phosphomonoesterases, phosphodiesterases, and various nucleases.

Alk_phosphatase domain (EC 3.1.3.1) has three zinc ions in the active site that are chelated by 10 residues. These residues are conserved in the 12 of 13 *Dmel* CSS gene products. Endonuclease_NS domain is found in non-specific endonucleases that degrade both single and double-stranded nucleic acids. Some genes encoding these enzymes are highly expressed in the digestive system, while one gene (*CG12917*) has enriched expression in testis. One gene (*Sid*), mainly expressed in fat body, is induced by stress response and could contribute to pathogen clearance [450]. *Dmel* Metallophos-containing genes encode several phosphoesterases such as purple acid phosphatase (CG1637, with Pur_ac_phosph_N and Metallophos_C domains), sphingomyelin phosphodiesterase (e.g., CG15534), and ecto-5′-nucleotidase (e.g., NT5E-2, with a C-terminal 5_nucleotid_C domain). PLC(g) is found in five genes—three with the GDPD (Glycerophosphoryl diester phosphodiesterase) domain and two with the PI-PLC-X (Phosphoinositide phospholipase C) domain. XendoU (2 genes), Exo_endo_phos (1 gene), Ribonuclease_T2 (1 gene), and His_Phos_2 (1 gene) are less abundant XC domains with phosphoesterase activity.

Sulfatase and DUF229. The Sulfatase domain catalyzes the hydrolysis of sulfate esters (EC 3.1.6.-). The *Dmel* gene *Sulf1* with this domain encodes an

enzyme that removes specific 6-*O*-sulfate groups from heparan sulfate in the extracellular space, which is functionally important for Wnt and Hedgehog signaling [451, 452]. Like the Sulfatase domain, the putative XC domain DUF229 belongs to the Pfam clan of Alk_phosphatase (Alkaline phosphatase-like). Thirteen of 16 DUF229 genes could encode catalytically active phosphatases based on conservation of three metal binding residues (two aspartic acids and one histidine) and one catalytic Ser/Thr residue [453].

Class E XC domains that are sugar hydrolase (EC 3.2.-.-) homologs. Six class E XC domains are found in sugar hydrolases or their inactive homologs. Glyco_hydro_18, Lysozyme(g), and Amylase are the most abundant ones, each with more than 10 genes found in the *Dmel* genome. The three less populated XC domains of sugar hydrolases are Glyco_hydro_16 (three genes), Glyco_hydro_20 (three genes), and Trehalase (one gene).

Glyco_hydro_18. A total of 17 *Dmel* genes encode proteins with the Glyco_hydro_18 domain [454]. While some of them are chitinases, others such as imaginal disc growth factors (IDGFs) have likely lost enzymatic activity due to the lack of active site residues (Asp and Glu in the DxxDxDxE motif). The nonenzymatic Glyco_hydro_18-containing proteins could maintain their ability to bind carbohydrate molecules or could have developed new functions in protein–protein interactions [455]. Chitinases and IDGFs are required for organization of ECM of cuticle [456], where chitin is the major structural component.

Lysozyme(g). Lysozymes are a diverse group of antibacterial peptidoglycan-hydrolyzing enzymes. *Dmel* has both c-type (chicken-type or conventional-type) lysozymes and i-type (invertebrate-type) lysozymes [457]. They correspond to Pfam domains Lys and Destabilase, respectively. Of the 12 Lys-containing genes, *LysB*, *LysD*, *LysE*, *LysS*, and *LysZ* are mainly expressed in the digestive system, while *LysP* is mainly expressed in the salivary gland. The muramidase activity of lysozymes relies on two acidic residues at the active site. Some i-type lysozymes, such as the destabilase from medicinal leech [458], also possess isopeptidase activity, which is dependent on a catalytic diad of serine and lysine residues at a structurally different location than the muramidase active site [459]. The five i-type lysozymes of *Dmel* lack active site residues of both muramidase and isopeptidase, suggesting that they are catalytically inactive.

Alpha-amylase. A majority of the 14 genes with an Alpha-amylase domain, such as *Amy-p* and *Amy-d*, show high expression levels in the digestive system, suggesting that they function as digestive enzymes. *Mal-B1* and *Mal-A5* are mainly expressed in salivary gland and fat body, respectively. All but one gene are predicted to encode active enzymes as they possess two conserved acidic residues in the active site. The catalytically inactive protein is encoded by *CD98hc* (*CD98 heavy chain*), where one of the acidic residue is changed to lysine. This type II TM protein, showing a nonspecific tissue expression, is likely involved in amino acid transport [460].

Glyco_hydro_16. Three GNBP (Gram-negative bacteria binding protein) genes (*GNBP1–3*) possess the Glyco_hydro_16 domain together with the CBM46 domain. GNBPs are important pattern recognition molecules for LPS and β-1,3-glucan in the *Dmel* innate immune response [137, 138]. The active site residues of Glyco_hydro_16 domain (two acid residues) in GNBPs are altered, suggesting loss of catalytic activity [138].

Glyco_hydro_20. The three genes with the Glyco_hydro_20 domain encode hexosaminidases, among which *fdl* is involved in N-glycan processing [461].

Trehalase. Trehalase domain catalyzes the hydrolysis of trehelose to glucose. Trehelose is the major sugar component of insect hemolymph. The two genes encoding trehelose synthesis (*Tps1*) and degradation (*Treh*, with the Trehalase domain) are important regulators of body water homeostasis [462].

Class E XC domains that are amidase homologs (EC 3.5.-.-). Four class E XC domains (Amidase_2, A_deaminase, CN_hydrolase, and Ceramidase_alk) are found in hydrolases of amide bonds other than peptide bonds. Three of them (Amidase_2, A_deaminase, and Ceramidase_alk) are zinc-dependent enzyme domains.

Amidase_2. Amidase_2 domain was found in 13 *Dmel* genes encoding peptidoglycan recognition proteins (PGRPs). These proteins, discovered in both invertebrates and vertebrates, are innate immunity molecules for pattern recognition of pathogens [463]. Catalytically active Amidase_2 domains possess two conserved histidines, one cysteine, and one tyrosine residue for zinc-binding and catalysis. Only five of the 13 Amidase_2 genes (*PGRP-SC1a*, *PGRP-SC1b*, *PGRP-SC2*, *PGRP-LB*, and *PGRP-SB1*) encode proteins that preserve these residues. They could act as immune effector proteins that hydrolyze the amide bond between MurNAc and L-alanine in bacterial peptidoglycans. Catalytically inactive PGRPs, capable of binding peptidoglycans, activate signaling pathways of immune response or induce proteolytic cascades that generate antimicrobial peptides.

A_deaminase. Six *Dmel* genes with XC A_deaminase (adenosine deaminase) domains are named Adgf (adenosine deaminase-related growth factor) [464]. Five of them, to the exclusion of *Adgf-E*, possess zinc binding and catalytic residues. These proteins could promote tissue proliferation by lowering the level of extracellular adenosine that has a negative effect on cell growth.

CN_hydrolase. Extracellular CN_hydrolase (Carbon-nitrogen hydrolase) domain [465] is present in four

Dmel genes. Three of them are predicted to encode GPI-anchored proteins. One gene *Btnd* (*Biotinidase*) encodes a protein repsonsible for cleaving biocytin (biotin-ε-lysine) to regenerate free biotin [466]. These genes could have a digestive role based on their expression patterns. Like A_deaminase domain, CN_hydrolase domain is also present in intracellular proteins without predicted signal peptide and TM segments.

Ceramidase_alk. A single extracellular ceramidase gene (*CDase*) possesses the Pfam Ceramidase_alk (neutral/alkaline non-lysosomal ceramidase) domain [467] that hydrolyzes ceramide to generate sphingosine and fatty acid.

Class E XC domains that are oxidoreductase homologs (EC 1.-.-.-). Eight XC domains (An_peroxidase, Lipo_10, GILT, Cu-oxidase(g), Cu2_monoox(g), Lysyl_oxidase, Sod_Cu, and DM13) are in enzymes with oxidoreductase activity that transfer electrons from donor (reductant) to acceptor (oxidant). Hemocyanin_M presents an interesting case where active enzymes are intracellular and catalytically inactive members are extracellular.

An_peroxidase. The heme-containing An_peroxidase domain acts on peroxide as an electron acceptor (EC 1.11.-.-) and is found in eight *Dmel* CSS proteins. Some An_peroxidase domains are found in chorion proteins (e.g., Pxd and CG4009) and could contribute to ECM cross linking and eggshell hardening [266]. Irc (Immune-regulated catalase) is a key component of host defense system and mediates homeostatic redox balance [468]. In contrast to other An_peroxidase-containing genes that do not have additional XC domains, the gene *Pxn* (*Peroxidasin*) possesses An_peroxidase together with VWC, Ig(g), and LRR(g) domains and plays important roles in ECM formation and immune response [89].

Lipo_10. The Lipo_10 domain (formerly named Chitin_bind_3) has lytic polysaccharide monooxygenase activity that oxidizes polysaccharides such as cellulose and chitin to facilitate their degradation [469]. Catalytically active members possess two histidines to bind copper in the active site [470]. Two of the six Lipo_10-containing genes (*CG4367* and *CG4362*) are highly expressed in the digestive system, suggesting a role in nutrient breakdown. One divergent and possibly inactive Lipo_10 domain was found in the *nahoda* gene with DOMON, EGF (g), Trypan_PARP, and SEA(g) domains.

GILT. GILT (Gamma-interferon-inducible lysosomal thiol reductase) domain with the thioredoxin fold was found in five *Dmel* genes. GILT1 is predicted to be a GPI-anchored protein and was found in cell membrane [471]. GILT2 and GILT3 with signal peptide predictions could locate in the lysosome and/or get secreted like mammalian GILT proteins. Two GILT genes *CG9427* and *CG41378* could encode proteins with different subcellular localizations. *CG9427* has two isoforms—one with a predicted signal peptide and the other without. *CG41378* isoforms could encode intracellular proteins as well as type II TM proteins with an N-terminal predicted TM segment. GILT proteins likely function in *Dmel* innate immune response [472].

Cu-oxidase(g). Four *Dmel* genes (*Mco1*, *MCO3*, *laccase2*, and *CG32557*) encode proteins with XC Cu-oxidase(g) domains. Structural analysis of a homologous protein revealed that each protein has three Cu-oxidase(g) domains (domains 1-3) with a β-barrel fold related to plastocyanin and azurin [473]. Domain 1 has a mononuclear copper chelated by two histidines, a cysteine, and a methionine. In addition, a trinuclear copper cluster is located in the interface between domain 1 and domain 3 and coordinated by eight histidines.

Cu2_monoox(g). Four *Dmel* genes encode proteins with the Cu2_monoox(g) domains [474]. These proteins contain two evolutionarily related jelly roll domains that each bind a copper ion. Two of these proteins also possess the DOMON domain that could bind heme. Tbh (tyramine β hydroxylase) [475] is the rate-limiting enzyme in the synthesis of octopamine, an invertebrate neurotransmitter. Phm (peptidylglycine-α-hydroxylating monooxygenase) catalyzes the hydroxylation of C-terminal glycine residues in signaling peptides [476]. The other two members, CG5235 and Olf413, are possible orthologs of mammalian protein monooxygenase DBH-like 1 (MOXD1), which has been shown to reside in the ER [477].

Lysyl_oxidase. Two genes (*lox* and *lox2*) with Lysyl_oxidase [478] domains exist in *Dmel*. They also possess SRCR(g) domains. They perform oxidative deamination of peptidyl lysine residues in collagen and elastin precursors to generate alpha-aminoadipic-delta-semialdehyde. Lysyl_oxidase domain binds copper for catalysis. The structure of a human lysyl oxidase homolog revealed that the Lysyl_oxidase domain adopts an IG-fold with three histines (in the sequence motif HxHxH) and a tyrosine as metal-binding residues in the active site [479]. These residues are preversed in *Dmel* Lox and Lox2.

Sod_Cu. Three genes (*Sod3*, *CG5948*, and *CG31028*) encode proteins with the IG-fold XC domain Sod_Cu, which is also present in two intracellular proteins (encoded by *Sod* and *Ccs*). A catalytically active Sod_Cu domain possesses a copper ion coordinated by three histidines and a zinc ion coordinated by three histidines and an aspartate [480] (Fig. 4E). *Sod3* has several isoforms encoding predicted secreted or GPI-anchored proteins with a single active Sod_Cu domain that preserves all metal-binding residues. *CG5948* and *CG31028* could encode proteins without catalytic activity as some of the metal-binding residues are altered.

DM13. DM13 is an XC domain with putative oxidoreductase activity [142]. It co-occurs with the

heme-binding DOMON domain in three *Dmel* genes. DM13 and DOMON could constitute an electron-transfer system that oxidatively modifies animal cell surface proteins [142]. The gene *knk* (*knickkopf*) encodes a GPI-anchored protein functioning in chitin organization in the cuticle and the tracheal system [223].

Hemocyanin_M. This domain is found in three prophenoloxidase genes (*PPO1*–3) without predicted signal peptide or TM segment. These genes likely encode intracellular active enzymes as they possess the six histidines for di-copper binding. These residues are not fully preserved in the seven secreted Hemocyanin_M-containing proteins with predicted signal peptides, suggesting that they are not capable of di-copper binding and catalysis. Instead, some of them function as storage proteins (Lsp1α, Lps1β, Lps1γ, and Lsp2) and their receptor (Fbp1) secreted by fat body cells [481, 482].

Class E XC domains that are transferase homologs (EC 2.-.-.-). G_glu_transpept, LCAT, and AIG2(g). Three of the six XC transferase domains are acyltransferases [G_glu_transpept, LCAT, and AIG2(g)]. Four *Dmel* genes have predicted XC G_glu_transpept (Gamma-glutamyltransferase) domains [483]. Gamma-glutamyltransferase breaks the gamma-glutamyl group from a substrate such as glutathione and transfers the glutamyl group to an acceptor such as an amino acid. The acceptor could also be a water molecule to generate glutamate (e.g., glutathione hydrolase). Two *Dmel* genes with LCAT (Lecithin–cholesterol acyltransferase) domains [484] exist. They transfer an acyl group from lecithin (phosphatidylcholine) to cholesterol to generate cholesterol ester, which is incorporated into lipoprotein particles. LCAT enzymes thus play important roles in lipoprotein metabolism. LCAT adopts the α/β hydrolase fold. The Pfam AIG2 clan has three domains (ChaC, AIG2_2, and GGACT) functioning as gamma-glutamyl cyclotransferases that are important in glutathione metabolism [485]. One *Dmel* AIG2(g)-containing protein (CG4306) has a predicted signal peptide and should be secreted, while several other AIG2(g)-containing proteins without predicted signal peptides are probably intracellular.

FAM20C, NDK, and Methyltransf_FA. Two phosphotransferase domains are found in *Dmel* CSS proteins. One is the FAM20C domain in a family of protein kinases responsible for phosphorylation of secreted proteins [486]. The other is the NDK (nucleotide diphosphate kinase) domain found in the product of the *awd* (*abnormal wing discs*) gene. It regulates endocytosis of various surface proteins and plays important roles in many developmental processes [487]. We also identified one XC methyltransferase domain (Methyltransf_FA). This domain generates methyl farnesoate (MF) from farnesoic acid (FA) in the biosynthetic pathway of juvenile hormone (JH) [488]. The structure and active site of Methyltransf_FA remain to be determined. In *Dmel*, the Methyltransf_FA domain resides in a single gene *NtR* that encodes a ligand-gated channel with the Neur_chan_LBD domain.

Class E XC domains that are lyase (EC 4.-.-.-) and isomerase (EC 5.-.-.-) homologs. XC lyase domains—Carb_anhydrase and NHL. There are the only two XC lyase domains. Carb_anhydrase (carbonic anhydrase, or carbonate dehydratase, EC 4.2.1.1) converts carbon dioxide and water to bicarbonate and protons, and vice versa [489]. Catalytically active members of this domain bind a zinc ion using three histidine residues. *Dmel* has 10 genes encoding proteins with XC Carb_anhydrase domains, four of which could be catalytically inactive due to substitutions at the metal-binding positions. The second lyase domain NHL, a beta-propeller domain found in the gene *Pal2* (*Peptidyl-α-hydroxyglycine-α-amidating lyase 2*), is responsile for C-terminal amidation of neuropeptides and peptide hormones [490].

XC isomerase domain—Pro_isomerase. Pro_isomerase (proline isomerase) is the only XC isomerase domain in *Dmel*. While many *Drosophila* XC enzyme domains are almost exclusively found in extracellular space, some can occur both extracellularly and intracellularly. Pro_isomerase is a versatile domain in terms of subcellular localization [491], as it is found in products of several genes with various subcellular localizations such as extracellular space, ER lumen, splicesome, and nucleus.

Class R XC domains—enzyme regulatory and inhibitory domains

This functional class includes 42 enzyme regulatory domains and inhibitor domains (Table 5). Enzyme regulatory domains were classified in this class based on literature analysis and domain architecture analysis. They often co-occur with the associated enzyme domains in the same protein products, such as CLIP [associated with Trypsin(g)] [492], Peptidase_M13_N (associated with Peptidase_M13), PPP-I(g) (associated with Peptidase_M14 and Peptidase_S8), ERAP1_C (associated with Peptidase_M1), and Alpha-amylase_C(g) (associated with Alpha-amylase). The most abundant enzyme regulatory domain is CLIP, a small cysteine-rich domain detected in 37 out of 259 Trypsin (g)-containing genes. This number could be an underestimate as some Trypsin(g)-containing proteins, such as Mas, Sb, Spirit, and CG18557, could have divergent CLIP domains not detected by the default score cutoffs. CLIP domains are located N-terminally to the Trypsin(g) domains. CLIP-containing serine proteases are synthesized as zymogens and activated by proteolytic cleavage in between the CLIP domain and the enzyme domain in protease cascades.

Another set of domains acting as enzyme inhibtors was also classified in class R based on literature analysis. These enzyme inhibitor domains, such as Kunitz_BPTI [493], Kazal(g) [494], and Serpin [495], generally do not co-occur with enzyme domains in the same protein. For example, none of the 26 Serpin-containing genes possesses other domains. Kunitz_BPTI, the most abundant class R domain in 48 *Dmel* genes, does not co-occur with other domains in most genes containing them. It is present in a few multi-domain genes with class P or other class R domains. For example, several Kunitz_BPTI-containing genes such as *fat-spondin* also have TSP_1, Spondin_N, and Reeler domains. One gene (*CG5639*) encoding a type I TM protein possesses multiple enzyme inhibitor domains including Kunitz_BPTI, Antastatin, Thyroglobulin_1, Lustrin_cystein, and WAP. Kazal(g) is the second most abundant class R domain identified in 43 *Dmel* genes. Like Kunitz_BPTI, most Kazal(g)-containing genes encode single-domain proteins, while a few of them also possess other enzyme inhibitor domains such as Pacifastin_I and Thyroglobulin_1, as well as class P domains such as EGF(g) and SPARC_Ca_bdg.

Table 5. Class R XC enzyme regulatory and inhibitory domains.

	XC domain	No. genes
1	Kunitz_BPTI	48
2	Kazal(g)	43
3	CLIP	37
4	Peptidase_M13_N	28
5	Serpin	26
6	PPP-I(g)	21
7	ERAP1_C	20
8	Alpha-amylase_C(g)	14
9	A1_Propeptide	12
10	TIL	12
11	Propeptide_C1(g)	11
12	Pep_M12B_propep	9
13	GD_N	8
14	Hemocyanin_N	7
15	Hemocyanin_C	7
16	Thiol-ester_cl	6
17	A_deaminase_N	6
18	WAP	6
19	Cystatin(g)	5
20	DPPIV_N	5
21	5_nucleotid_C	4
22	ADAM_CR	4
23	TIMP-like(g)	3
24	P_proprotein	3
25	ADAM17_MPD	3
26	Pacifastin_I	3
27	Hexosaminidase(g)	3
28	CarboxypepD_reg(g)	3
29	Thyroglobulin_1	3
30	CBM_20	3
31	PG_binding_1	2
32	PA	2
33	Hemopexin	2
34	Peptidase_M24_C	1
35	AMP_N-like(g)	1
36	Lustrin_cystein	1
37	Ceramidse_alk_C	1
38	Antistasin	1
39	DUF3740	1
40	DUF4976	1
41	Inhibitor_I68	1
42	Metallophos_C	1

Class U XC domains with unknown molecular function

Thirty-eight XC domains were classified in class U as they have unknown molecular function. Some of them have been experimentally shown to occur extracellularly, while others were predicted to be extracellular based on signal peptide and/or TM segment preditions. Most of them (26 out of 38) are named DUF (Domain of Unknown Function) (Table 6). About one third of the these domains (13 out of 38) are cysteine-rich domains, and eleven of them are domains with mostly low complexity or disordered regions. This class also includes several Pfam families corresponding to short peptide motifs. Two of them, GYR and YLP, could be involved in cuticle assembly and protein–protein interactions [496]. Residues in PT (repeats of XPTX) and Pentapeptide_2 (repeats of XNXGX) motifs could be involved in post-translational modifications such as glycosylation.

The most abundant class U domain is DM4_12, found in 36 *Dmel* genes. The majority of them have not been experimentally studied. The gene *Desi* (*Desiccate*) is expressed in the epidermis of *Dmel* larvae and plays a role in resistance to desiccation stress [497]. It is also expressed in the adult labellum and modulates taste sensitivities of sensilla to promote water intake in response to desiccation [498]. Another DM4_12-containing gene *geko* is involved in olfactory responses to ethanol [499]. The second most abundant class U domain is DUF1676, which is present in the Osiris gene family [500]. The Osiris genes have been associated with resistance to octanoic acid in *Drosophila sechellia* [501]. They play important roles in insect development, and are also involved in phenotypic plasticity as well as stress and immune responses [502–504]. Interestingly, DM4_12 and DUF1676 co-occur in two *Dmel* genes (*Osi17* and *Osi24*), suggesting that they could have related functions.

Besides DM4_12 and DUF1676, several class U domains are also involved in stress or immune response. The DIM domain is found in 14 *Dmel* genes. They are predicted to be secreted or GPI-anchored small proteins with two conserved cysteines. DIM-containing proteins are induced by infection, and they could be effectors (antimicrobial peptides) in immune response [505,506]. The Pfam GRP domain is found in six genes encoding proteins enriched with glycine and tyrosine. One of them is *Listericin*, whose expression is induced in response to infection by the gram-positive bacterium *Listeria monocytogenes*.

Listericin could encode a protein with antimicrobial activity [507]. The Pfam domain Turandot is found in eight *Dmel* genes that are involved in stress response. They are induced by various stressful conditions such as bacterial infection, heat, cold, and mechanical stimulus [508].

Assignment of cell membrane topology (CMT) categories to CSS proteins with *Drosophila* XC domains

For 2509 *Dmel* genes encoding proteins with XC domains, we manually assigned CMT categories to their protein products by analysis of prediction results of signal peptide, TM segment, and GPI signal sequence, as well as domain contents and literature. Five CMT categories include the following: E, secreted extracellular protein; S, type I or type III TM protein with a single TM and cytosolic C-terminus; T, type II TM protein with a single TM and cytosolic N-terminus; M, multi-pass TM protein; and G, GPI-anchored protein. The number of XC-domain-containing genes in each CMT category is shown as a Venn diagram in Fig. 9a. Overlapping areas correspond to genes with isoforms of different CMT assignments. CMT category E (secreted) is the most populated, as more than two third of the genes (1739 out of 2509) encoding predicted secreted proteins. CMT categories S (type I/III single-TM) and M (multi-pass TM) have similar numbers of genes, while CMT categories G (GPI-anchored) and T (type II single-TM) are less popular.

CMT category S is popular among several most abundant class P (protein–protein interaction) XC domains such as Ig(g), EGF(g), LRR(g), and fn3(g) (Table 7), as they tend to occur in type I TM cell surface receptors or adhesion molecules. The majority of uPAR_Ly6_toxin(g)-containing proteins are assigned CMT category G, while XC domains Tetraspanin and ASC are exclusively found in multi-pass TM proteins (CMT category M). The most populated CMT category is E (secreted) for the majority of the abundant XC domains in functional classes S (signaling molecules), B (non-protein binding), E (enzyme homolog), R (enzyme regulatory/inhibitory), and U (unknown molecular function) (Table 7). The exceptions are two class B domains PBP(g) and Periplas_BP(g) and the class U domain DUF1676, for which the most populated CMT category is M.

Genes encoding proteins with different signal peptides, CMT categories, and XC domain contents

Eukaryotic genome transcripts can be modulated through events such as alternative splicing [509], alternative promoter usage [510], alternative polyadenylation [511], and RNA editing [512]. The resulting difference in transcripts could lead to different protein products for a protein-coding gene. The repertoire of the protein products can be further expanded by translational events such as alternative translation initiation [513] and stop codon readthrough [514], as well as posttranslational events such as phosphorylation, glycosylation, and proteolysis [515].

Manual inspection of *Dmel* CSS proteins revealed that alternative splicing could affect N-terminal signal peptides. The gene *Ama* is an example where two isoforms have signal peptides of different lengths. The gene *Ptth* encodes three isoforms with signal peptides differing in their N-termini. The isoform PE has an eight-residue extension at the N-terminus compared to the isoform PG, while the isoform PF has an additional 26 residues added to the N-terminus of PE. Both *Sema2a* and *NimC1* have two isoforms of different signal peptides encoded by two non-overlapping exons. The *ptp99A* gene encodes one isoform (PF) that uses a different signal peptide than three other isoforms (PA, PB, and PC). This PF isoform contains an extra fn3(g) domain compared to isoforms PA, PB, and PC. This gene also encodes a putative intracellular isoform (PG) without signal peptide and TM segment.

Table 6. Class U putative XC domains of unknown molecular function

	XC domain	No. genes
1	DM4_12	36
2	DUF1676	22
3	DUF725	14
4	DIM	14
5	YLP	13
6	GYR	13
7	L71	11
8	DUF745	11
9	DUF4794	11
10	MBF2	9
11	Turandot	8
12	GRP	6
13	DUF4766	6
14	DUF4773	5
15	ACP53EA	4
16	DUF4786	3
17	DUF4813	3
18	DUF4816	3
19	DUF4779	3
20	DUF4768	2
21	PT	2
22	DUF4789	2
23	DUF4758	2
24	DUF4803	2
25	DUF4744	1
26	DUF1180	1
27	DUF2054	1
28	DUF1619	1
29	DUF4774	1
30	DUF4787	1
31	Pentapeptide_2	1
32	DUF4735	1
33	DUF3105	1
34	OAF	1
35	DUF4512	1
36	DUF2489	1
37	DUF3472	1
38	DUF2368	1

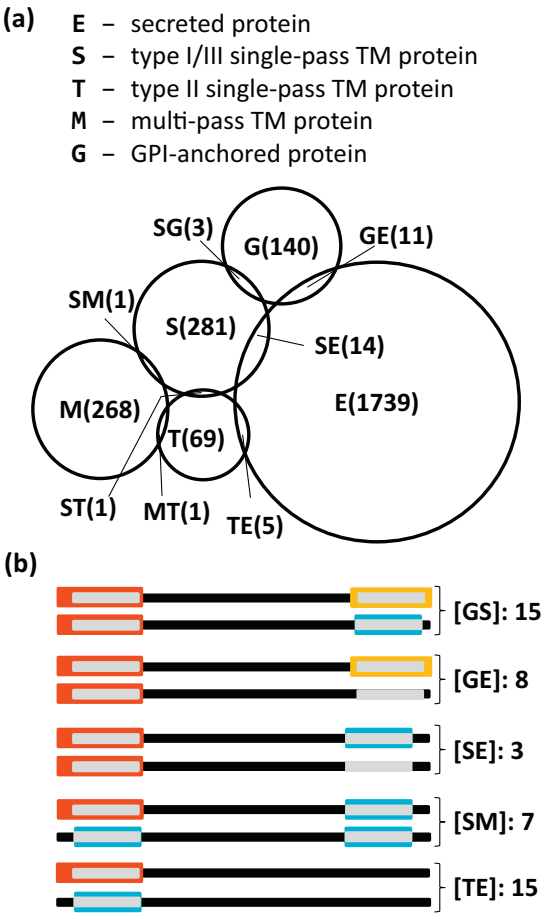


Fig. 9. Statistics of CMT assignment. (a) CMT category notations and the Venn diagram of the assignment of CMT categories. Each circle corresponds to the set of *Dmel* genes with product(s) of an assignment (S, T, E, M, and G), and the number of genes is shown inside the circle. Genes with products of more than one assignment are shown as overlapping regions, and their numbers are shown outside the circles. (b) Illustrations of unresolved assignments. Hydrophobic segments are shown as gray boxes. A hydrophobic segment could be part of a signal peptide (red box), a TM segment (blue box), a GPI signal sequence (orange box), or none of them.

Alternative splicing could also modulate the TM domain and GPI signaling sequence. One example is the *sns* gene that encodes three isoforms. Isoforms PA and PC have a single, but different TM after the XC domains, while the PB isoform contains both the TMs. Another example of isoforms with different membrane attachment is *Fas2*. Its isoforms PA, PD, PF, PG, and PH are predicted to be type I TM proteins, while isoforms PB and PC are predicted to be GPI-anchored proteins. The gene *beat-IV* has two isoforms. One isoform (PB) has a C-terminal segment predicted by PredGPI as a probable GPI-anchor sequence, while the other isoform does not have this segment and could be secreted instead.

Table 7. CMT assignment for popular XC domains in Fig. 2.

XC domain	CMT category					
	E	S	T	M	G	Un
Ig(g)	27	**73**	3	3	29	10
EGF(g)	35	**46**	2	4	2	2
LRR(g)	13	**48**	0	5	4	4
fn3(g)	8	**45**	0	1	8	3
Beta_propeller_XC(g)	18	**21**	5	0	2	1
uPAR_Ly6_toxin(g)	2	7	0	0	**29**	1
Ldl_recept_a	11	**20**	4	2	0	1
Tetraspannin	0	0	0	**38**	0	0
LRRNT	4	**25**	0	3	0	2
CUB(g)	**18**	12	0	1	0	1
ASC	0	0	0	**31**	0	0
CAP	**27**	0	0	0	2	1
Cystine-knot(g)	**24**	1	0	0	0	2
SVWC	**14**	0	0	0	0	0
Insulin(g)	**7**	0	0	0	0	0
wnt	**7**	0	0	0	0	0
Omega_toxin(g)	**6**	0	0	0	0	0
LIG(g)	**119**	0	0	0	0	0
Chitin_bind_4	**116**	0	0	0	1	1
CBM_14_19(g)	**110**	0	0	0	0	0
PBP(g)	1	0	0	**74**	2	0
PBP_GOBP	**50**	0	0	0	1	0
Lectin_C(g)	**38**	4	0	0	1	0
Aha1_BPI(g)	**31**	0	0	0	0	0
Periplas_BP(g)	1	5	0	**22**	0	0
DUF243	**26**	0	0	0	0	0
Trypsin(g)	**246**	0	4	0	7	2
Lipase	**30**	0	0	0	0	0
Peptidase_M13	**22**	0	7	0	0	0
Abhydrolase(g)	**24**	0	1	0	0	0
COesterase	**15**	6	1	0	2	0
Kunitz_BPTI	**45**	2	0	0	1	0
Kazal(g)	**40**	0	0	0	1	2
CLIP	**37**	0	0	0	0	0
Peptidase_M13_N	**22**	0	7	0	0	0
Serpin	**26**	0	0	0	0	0
DM4_12	**32**	0	0	2	0	2
DUF1676	0	0	0	**22**	0	0
DUF725	**14**	0	0	0	0	0
DIM	**13**	0	0	0	1	0
YLP	**13**	0	0	0	0	0
GYR	**13**	0	0	0	0	0

The table is ordered by domain classes P (red), S (dark red), B (cyan), E (magenta), R (orange), and U (blue). The most populated CMT category for each XC domain is in bold and gray background. "Un" stands for unresolved assignment.

The gene *chp* uses stop codon readthrough to create protein products with potentially different CMT categories. Four isoforms (PA, PC, PE, and PF) of *chp* have the same short hydrophobic segment at their C-termini, suggesting that they are GPI-anchored, as reported by experiments [516]. One isoform (PD) gets it C-terminus extended by stop codon readthrough to include a predicted TM segment (by both Phobius and TMHMM) as well as 19 residues after it.

Domain contents can also be modulated by alternative splicing. For example, the gene *Pxn* (*Peroxidasin*)

has two isoforms. One has the An_peroxidase domain while the other does not have it. As another example, the *NetA* isoform PA has a C-terminal NTR domain which is missing in the PB isoform.

Assignment of unresolved CMT categories

One recurring difficulty in CMT assignment is the C-terminal hydrophobic segment that can be predicted both as a GPI signal sequence and as a TM segment. For some cases with contradictory predictions, assignments were made based on experimental evidence available for the *Dmel* protein or its orthologs in other organisms. For example, the *Con* (*Connectin*) gene products have a C-terminal segment with both TM and GPI signal predictions. We assigned the CMT category of G based on experimental studies [180]. Another example is the *Cont* (*Contactin*) gene, which also encodes a protein with both TM and GPI signal predictions and is assigned CMT category of G based on experimental studies of the *Dmel* gene [517] and its vertebrate ortholog [518]. In cases without experimental evidence, we assigned an unresolved class [GS] (Fig. 9b). Fifteen genes were assigned the unresolved CMT category of [GS], including four Ig(g)-domain containing genes *dpr11*, *dpr20*, *beat-Va*, and *beat-VII*. For other genes, inconsistent prediction results among *Drosophila* orthologs were observed for the C-terminal hydrophobic segment that could be predicted to be a TM segment, the GPI signal sequence, or neither. This inconsistency could lead to unresolved assignments of [SE] and [GE], in addition to [GS] (Fig. 9b). Future experimental studies could help resolve some of these cases.

Likewise, an N-terminal hydrophobic segment can be contradictorily predicted to be part of a signal peptide and a TM segment by different programs, or inconsistently predicted to be part of a signal peptide or a TM segment among orthologs. In absence of additional predicted TM segments, this can result in unresolved CMT assignment of [TE] (Fig. 9b). For example, the protein product of the *Dmel sog* gene possesses an N-terminal hydrophobic segment (amino acid residues 55 to 73) predicted to be a TM segment by Phobius and TMHMM. It is also predicted to be part of a long signal peptide by SignalP 4.0 (truncation cutoff: 200). Moreover, the TM segment or signal peptide predictions are not consistent among *Drosophila* orthologs. The manual assignment of CMT category for the Sog protein is [TE]. Uncertainty of N-terminal hydrophobic segment predictions coupled with an additional C-terminal predicted TM segment can lead to unresolved CMT assignment of [SM] (Fig. 9B). One example is the *sev* (*sevenless*) gene, which encodes an RTK. More than 100 residues exist before the N-terminal hydrophobic segment, which is predicted to be a TM by Phobius. In contrast, the mammalian orthologs of Sevenless have a much shorter sequence before the N-terminal hydrophobic segment and are predicted to be type I TM proteins. No experimental evidence was found for *Dmel* Sevenless regarding whether the its N-terminal hydrophobic segment is cleaved by signal peptidase or serves as a TM segment. Its CMT category is thus assigned as [SM] since it contains a predicted TM segment near the C-terminus.

Dmel CSS proteins with potentially incorrect gene models

We observed a number of cases of potentially incorrect gene models and protein products of *Dmel* in FlyBase (version 2015_03) due to incomplete N-terminus. Missing exons and wrong translation start site could lead to shortened protein sequences at the N-terminus and result in no predictions of signal peptides for CSS proteins. One example is Beat-Vc. While several close orthologs of *Dmel* Beat-Vc have predicted signal peptides, it is missing for *Dmel* Beat-Vc. A translation of its mRNA indeed recovered the N-terminal sequence segment (MSLLRLLGALL-LATFNG) highly similar to its orthologs in other *Drosophila* species. This segment constitutes most of the missing signal peptide (Table 8). We made potential corrections for a total of 39 *Dmel* genes where the missing signal peptide can be recovered for their proteins (Table 8).

The FlyXCDB database

We developed FlyXCDB (http://prodata.swmed.edu/FlyXCDB) that reports numerous computational predictions and serves as a resource to study XC domains in *Dmel* CSS proteins. This database contains a main information table for about 2,500 *Dmel* genes with XC domains. The information in this table for any gene includes CMT assignment for its product(s), the number of orthologous *Drosophila* proteins with a link to their alignment, summary scores of signal peptide/TM segment/GPI signal sequence predictions (SP, SP200, SP_PHO, ave_SP, TM_PHO, and ave_GPI, see Materials and Methods for their definitions), XC domains and other domains detected by HMMER and HHsearch, and FlyBase (2015_03) Gene Ontology cellular component terms. Each gene also has a link to a dedicated page containing detailed information about the predictions for its orthologous proteins. A separate web page stores the information table for all *Dmel* protein-coding genes, which is slower to load.

FlyXCDB also has a web page dedicated to XC domain classification with a table that includes manually defined *Drosophila* XC domains classified into six classes. For each XC domain, the table has its classification code (P, S, B, E, R, or U), the number of associated *Dmel* genes, and the Pfam domain(s) contributing to the XC domain. The number of associated *Dmel* genes is linked to a separate page that contains the information table for genes associated

Table 8. Cases of gene model corrections for possibly missing N-terminal segments.

Gene	Isoform(s)	Corrected N-terminus[a]
beat-Vc	PA	**MSLLRLLGALLLATFNG**MAPVPAEGLHLSNLSVPRIIDVAQKAKLFCSYAM
ItgaPS5	PA	**MHRLLFLIFLALKYQSNA**MNFSPLPNRVIDAPKHLKTRMIQVRSSYFGYSL
Jon74E	PA,PB	**MQISTILVFLLILVQGRSISCLD**MGHGIGGRIAGGELARANQFPYQVGLSI
Mal-B2	PC,PD	**MRAPLIQILLFSLLNSGSI**MAGLVKSDTEDFIDWWPHTVFYQIYPRSFKDS
Spn28Db	PA	**MQGNNKIKYLVLLLIATSVLGKFKLNLLELV**MDKAESNFIASPLCIEIGIS
upd2	PA	**MPAFTLNASQQSQSSRSSRSAHSSWSCHSRQVLLVI**MILSVVMPFTKARHL
CG10041	PA	**MCCWQSLCSIAWFAA**MSAAQETLSDTPQNSTPLLATTVSTTKVISFRPRYP
CG11570	PA	**MRQGSVVIYLGLLAFIAIVDCQENNNLVSIVTDHSIQCPPFDDPNHNV**MLP
CG11836	PI	**MSLKYIIFFWMLTANYSSVLSLEYSKGFNESDAINTIHT[8]FLFDTI**FRI
CG12662	PA	**MGYYRNLLGFITLFGFLQLTKQHCY**MMKLVGAKSWDTPSYGLKLNMFEKNN
CG13247	PA	**MDMMTATWMLCLVLCSTA**MATEEESGSGLFNVTGNPLDTEAAPKLPSEVQK
CG14495	PA,PB	**MEALRRFLRTETRCRLNGIHTLILCVLFLGFSVAGE**MPDLFTPEPDLTIDE
CG14720	PA	**MRLEVAIYAFLSC**MHLCSADGHLKRLSRSLIFPPTSPTRVQFIGGIGIPVE
CG15358	PB,PC	**MQKLSIWLLLALFAWNAHAPSAA**MPSVCLLQDAPQQCGEFCLTALSPMLDH
CG17189	PA	**MLLIGLL**MLLHAGLQGAQCVAFYTEKPSYIESCKIYEPEFTKCSTRSIQAF
CG17279	PB,PC	**MRLIVFVCCLW**MAPSLGQLPPEIEKCRAGDSICIAETVTRILRLYPKGLPS
CG30187	PC,PD	**MQTRLAWIPVIFWFLKDVGASIFLDQICGINIALKITGGHNAAFQNSVW**MA
CG30270	PC	**MQLAKYSIFLFVLCGTLIPVEIVARSLEKKQPRGSVART[23]GGGIPE**MP
CG30280	PB,PC	**MFPGVPVFSVCIALSV**MIFCHAETLNLFGNLEAIYEQAENALSTLQESLLQ
CG31465	PA,PB,PC	**MSRHLFHIFCILAITRPIPISCTEAKDQELS**MSHKEASSWLRTQDDLVPKH
CG31664	PB,PC	**MNTRCTLLFTLATLLAFGKPAHAYKTYYLMGEEDKIHLD[32]VSDLLE**MF
CG32413	PC	**MLGRIALFLTIIYVVFKA**MFLRWVIEKPPFKFDDDEEHFNATLAKLLKPRS
CG32984	PA	**MMLKMKTFTSVSYKKCILCIFILFLLIVWKITSYGVNDV[14]NSSNCQ**MA
CG32985	PA	**MARSISYSFKRLIVSFLAVTLICLWS**MNSSRVDENFSARDSQEVEKLFYVE
CG33645	PB	**MNFLLLFFSATLIFNS**MRCEERNFRVYIKEVNITHLDTDLYEKFECKVYQV
CG33768	PA	**MLKIVCVL**MVIFVSQAVSSSVTLNRVQCEKNAKFFATLNVTSVNSTIYADI
CG34234	PA	**MRRCKTIQ**MLFLCLMMRMRESHERPTPKNLLEQMLPADTFDVIRRIPRSDE
CG34303	PA	**MFSVRVAFLCIFCFNIFGCIPVSKIPYIKSFGLKSIVIS[16]APELGK**MK
CG34428	PA	**MKFSIVFILVSCLLYQV**MASLGSLFKHQRSKRAPIPWLIYPTTSPTRVQFI
CG42685	PA	**MFLYFMTISMSLVLI**MENSGTPLMSLLDGTFLEAYNRSTESPEGGPTSHPW
CG42878	PA	**MQCRGLSLAVLCLLGYTRAMLHGPNYCTHGDL**MVKCIPVCPKICADFLYRQ
CG43294	PA	**MLNKLPLIVLLCAVFSGFC**MAQTVQRWPFPNDCHRYYETRLLDCPPEFYWN
CG4367	PA	**MNTYKFILFLS**MGFLNRMGELEGHGMMLSPTGRSSRWRYDNSAPTNYDDNA
CG5267	PA	**MMAWPKVYFVVQLIGILFQKHVL**MAEDSAEDQITWPDDIAEPTLESELTWP
CG5897	PB	**MRVPPCGFHLCVILAIVAEIRGFS**MEDKCKLWAGTGYIGDPSDCQAWGYCQ
CG7653	PB	**MFKCLLSERENCYKNVALYGLLLLLCLCQVACSAKTQLY[143]WIRNT**MP
CG7763	PA,PB	**MQKSIIFLVLPLCLSSSYSAACEGVESDSQCAAYCYGVLNPCIAS**MGNLQR
CG9168	PA	**MSLGRNVFLLCSIQFLLCKAKLSLHSHYPNVAKTLENKKFYVDSPSCK**MPE
CG9500	PA	**MRSVWIYVAFLGLSFYSTEA**MDSESFQNDTAIRNKPELKSLYKLVLALLEE

[a] Inferred missing N-terminal residues are shown in bold red letters. The number of omitted letters in long missing sequences is shown in brackets. Predicted signal peptide and TM segment are underlined and double-underlined, respectively.

with that XC domain, in the same format as the main gene information table of FlyXCDB. One example of the gene information table is shown for the class B lipid-binding XC domain Calycin(g) (Fig. 10). We detected this domain in 11 genes potentially encoding *Dmel* CSS proteins, including *Glaz* and *Nlaz* that have been experimentally shown to encode secreted proteins. Four genes (*CG5399*, *CG3706*, *CG34256*, and *CG43050*) possess divergent Calycin(g) domains only detected by HHsearch. They encode candidate *Dmel* lipid-binding proteins that could be targets for future experimental studies.

Materials and Methods

Defining orthologous protein groups for *Dmel* protein-coding genes

For each of the protein-coding genes of *Dmel* (FlyBase version FB2015_03), we aim to study if it encodes CSS protein(s) and if any known XC domains are present in its protein product(s) by analyzing predictions of signal sequences and TM segments, domain contents, gene ontology, and literature

***Dmel* genes with XC domain - Calycin(g)**

Number of genes: 11

#	CMT	GeneID	Name	#SEQ	ave_SP	SP	SP200	SP_PHO	TM_PHO	ave_GPI	XC	Others	XC_HH
1	CMT_E	FBgn0033799	GLaz	13	**0.923**	0.923	0.923	0.923	0.000	0.000	Calycin[Lipocalin:1, Lipocalin_2:0.692]	N/A	N/A
2	CMT_E	FBgn0053126	NLaz	13	**0.974**	0.923	1.000	1.000	0.077	0.000	Calycin[Lipocalin_2:1]	N/A	N/A
3	CMT_E	FBgn0030334	Karl	12	**0.583**	0.500	0.750	0.500	0.333	0.000	Calycin[Lipocalin:0.833]	N/A	N/A
4	CMT_G	FBgn0264775	CG44013	5	**0.800**	0.800	0.800	0.800	**0.600**	0.600	Calycin[Lipocalin:0.8]	N/A	N/A
5	CMT_G	FBgn0264776	CG44014	12	**0.861**	0.833	1.000	0.750	**0.833**	0.792	Calycin[Lipocalin:0.667, Lipocalin_2:0.167]	N/A	N/A
6	CMT_G	FBgn0051446	CG31446	14	**1.000**	1.000	1.000	1.000	**0.857**	0.929	Calycin[Lipocalin:0.214, Lipocalin_2:0.643]	N/A	N/A
7	CMT_E	FBgn0051659	CG31659	14	**0.810**	0.786	0.857	0.786	0.071	0.000	Calycin[Triabin:0.5, Lipocalin:0.429]	N/A	N/A
8	CMT_G	FBgn0038353	CG5399	12	**0.917**	0.917	0.917	0.917	**1.000**	1.000	N/A	N/A	Calycin[Lipocalin_2, Lipocalin, Triabin, Nitrophorin, VDE, ApoM]
9	CMT_E	FBgn0040342	CG3706	12	0.278	0.000	0.167	0.667	**1.000**	0.000	N/A	N/A	Calycin[CrtC, Svf1, DUF2804]
10	CMT_E	FBgn0085285	CG34256	15	**0.711**	0.267	0.933	0.933	0.000	0.100	N/A	N/A	Calycin[Lipocalin_2, Triabin, VDE]
11	CMT_E	FBgn0262352	CG43050	16	**0.625**	0.625	0.625	0.625	0.000	0.031	N/A	N/A	Calycin[Lipocalin_2, Triabin, Nitrophorin, VDE, ApoM]

Fig. 10. FlyXCDB gene information table for class B XC domain Calycin(g). The "CMT" column shows the assigned CMT categories. The "GeneID" column has FlyBase gene ids linked to the gene pages in FlyBase. The "Name" column has gene symbols linked to FlyXCDB pages with detailed information of predictions. The "#SEQ" column contains the numbers of *Drosophila* nonredundant orthologs with links to their alignments. The columns of "ave_SP," "SP," "SP200," "SP_PHO," "TM_PHO," and "ave_GPI" store the prediction scores of signal peptide, TM segment, and GPI signal sequence (see Materials and Methods). The "XC" column contains XC domains detected by HMMER and the detection fractions among orthologs. Pfam domains belonging to the same XC domain are grouped in brackets with the XC domain name before the brackets. The "Others" column contains other Pfam domains detected by HMMER. The "XC_HH" column contains XC domains detected by HHsearch. The "Others_HH" column (containing other domains detected by HHsearch) and the "GO_CC" column (gene ontology cellular component terms) are not shown in this figure due to space limit.

(Fig. 1A). To aid the prediction power, we applied a comparative genomics approach to studying the protein products of each *Dmel* gene as well as the protein products of its orthologous genes in the 12 *Drosophila* species with sequenced whole genomes—*D. ananassae* (abbreviation: *Dana*), *D. erecta* (*Dere*), *D. grimshawi* (*Dgri*), *D. melanogaster* (*Dmel*), *D. mojavensis* (*Dmoj*), *D. persimilis* (*Dper*), *D. pseudoobscura* (*Dpse*), *D. sechellia* (*Dsec*), *D. simulans* (*Dsim*), *D. yakuba* (*Dyak*), *D. virilis* (*Dvir*), and *D. willistoni* (*Dwil*). The orthologous relationships among the genes of the 12 *Drosophila* genomes were predicted by OrthoDB [519], with each gene represented by its longest protein product. For any protein-coding gene encoding multiple protein isoforms, we examined its protein products and removed sequence redundancy by keeping only non-identical protein products. The orthologous protein group of a *Dmel* protein-coding gene consists of its non-redundant protein products and the non-redundant protein products of its orthologous genes from the 12 *Drosophila* genomes. For each protein-coding *Dmel* gene, a multiple sequence alignment was obtained by MAFFT [520] for all members of the orthologous protein group.

Predictions of signal peptide, TM segment, GPI signal sequence, and Pfam domains

SignalP (version: 4.0) [521] and Phobius [522] were used to predict N-terminal signal peptide for secretory pathway targeting for all members in each orthologous group. For SignalP, two separate predictions were made with truncation cutoffs set to 70 (default setting) and 200 (a setting that allows prediction of long signal peptide up to 200-amino-acid residues). TM segments were predicted by Phobius and TMHMM [523]. PredGPI [524] and FragAnchor [525] were used to predict C-terminal GPI signal sequences.

To aid in the discovery of *Dmel* CSS proteins and the XC domains in them, we developed several scores based on predictions of signal peptide, TM segment, and GPI signal sequence. For any *Dmel* gene, the SP score is the fraction of signal peptide prediction by SignalP4.0 with default setting (truncation cutoff: 70) for proteins in its orthologous group (the number of predicted signal peptides divided by the total number of proteins in the orthologous group of that gene). The SP200 score is the fraction of signal

peptide prediction by SignalP4.0 with a truncation cutoff of 200 residues. The SP_PHO score is the fraction of signal peptide prediction by Phobius. The ave_SP score is the average of SP, SP200, and SP_PHO. TM_PHO is the average number of predicted TMs by Phobius. The ave_GPI score is the average fraction of GPI signal sequences predicted by PredGDI and fragAnchor. These scores (SP, SP200, SP_PHO, ave_SP, TM_PHO, and ave_GPI) are reported in gene information tables in FlyXCDB.

For protein domain detection, HMMER [526] was used to search each protein in an orthologous group against the Pfam database (version: 28) [527]. For any detected Pfam domain type, HMMER reports a global, full-length e-value based on combined similarity of individual domains as well as domain-level e-values for detected individual domain regions. By default, the Pfam domain hit to a region was reported as a true positive if all three conditions are satisfied: (1) the full-length e-value is no worse than 0.01, (2) the domain-level e-value is no worse than 0.1, and (3) the coverage of the Pfam HMM model is no less than 0.5. One exception is for hits with superior domain-level e-values (1e–10 or better), for which the coverage requirement is relaxed. In practice, we found that some domains with a significant portion of compositional biased regions (such as low complexity regions, coiled coil regions, TM segments, and signal peptides) were detected under these settings. We manually inspected these spurious Pfam hits in potential extracellular regions and set the domain-level e-value cutoff to 1e−4 for these compositionally biased Pfam domains. HMMER hits of different homologous Pfam domains were often reported in the same region. In these situations, the best hit in the region is kept. To do this, we applied a domain hit filtering procedure. Any domain hit is filtered out if there is another domain hit with a better e-value and covering at least 70% of the region. HHsearch [528] was applied to each non-redundant protein of *Dmel* against the Pfam database for prediction of divergent domains. By default, true positives were considered for hits with probability scores no less than 95 and the domain coverages no less than 0.5. HHsearch was also used against the PDB database for identification of homologs with structures.

Obtaining an initial set of candidate Pfam XC domains

The SMART database [36] (version 7.0) has a set of SMART domains defined as extracellular. For most of them, we mapped SMART-defined XC domains to Pfam [35] (version 28) domains from the links provided by the SMART database. In cases where the links to Pfam domains are missing, we performed HMMER searches against the Pfam database using the SMART domain sequences as queries to help identify the corresponding Pfam domains. This procedure created an initial set of candidate Pfam domains. Some of these domains were later removed by manual analysis if one of the following conditions applied: (1) The Pfam domain was not detected in *Drosophila* proteins (e.g., B_lectin, Autotransporter, and C1q). (2) The Pfam domain was only identified in *Drosophila* intracellular proteins (e.g., AIP3 and Peptidase_C2). (3) The Pfam domain is mostly made up of TM segments and does not have an extracellular independent folding unit (e.g., 7TM_GPCR_Srsx).

Analysis of candidate *Dmel* genes encoding CSS proteins to expand the set of XC domains

We added more Pfam domains to the XC domain set based on analysis of HMMER and HHsearch results of proteins from candidate *Drosophila* CSS genes. The set of candidate *Dmel* CSS genes (genes potentially encoding CSS proteins) consisted of *Dmel* genes satisfying one of these conditions: (1) the ave_SP score was no less than 0.4. (2). The TM_PHO score was no less than 0.5. (3) The ave_GPI score was no less than 0.4. (4) One or more proteins in the orthologous group of a *Dmel* gene had a predicted XC domain initially mapped from the SMART database XC domain set. Additional *Drosophila* XC domains were assigned by manual analysis of these candidate *Dmel* CSS genes, with the help of FlyBase (version 2015_03) gene ontology terms of cellular component (ftp://ftp.flybase.net/releases/FB2015_03/precomputed_files/ontologies/go-basic.obo.gz) as well as literature mining.

Classification of *Drosophila* XC domains

Drosophila XC domains were manually classified into six classes mainly according to their molecular functions. We relied on manual literature mining and structure analysis in assigning the classes. Class P XC domains are those likely involved in protein–protein interactions, which are crucial for cell adhesion, receptor-protein ligand interactions, and establishment of ECM. Assignments of some domains in this class, such as Ig(g), fn3(g), EGF(g), and LRR(g), were based on experimental studies supporting their roles in protein–protein interactions [42,43]. Protein–protein interactions involving some domains were additionally supported by three-dimensional structures, such as Cadherin (g) [93]. Some XC domains, such as Vitelline_membr and and Chorion_3 [260,262], were assigned class P as they were found in major protein components of ECMs. Class S domains, defined as those mainly found in XC signaling molecules, are also involved in protein–protein interactions. We made a separate class for them as they are involved in interactions with cell surface receptors in signaling pathways. Domains in this class, together with signaling molecules without known domains (such as a number of neuropeptides),

are described in section Class S XC domains mainly found in extracellular signaling molecules of Results and Discussion. Literature and structural analyses were used in assigning class B (likely involved in binding non-protein molecules and groups) and class E (enzyme homologs) domains. Class R domains are those that act as enzyme inhibitors or potentially serve as as enzyme regulatory domains. Enzyme inhibitor domains, such as Serpin and Kazal(g), were assigned class R based on literature. Enzyme regulatory domains were assigned class R based on literature analysis as well as domain architecture analysis, as they often co-occur with enzyme domains. The rest of the XC domains, without known molecular function, were assigned class U.

Enzyme domain EC number and active site analysis

For each enzyme domain, we assigned the EC (Enzyme Commission) number based on information from the Pfam database and literature analysis. For a majority of the enzyme domains, catalytically important residues were defined as those involved in catalysis as well as metal binding residues in the active site, if available. Definitions of catalytically important residues were conducted manually by inspecting structures of the enzyme domains and literature mining, and the positions of these residues were assigned to a known structure of the enzyme domain. *Dmel* gene products with the XC enzyme domain were then aligned to the sequence of the structure homolog. If any isoform of a gene contains an intact set of catalytically important residues, the gene is considered to encode catalytically active enzymes. Otherwise, the gene is considered to encode catalytically inactive enzyme domain. The numbers of genes encoding catalytically active enzymes are undetermined for two enzyme domains without known structures and experimental knowledge of their active sites (DM13 and Methyltransf_FA).

CMT category assignment

The previously defined prediction scores (SP, SP200, SP_PHO, ave_SP, TM_PHO, and ave_GPI) were useful in aiding manual assignment of CMT categories to *Dmel* gene products. For example, secreted proteins often have the ave_SP score near 1 and the TM_PHO score near zero, and type I TM proteins often have both the ave_SP score and the TM_PHO score near 1. However, we still mainly relied on manaul analysis for CMT assignment, due to difficulties caused by potentially incorrect predictions, inconsistent predictions among *Drosophila* orthologs, occurrence of multiple isoforms in a gene, and endomembrane (such as ER and Golgi) proteins with signal peptide/TM segment predictions.

Acknowledgments

This work was supported in part by the National Institutes of Health (GM094575 to N.V.G.) and the Welch Foundation (I-1505 to N.V.G.).

Received 11 March 2018;
Received in revised form 31 May 2018;
Accepted 2 June 2018
Available online 8 June 2018

Keywords:
Drosophila cell surface and secreted proteins;
extracellular domain definition and classification;
cell adhesion and signaling;
extracellular enzymes;
cell membrane topology

Abbreviations used:
CSS, cell surface and secreted; ECM, extracellular matrix; SRP, signal recognition particle; ER, endoplasmic reticulum; TM, transmembrane; GPI, glycophosphatidylinositol; *Dmel*, *Drosophila melanogaster*; IG-fold, immunoglobulin-like fold; DIPs, Dpr-interacting proteins; RTKs, receptor tyrosine kinases; RTPs, receptor tyrosine phosphatases; GPCRs, G protein-coupled receptors; TEPs, thioester-containing proteins; ZP, Zona_pellucida; TNF, tumor necrosis factor; LPS, lipopolysaccharide; GABA, gamma-aminobutyric-acid; LDL, low-density lipoprotein; iGluRs, ionotropic glutamate receptors; BM, basement membrane; HSPG, heparan sulfate proteoglycan; Ptth, prothoracicotropic hormone; FGF, fibroblast growth factor; CRF, corticotropin-releasing factor; PGRPs, peptidoglycan recognition proteins; FA, farnesoic acid; CMT, cell membrane topology.

References

[1] J.P. da Cunha, P.A. Galante, J.E. de Souza, R.F. de Souza, P.M. Carvalho, D.T. Ohara, R.P. Moura, S.M. Oba-Shinja, S.K. Marie, W.A. Silva Jr., R.O. Perez, B. Stransky, M. Pieprzyk, J. Moore, O. Caballero, J. Gama-Rodrigues, A. Habr-Gama, W.P. Kuo, A.J. Simpson, A.A. Camargo, L.J. Old, S.J. de Souza, Bioinformatics construction of the human cell surfaceome, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 16752–16757.

[2] L. Fagerberg, K. Jonasson, G. von Heijne, M. Uhlen, L. Berglund, Prediction of the human membrane proteome, Proteomics 10 (2010) 1141–1149.

[3] J. Meinken, G. Walker, C.R. Cooper, X.J. Min, MetazSecKB: the human and animal secretome and subcellular proteome knowledgebase, Database (Oxford) 2015 (2015).

[4] M. Uhlen, L. Fagerberg, B.M. Hallstrom, C. Lindskog, P. Oksvold, A. Mardinoglu, A. Sivertsson, C. Kampf, E. Sjostedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C.A. Szigyarto, J. Odeberg, D. Djureinovic, J.O. Takanen, S.

Hober, T. Alm, P.H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J.M. Schwenk, M. Hamsten, K. von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. von Heijne, J. Nielsen, F. Ponten, Proteomics. Tissue-based map of the human proteome, Science 347 (2015) 1260419.

[5] D. Bausch-Fluck, A. Hofmann, T. Bock, A.P. Frei, F. Cerciello, A. Jacobs, H. Moest, U. Omasits, R.L. Gundry, C. Yoon, R. Schiess, A. Schmidt, P. Mirkowska, A. Hartlova, J.E. Van Eyk, J.P. Bourquin, R. Aebersold, K.R. Boheler, P. Zandstra, B. Wollscheid, A mass spectrometric-derived cell surface protein atlas, PLoS One 10 (2015), e0121314. .

[6] I. Ben-Shlomo, S. Yu Hsu, R. Rauch, H.W. Kowalski, A.J. Hsueh, Signaling receptome: a genomic and evolutionary perspective of plasma membrane receptors involved in signal transduction, Sci. STKE 2003 (2003) RE9.

[7] L.F. Reichardt, K.J. Tomaselli, matrix molecules and their receptors: functions in neural development, Annu. Rev. Neurosci. 14 (1991) 531–570.

[8] S. Akira, S. Uematsu, O. Takeuchi, Pathogen recognition and innate immunity, Cell 124 (2006) 783–801.

[9] B. Lemaitre, I. Miguel-Aliaga, The digestive tract of *Drosophila melanogaster*, Annu. Rev. Genet. 47 (2013) 377–404.

[10] T. Rozario, D.W. DeSimone, The extracellular matrix in development and morphogenesis: a dynamic view, Dev. Biol. 341 (2010) 126–140.

[11] S. Shao, R.S. Hegde, Membrane protein insertion at the endoplasmic reticulum, Annu. Rev. Cell Dev. Biol. 27 (2011) 25–56.

[12] R.J. Keenan, D.M. Freymann, R.M. Stroud, P. Walter, The signal recognition particle, Annu. Rev. Biochem. 70 (2001) 755–775.

[13] D.J. Schnell, D.N. Hebert, Protein translocons: multifunctional mediators of protein translocation across membranes, Cell 112 (2003) 491–505.

[14] S.H. White, G. von Heijne, The machinery of membrane protein assembly, Curr. Opin. Struct. Biol. 14 (2004) 397–404.

[15] L. Kall, A. Krogh, E.L. Sonnhammer, A combined transmembrane topology and signal peptide prediction method, J. Mol. Biol. 338 (2004) 1027–1036.

[16] C. Gao, Y. Cai, Y. Wang, B.H. Kang, F. Aniento, D.G. Robinson, L. Jiang, Retention mechanisms for ER and Golgi membrane proteins, Trends Plant Sci. 19 (2014) 508–515.

[17] V. Goder, M. Spiess, Topogenesis of membrane proteins: determinants and dynamics, FEBS Lett. 504 (2001) 87–93.

[18] M. Higy, T. Junne, M. Spiess, Topogenesis of membrane proteins at the endoplasmic reticulum, Biochemistry 43 (2004) 12716–12722.

[19] G. von Heijne, Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule, J. Mol. Biol. 225 (1992) 487–494.

[20] S. Mayor, H. Riezman, Sorting GPI-anchored proteins, Nat. Rev. Mol. Cell Biol. 5 (2004) 110–120.

[21] D.A. Brown, J.K. Rose, Sorting of GPI-anchored proteins to glycolipid-enriched membrane subdomains during transport to the apical cell surface, Cell 68 (1992) 533–544.

[22] R.G. Spiro, Protein glycosylation: nature, distribution, enzymatic formation, and disease implications of glycopeptide bonds, Glycobiology 12 (2002) 43R–56R.

[23] A. Zhou, G. Webb, X. Zhu, D.F. Steiner, Proteolytic processing in the secretory pathway, J. Biol. Chem. 274 (1999) 20745–20748.

[24] F.R. Maxfield, T.E. McGraw, Endocytic recycling, Nat. Rev. Mol. Cell Biol. 5 (2004) 121–132.

[25] M.N. Seaman, The retromer complex—endosomal protein recycling and beyond, J. Cell Sci. 125 (2012) 4693–4702.

[26] B.D. Grant, J.G. Donaldson, Pathways and mechanisms of endocytic recycling, Nat. Rev. Mol. Cell Biol. 10 (2009) 597–608.

[27] K.J. Brown, C.A. Formolo, H. Seol, R.L. Marathi, S. Duguez, E. An, D. Pillai, J. Nazarian, B.R. Rood, Y. Hathout, Advances in the proteomic investigation of the cell secretome, Expert Rev. Proteomics 9 (2012) 337–345.

[28] R.O. Hynes, A. Naba, Overview of the matrisome—an inventory of extracellular matrix constituents and functions, Cold Spring Harb. Perspect. Biol. 4 (2012) a004903.

[29] E. Hohenester, J. Engel, Domain structure and organisation in extracellular matrix proteins, Matrix Biol. 21 (2002) 115–128.

[30] K.M. Beckingham, J.D. Armstrong, M.J. Texada, R. Munjaal, D. A. Baker, *Drosophila melanogaster*—the model organism of choice for the complex biology of multi-cellular organisms, Gravit. Space Biol. Bull. 18 (2005) 17–29.

[31] G.M. Rubin, *Drosophila melanogaster* as an experimental organism, Science 240 (1988) 1453–1459.

[32] M.D. Adams, S.E. Celniker, R.A. Holt, C.A. Evans, J.D. Gocayne, P.G. Amanatides, S.E. Scherer, P.W. Li, R.A. Hoskins, R.F. Galle, R.A. George, S.E. Lewis, S. Richards, M. Ashburner, S.N. Henderson, G.G. Sutton, J.R. Wortman, M.D. Yandell, Q. Zhang, L.X. Chen, R.C. Brandon, Y.H. Rogers, R.G. Blazej, M. Champe, B.D. Pfeiffer, K.H. Wan, C. Doyle, E.G. Baxter, G. Helt, C.R. Nelson, G.L. Gabor, J.F. Abril, A. Agbayani, H.J. An, C. Andrews-Pfannkoch, D. Baldwin, R.M. Ballew, A. Basu, J. Baxendale, L. Bayraktaroglu, E.M. Beasley, K.Y. Beeson, P.V. Benos, B.P. Berman, D. Bhandari, S. Bolshakov, D. Borkova, M.R. Botchan, J. Bouck, P. Brokstein, P. Brottier, K.C. Burtis, D.A. Busam, H. Butler, E. Cadieu, A. Center, I. Chandra, J.M. Cherry, S. Cawley, C. Dahlke, L.B. Davenport, P. Davies, B. de Pablos, A. Delcher, Z. Deng, A.D. Mays, I. Dew, S.M. Dietz, K. Dodson, L.E. Doup, M. Downes, S. Dugan-Rocha, B.C. Dunkov, P. Dunn, K.J. Durbin, C.C. Evangelista, C. Ferraz, S. Ferriera, W. Fleischmann, C. Fosler, A.E. Gabrielian, N.S. Garg, W.M. Gelbart, K. Glasser, A. Glodek, F. Gong, J.H. Gorrell, Z. Gu, P. Guan, M. Harris, N.L. Harris, D. Harvey, T.J. Heiman, J.R. Hernandez, J. Houck, D. Hostin, K.A. Houston, T.J. Howland, M.H. Wei, C. Ibegwam, et al., The genome sequence of *Drosophila melanogaster*, Science 287 (2000) 2185–2195.

[33] Drosophila 12 Genomes, C, A.G. Clark, M.B. Eisen, D.R. Smith, C.M. Bergman, B. Oliver, T.A. Markow, T.C. Kaufman, M. Kellis, W. Gelbart, V.N. Iyer, D.A. Pollard, T.B. Sackton, A.M. Larracuente, N.D. Singh, J.P. Abad, D.N. Abt, B. Adryan, M. Aguade, H. Akashi, W.W. Anderson, C.F. Aquadro, D.H. Ardell, R. Arguello, C.G. Artieri, D.A. Barbash, D. Barker, P. Barsanti, P. Batterham, S. Batzoglou, D. Begun, A. Bhutkar, E. Blanco, S.A. Bosak, R.K. Bradley, A.D. Brand, M.R. Brent, A.N. Brooks, R.H. Brown, R.K. Butlin, C. Caggese, B.R. Calvi, A. Bernardo de Carvalho, A. Caspi, S. Castrezana, S.E. Celniker, J.L. Chang, C. Chapple, S. Chatterji, A. Chinwalla, A. Civetta, S.W. Clifton, J.M. Comeron, J.C. Costello, J.A. Coyne, J. Daub, R.G. David, A.L. Delcher, K. Delehaunty, C.B. Do, H. Ebling, K. Edwards, T. Eickbush, J.D. Evans, A. Filipski, S. Findeiss, E. Freyhult, L. Fulton, R. Fulton, A.C. Garcia, A. Gardiner, D.A. Garfield, B.E. Garvin, G. Gibson, D. Gilbert, S. Gnerre, J. Godfrey, R. Good, V. Gotea, B. Gravely, A.J. Greenberg, S. Griffiths-Jones, S. Gross, R. Guigo, E.A. Gustafson, W. Haerty, M.W. Hahn, D.L. Halligan, A.L.

Halpern, G.M. Halter, M.V. Han, A. Heger, L. Hillier, A.S. Hinrichs, I. Holmes, R.A. Hoskins, M.J. Hubisz, D. Hultmark, M. A. Huntley, D.B. Jaffe, et al., Evolution of genes and genomes on the *Drosophila* phylogeny, Nature 450 (2007) 203–218.

[34] P. McQuilton, S.E. St Pierre, J. Thurmond, C. FlyBase, FlyBase 101—the basics of navigating FlyBase, Nucleic Acids Res. 40 (2012) D706–D714.

[35] R.D. Finn, A. Bateman, J. Clements, P. Coggill, R.Y. Eberhardt, S.R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E.L. Sonnhammer, J. Tate, M. Punta, Pfam: the protein families database, Nucleic Acids Res. 42 (2014) D222–D230.

[36] I. Letunic, T. Doerks, P. Bork, SMART 7: recent updates to the protein domain annotation resource, Nucleic Acids Res. 40 (2012) D302–D305.

[37] R.D. Finn, J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S.R. Eddy, E.L. Sonnhammer, A. Bateman, Pfam: clans, web tools and services, Nucleic Acids Res. 34 (2006) D247–D251.

[38] B.D. Angst, C. Marcozzi, A.I. Magee, The cadherin superfamily: diversity in form and function, J. Cell Sci. 114 (2001) 629–641.

[39] G. Dennis Jr., B.T. Sherman, D.A. Hosack, J. Yang, W. Gao, H.C. Lane, R.A. Lempicki, DAVID: Database for Annotation, Visualization, and Integrated Discovery, Genome Biol. 4 (2003) P3.

[40] C. Vogel, S.A. Teichmann, J. Pereira-Leal, The relationship between domain duplication and recombination, J. Mol. Biol. 346 (2005) 355–365.

[41] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. 13 (2003) 2498–2504.

[42] R.A. Carrillo, E. Ozkan, K.P. Menon, S. Nagarkar-Jaiswal, P.T. Lee, M. Jeon, M.E. Birnbaum, H.J. Bellen, K.C. Garcia, K. Zinn, Control of synaptic connectivity by a network of *Drosophila* IgSF cell surface proteins, Cell 163 (2015) 1770–1782.

[43] B.Z. Shilo, Signaling by the *Drosophila* epidermal growth factor receptor pathway during development, Exp. Cell Res. 284 (2003) 140–149.

[44] C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, J.M. Thornton, CATH—a hierarchic classification of protein domain structures, Structure 5 (1997) 1093–1108.

[45] H. Cheng, R.D. Schaeffer, Y. Liao, L.N. Kinch, J. Pei, S. Shi, B.H. Kim, N.V. Grishin, ECOD: an evolutionary classification of protein domains, PLoS Comput. Biol. 10 (2014), e1003926.

[46] Y. Harpaz, C. Chothia, Many of the immunoglobulin superfamily domains in cell adhesion molecules and surface receptors belong to a new structural set which is close to that containing variable domains, J. Mol. Biol. 238 (1994) 528–539.

[47] C. Vogel, S.A. Teichmann, C. Chothia, The immunoglobulin superfamily in *Drosophila melanogaster* and *Caenorhabditis elegans* and the evolution of complexity, Development 130 (2003) 6317–6328.

[48] M. Nakamura, D. Baldwin, S. Hannaford, J. Palka, C. Montell, Defective proboscis extension response (DPR), a member of the Ig superfamily required for the gustatory response to salt, J. Neurosci. 22 (2002) 3463–3472.

[49] E. Ozkan, R.A. Carrillo, C.L. Eastman, R. Weiszmann, D. Waghray, K.G. Johnson, K. Zinn, S.E. Celniker, K.C. Garcia, An extracellular interactome of immunoglobulin and LRR proteins reveals receptor-ligand networks, Cell 154 (2013) 228–239.

[50] L. Tan, K.X. Zhang, M.Y. Pecot, S. Nagarkar-Jaiswal, P.T. Lee, S.Y. Takemura, J.M. McEwen, A. Nern, S. Xu, W. Tadros, Z. Chen, K. Zinn, H.J. Bellen, M. Morey, S.L. Zipursky, Ig superfamily ligand and receptor pairs expressed in synaptic partners in *Drosophila*, Cell 163 (2015) 1756–1769.

[51] G.C. Pipes, Q. Lin, S.E. Riley, C.S. Goodman, The Beat generation: a multigene family encoding IgSF proteins related to the Beat axon guidance molecule in *Drosophila*, Development 128 (2001) 4545–4552.

[52] D. Fambrough, C.S. Goodman, The Drosophila beaten path gene encodes a novel secreted protein that regulates defasciculation at motor axon choice points, Cell 87 (1996) 1049–1058.

[53] M. Siebert, D. Banovic, B. Goellner, H. Aberle, *Drosophila* motor axons recognize and follow a Sidestep-labeled substrate pathway to reach their target fields, Genes Dev. 23 (2009) 1052–1062.

[54] D. Schmucker, J.C. Clemens, H. Shu, C.A. Worby, J. Xiao, M. Muda, J.E. Dixon, S.L. Zipursky, Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity, Cell 101 (2000) 671–684.

[55] G. Neves, J. Zucker, M. Daly, A. Chess, Stochastic yet biased expression of multiple Dscam splice variants by individual cells, Nat. Genet. 36 (2004) 240–246.

[56] D. Hattori, S.S. Millard, W.M. Wojtowicz, S.L. Zipursky, Dscam-mediated cell recognition regulates neural circuit formation, Annu. Rev. Cell Dev. Biol. 24 (2008) 597–620.

[57] W. Tadros, S. Xu, O. Akin, C.H. Yi, G.J. Shin, S.S. Millard, S.L. Zipursky, Dscam proteins direct dendritic targeting through adhesion, Neuron 89 (2016) 480–493.

[58] T. Zeev-Ben-Mordehai, A. Paz, Y. Peleg, L. Toker, S.G. Wolf, E.H. Rydberg, J.L. Sussman, I. Silman, Amalgam, an axon guidance *Drosophila* adhesion protein belonging to the immunoglobulin superfamily: over-expression, purification and biophysical characterization, Protein Expr. Purif. 63 (2009) 147–157.

[59] H. Kose, D. Rose, X. Zhu, A. Chiba, Homophilic synaptic target recognition mediated by immunoglobulin-like cell adhesion molecule Fasciclin III, Development 124 (1997) 4143–4152.

[60] G. Grenningloh, E.J. Rehm, C.S. Goodman, Genetic analysis of growth cone guidance in *Drosophila*: fasciclin II functions as a neuronal recognition molecule, Cell 67 (1991) 45–57.

[61] C. Shelton, K.S. Kocherlakota, S. Zhuang, S.M. Abmayr, The immunoglobulin superfamily member Hbs functions redundantly with Sns in interactions between founder and fusion-competent myoblasts, Development 136 (2009) 1159–1168.

[62] S. Zhuang, H. Shao, F. Guo, R. Trimble, E. Pearce, S.M. Abmayr, Sns and Kirre, the *Drosophila* orthologs of Nephrin and Neph1, direct adhesion, fusion and formation of a slit diaphragm-like structure in insect nephrocytes, Development 136 (2009) 2335–2344.

[63] H. Peradziryi, N.A. Kaplan, M. Podleschny, X. Liu, P. Wehner, A. Borchers, N.S. Tolwinski, PTK7/Otk interacts with Wnts and inhibits canonical Wnt signalling, EMBO J. 30 (2011) 3729–3740.

[64] P. Cafferty, L. Yu, Y. Rao, The receptor tyrosine kinase Off-track is required for layer-specific neuronal connectivity in *Drosophila*, Development 131 (2004) 5287–5295.
[65] K. Linnemannstons, C. Ripp, M. Honemann-Capito, K. Brechtel-Curth, M. Hedderich, A. Wodarz, The PTK7-related transmembrane proteins off-track and off-track 2 are co-receptors for *Drosophila* Wnt2 required for male fertility, PLoS Genet. 10 (2014), e1004443. .
[66] T.A. Springer, An extracellular beta-propeller module predicted in lipoprotein and scavenger receptors, tyrosine kinases, epidermal growth factor precursor, and extracellular matrix components, J. Mol. Biol. 283 (1998) 837–862.
[67] J.P. Himanen, K.R. Rajashankar, M. Lackmann, C.A. Cowan, M. Henkemeyer, D.B. Nikolov, Crystal structure of an Eph receptor–ephrin complex, Nature 414 (2001) 933–938.
[68] K. Anamika, K.R. Abhinandan, K. Deshmukh, N. Srinivasan, Classification of nonenzymatic homologues of protein kinases, Comp. Funct. Genomics 365637 (2009).
[69] T. Hatzihristidis, N. Desai, A.P. Hutchins, T.C. Meng, M.L. Tremblay, D. Miranda-Saavedra, A *Drosophila*-centric view of protein tyrosine phosphatases, FEBS Lett. 589 (2015) 951–966.
[70] J. den Hertog, C. Blanchetot, A. Buist, J. Overvoorde, A. van der Sar, L.G. Tertoolen, Receptor protein-tyrosine phosphatase signalling in development, Int. J. Dev. Biol. 43 (1999) 723–733.
[71] A.A. Zarin, J.P. Labrador, Motor axon guidance in *Drosophila*, Semin. Cell Dev. Biol. (2017) (Article in press).
[72] M. Jeon, H. Nguyen, S. Bahri, K. Zinn, Redundancy and compensation in axon guidance: genetic analysis of the *Drosophila* Ptp10D/Ptp4E receptor tyrosine phosphatase subfamily, Neural Dev. 3 (2008) 3.
[73] A.N. Mohebiany, R.M. Nikolaienko, S. Bouyain, S. Harroch, Receptor-type tyrosine phosphatase ligands: looking for the needle in the haystack, FEBS J. 280 (2013) 388–400.
[74] A.N. Fox, K. Zinn, The heparan sulfate proteoglycan syndecan is an in vivo ligand for the *Drosophila* LAR receptor tyrosine phosphatase, Curr. Biol. 15 (2005) 1701–1711.
[75] H.K. Lee, A. Cording, J. Vielmetter, K. Zinn, Interactions between a receptor tyrosine phosphatase and a cell surface ligand regulate axon guidance and glial-neuronal communication, Neuron 78 (2013) 813–826.
[76] M. Yamamoto, S. Ohsawa, K. Kunimasa, T. Igaki, The ligand Sas and its receptor PTP10D drive tumour-suppressive cell competition, Nature 542 (2017) 246–250.
[77] T.A. Evans, Embryonic axon guidance: insights from *Drosophila* and other insects, Curr. Opin. Insect Sci. 18 (2016) 11–16.
[78] H. Blockus, A. Chedotal, Slit-Robo signaling, Development 143 (2016) 3037–3044.
[79] M. Hiramoto, Y. Hiromi, E. Giniger, Y. Hotta, The *Drosophila* Netrin receptor Frazzled guides axons by controlling Netrin distribution, Nature 406 (2000) 886–889.
[80] K. Keleman, B.J. Dickson, Short- and long-range repulsion by the *Drosophila* Unc5 netrin receptor, Neuron 32 (2001) 605–617.
[81] C. Dostert, E. Jouanguy, P. Irving, L. Troxler, D. Galiana-Arnoux, C. Hetru, J.A. Hoffmann, J.L. Imler, The Jak–STAT signaling pathway is required but not sufficient for the antiviral response of *Drosophila*, Nat. Immunol. 6 (2005) 946–953.
[82] C. Ghiglione, O. Devergne, E. Georgenthum, F. Carballes, C. Medioni, D. Cerezo, S. Noselli, The *Drosophila* cytokine receptor Domeless controls border cell migration and epithelial polarization during oogenesis, Development 129 (2002) 5437–5447.
[83] K.H. Fisher, W. Stec, S. Brown, M.P. Zeidler, Mechanisms of JAK/STAT pathway negative regulation by the short coreceptor Eye Transformer/Latran, Mol. Biol. Cell 27 (2016) 434–441.
[84] N. Okamoto, R. Nakamori, T. Murai, Y. Yamauchi, A. Masuda, T. Nishimura, A secreted decoy of InR antagonizes insulin/IGF signaling to restrict body growth in *Drosophila*, Genes Dev. 27 (2013) 87–97.
[85] N. Alic, M.P. Hoddinott, G. Vinti, L. Partridge, Lifespan extension by increased expression of the *Drosophila* homologue of the IGFBP7 tumour suppressor, Aging Cell 10 (2011) 137–147.
[86] M. Grigorian, T. Liu, U. Banerjee, V. Hartenstein, The proteoglycan Trol controls the architecture of the extracellular matrix and balances proliferation and differentiation of blood progenitors in the *Drosophila* lymph gland, Dev. Biol. 384 (2013) 301–312.
[87] M. Nakayama, E. Suzuki, S. Tsunoda, C. Hama, The matrix proteins hasp and hig exhibit segregated distribution within synaptic clefts and play distinct roles in synaptogenesis, J. Neurosci. 36 (2016) 590–606.
[88] I.A. Kramerova, N. Kawaguchi, L.I. Fessler, R.E. Nelson, Y. Chen, A.A. Kramerov, M. Kusche-Gullberg, J.M. Kramer, B.D. Ackley, A.L. Sieron, D.J. Prockop, J.H. Fessler, Papilin in development; a pericellular protein with a homology to the ADAMTS metalloproteinases, Development 127 (2000) 5475–5485.
[89] M. Soudi, M. Zamocky, C. Jakopitsch, P.G. Furtmuller, C. Obinger, Molecular evolution, structure, and function of peroxidasins, Chem. Biodivers. 9 (2012) 1776–1793.
[90] R.R. Anholt, Olfactomedin proteins: central players in development and disease, Front. Cell Dev. Biol. 2 (2014) 6.
[91] L.C. Zeng, Z.G. Han, W.J. Ma, Elucidation of subfamily segregation and intramolecular coevolution of the olfactomedin-like proteins by comprehensive phylogenetic analysis and gene expression pattern assessment, FEBS Lett. 579 (2005) 5443–5453.
[92] P. Hulpiau, F. van Roy, Molecular evolution of the cadherin superfamily, Int. J. Biochem. Cell Biol. 41 (2009) 349–369.
[93] H. Oda, M. Takeichi, Evolution: structural and functional diversity of cadherin at the adherens junction, J. Cell Biol. 193 (2011) 1137–1146.
[94] J.M. Carvajal-Gonzalez, M. Mlodzik, Mechanisms of planar cell polarity establishment in *Drosophila*, F1000Prime Rep. 6 (2014) 98.
[95] H. Kawamori, M. Tai, M. Sato, T. Yasugi, T. Tabata, Fat/Hippo pathway regulates the progress of neural differentiation signaling in the *Drosophila* optic lobe, Develop. Growth Differ. 53 (2011) 653–667.
[96] T. Usui, Y. Shima, Y. Shimada, S. Hirano, R.W. Burgess, T.L. Schwarz, M. Takeichi, T. Uemura, Flamingo, a seven-pass transmembrane cadherin, regulates planar cell polarity under the control of Frizzled, Cell 98 (1999) 585–595.
[97] S. Prakash, H.M. McLendon, C.I. Dubreuil, A. Ghose, J. Hwa, K.A. Dennehy, K.M. Tomalty, K.L. Clark, D. Van Vactor, T.R. Clandinin, Complex interactions amongst N-cadherin, DLAR, and Liprin-alpha regulate *Drosophila* photoreceptor axon targeting, Dev. Biol. 336 (2009) 10–19.
[98] P.L. Chen, T.R. Clandinin, The cadherin Flamingo mediates level-dependent interactions that guide photoreceptor target choice in *Drosophila*, Neuron 58 (2008) 26–33.

[99] J. Pei, N.V. Grishin, Expansion of divergent SEA domains in cell surface proteins and nucleoporin 54, Protein Sci. 26 (2017) 617–630.
[100] S. Wang, V. Tsarouhas, N. Xylourgidis, N. Sabri, K. Tiklova, N. Nautiyal, M. Gallio, C. Samakovlis, The tyrosine kinase Stitcher activates Grainy head and epidermal wound healing in *Drosophila*, Nat. Cell Biol. 11 (2009) 890–895.
[101] A.W. Dodds, S.K. Law, The phylogeny and evolution of the thioester bond-containing proteins C3, C4 and alpha 2-macroglobulin, Immunol. Rev. 166 (1998) 15–26.
[102] M. Lagueux, E. Perrodou, E.A. Levashina, M. Capovilla, J.A. Hoffmann, Constitutive expression of a complement-like protein in toll and JAK gain-of-function mutants of *Drosophila*, Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 11427–11432.
[103] U. Theopold, M. Schmid, Thioester-containing proteins: At the crossroads of immune effector mechanisms, Virulence (2017) 1–3.
[104] S. Hall, C. Bone, K. Oshima, L. Zhang, M. McGraw, B. Lucas, R.G. Fehon, R.E. t Ward, Macroglobulin complement-related encodes a protein required for septate junction organization and paracellular barrier function in *Drosophila*, Development 141 (2014) 889–898.
[105] M. Williams, R. Baxter, The structure and function of thioester-containing proteins in arthropods, Biophys. Rev. 6 (2014) 261–272.
[106] I.D. Campbell, M.J. Humphries, Integrin structure, activation, and interactions, Cold Spring Harb. Perspect. Biol. 3 (2011).
[107] D.L. Brower, Platelets with wings: the maturation of *Drosophila* integrin biology, Curr. Opin. Cell Biol. 15 (2003) 607–613.
[108] J.P. Xiong, T. Stehle, R. Zhang, A. Joachimiak, M. Frech, S.L. Goodman, M.A. Arnaout, Crystal structure of the extracellular segment of integrin alpha Vbeta3 in complex with an Arg–Gly–Asp ligand, Science 296 (2002) 151–155.
[109] L.C. Dekkers, M.C. van der Plas, P.B. van Loenen, J.T. den Dunnen, G.J. van Ommen, L.G. Fradkin, J.N. Noordermeer, Embryonic expression patterns of the *Drosophila* dystrophin-associated glycoprotein complex orthologs, Gene Expr. Patterns 4 (2004) 153–159.
[110] L. Jovine, C.C. Darie, E.S. Litscher, P.M. Wassarman, Zona pellucida domain proteins, Annu. Rev. Biochem. 74 (2005) 83–114.
[111] L. Han, M. Monne, H. Okumura, T. Schwend, A.L. Cherry, D. Flot, T. Matsuda, L. Jovine, Insights into egg coat assembly and egg-sperm interaction from the X-ray structure of full-length ZP3, Cell 143 (2010) 404–415.
[112] A. Jazwinska, M. Affolter, A family of genes encoding zona pellucida (ZP) domain proteins is expressed in various epithelial tissues during *Drosophila* embryogenesis, Gene Expr. Patterns 4 (2004) 413–421.
[113] I. Fernandes, H. Chanut-Delalande, P. Ferrer, Y. Latapie, L. Waltzer, M. Affolter, F. Payre, S. Plaza, Zona pellucida domain proteins remodel the apical compartment for localized cell shape changes, Dev. Cell 18 (2010) 64–76.
[114] F. Roch, C.R. Alonso, M. Akam, *Drosophila* miniature and dusky encode ZP proteins required for cytoskeletal reorganisation during wing morphogenesis, J. Cell Sci. 116 (2003) 1199–1207.
[115] K.F. Chen, N. Peschel, R. Zavodska, H. Sehadova, R. Stanewsky, QUASIMODO, a Novel GPI-anchored zona pellucida protein involved in light input to the *Drosophila* circadian clock, Curr. Biol. 21 (2011) 719–729.
[116] Y.D. Chung, J. Zhu, Y. Han, M.J. Kernan, nompA encodes a PNS-specific, ZP domain protein required to connect mechanosensory dendrites to sensory structures, Neuron 29 (2001) 415–428.
[117] P.C. Harris, S. Rossetti, Molecular diagnostics for autosomal dominant polycystic kidney disease, Nat. Rev. Nephrol. 6 (2010) 197–206.
[118] G.W. Moy, L.M. Mendoza, J.R. Schulz, W.J. Swanson, C.G. Glabe, V.D. Vacquier, The sea urchin sperm receptor for egg jelly is a modular protein with extensive homology to the human polycystic kidney disease protein, PKD1, J. Cell Biol. 133 (1996) 809–817.
[119] M. Nagae, K. Nishikawa, N. Yasui, M. Yamasaki, T. Nogi, J. Takagi, Structure of the F-spondin reeler domain reveals a unique beta-sandwich fold with a deformable disulfide-bonded loop, Acta Crystallogr. D Biol. Crystallogr. 64 (2008) 1138–1145.
[120] K. Tan, M. Duquette, J.H. Liu, J. Lawler, J.H. Wang, The crystal structure of the heparin-binding reelin-N domain of f-spondin, J. Mol. Biol. 381 (2008) 1213–1223.
[121] K. Tan, J. Lawler, The structure of the Ca(2)+-binding, glycosylated F-spondin domain of F-spondin—A C2-domain variant in an extracellular matrix protein, BMC Struct. Biol. 11 (2011) 22.
[122] Y. Li, C. Cao, W. Jia, L. Yu, M. Mo, Q. Wang, Y. Huang, J.M. Lim, M. Ishihara, L. Wells, P. Azadi, H. Robinson, Y.W. He, L. Zhang, R.A. Mariuzza, Structure of the F-spondin domain of mindin, an integrin ligand and pattern recognition molecule, EMBO J. 28 (2009) 286–297.
[123] A.G. Uren, D.L. Vaux, TRAF proteins and meprins share a conserved domain, Trends Biochem. Sci. 21 (1996) 244–245.
[124] P. Arnold, A. Otte, C. Becker-Pauly, Meprin metalloproteases: molecular regulation and function in inflammation and fibrosis, Biochim. Biophys. Acta 1864 (2017) 2096–2104.
[125] J. Inoue, T. Ishida, N. Tsukamoto, N. Kobayashi, A. Naito, S. Azuma, T. Yamamoto, Tumor necrosis factor receptor-associated factor (TRAF) family: adapter proteins that mediate cytokine signaling, Exp. Cell Res. 254 (2000) 14–24.
[126] M.L. Winberg, J.N. Noordermeer, L. Tamagnone, P.M. Comoglio, M.K. Spriggs, M. Tessier-Lavigne, C.S. Goodman, Plexin A is a neuronal semaphorin receptor that controls axon guidance, Cell 95 (1998) 903–916.
[127] Y. Izumi, Y. Yanagihashi, M. Furuse, A novel protein complex, Mesh-Ssk, is required for septate junction formation in the *Drosophila* midgut, J. Cell Sci. 125 (2012) 4923–4933.
[128] I. Callebaut, D. Gilges, I. Vigon, J.P. Mornon, HYR, an extracellular module involved in cellular adhesion and related to the immunoglobulin-like fold, Protein Sci. 9 (2000) 1382–1390.
[129] T. Shinoda, H. Ogawa, F. Cornelius, C. Toyoshima, Crystal structure of the sodium-potassium pump at 2.4 A resolution, Nature 459 (2009) 446–450.
[130] X. Huang, J.T. Warren, J. Buchanan, L.I. Gilbert, M.P. Scott, *Drosophila* Niemann-Pick type C-2 genes control sterol homeostasis and steroid biosynthesis: a model of human neurodegenerative disease, Development 134 (2007) 3733–3742.
[131] X.Z. Shi, X. Zhong, X.Q. Yu, *Drosophila melanogaster* NPC2 proteins bind bacterial cell wall components and may function in immune signal pathways, Insect Biochem. Mol. Biol. 42 (2012) 545–556.
[132] G.D. Findlay, X. Yi, M.J. Maccoss, W.J. Swanson, Proteomics reveals novel *Drosophila* seminal fluid proteins transferred at mating, PLoS Biol. 6 (2008), e178.

[133] A. Xu, S.K. Park, S. D'Mello, E. Kim, Q. Wang, C.W. Pikielny, Novel genes expressed in subsets of chemosensory sensilla on the front legs of male *Drosophila melanogaster*, Cell Tissue Res. 307 (2002) 381–392.
[134] E. Starostina, A. Xu, H. Lin, C.W. Pikielny, A *Drosophila* protein family implicated in pheromone perception is related to Tay-Sachs GM2-activator protein, J. Biol. Chem. 284 (2009) 585–594.
[135] K. Brejc, W.J. van Dijk, R.V. Klaassen, M. Schuurmans, J. van Der Oostn, A.B. Smit, T.K. Sixma, Crystal structure of an ACh-binding protein reveals the ligand-binding domain of nicotinic receptors, Nature 411 (2001) 269–276.
[136] G. Gisselmann, J. Plonka, H. Pusch, H. Hatt, *Drosophila melanogaster* GRD and LCCH3 subunits form heteromultimeric GABA-gated cation channels, Br. J. Pharmacol. 142 (2004) 409–413.
[137] E. Warr, S. Das, Y. Dong, G. Dimopoulos, The Gram-negative bacteria-binding protein gene family: its role in the innate immune system of *Anopheles gambiae* and in anti-*Plasmodium* defence, Insect Mol. Biol. 17 (2008) 39–51.
[138] Y.S. Kim, J.H. Ryu, S.J. Han, K.H. Choi, K.B. Nam, I.H. Jang, B. Lemaitre, P.T. Brey, W.J. Lee, Gram-negative bacteria-binding protein, a pattern recognition receptor for lipopolysaccharide and beta-1,3-glucan that mediates the signaling for the induction of innate immune genes in *Drosophila melanogaster* cells, J. Biol. Chem. 275 (2000) 32721–32727.
[139] L. Serre, B. Vallee, N. Bureaud, F. Schoentgen, C. Zelwer, Crystal structure of the phosphatidylethanolamine-binding protein from bovine brain: a novel structural class of phospholipid-binding proteins, Structure 6 (1998) 1255–1265.
[140] C.W. Pikielny, G. Hasan, F. Rouyer, M. Rosbash, Members of a family of *Drosophila* putative odorant-binding proteins are expressed in different subsets of olfactory hairs, Neuron 12 (1994) 35–49.
[141] C.R. Chillakuri, D. Sheppard, M.X. Ilagan, L.R. Holt, F. Abbott, S. Liang, R. Kopan, P.A. Handford, S.M. Lea, Structural analysis uncovers lipid-binding properties of Notch ligands, Cell Rep. 5 (2013) 861–867.
[142] L.M. Iyer, V. Anantharaman, L. Aravind, The DOMON domains are involved in heme and sugar recognition, Bioinformatics 23 (2007) 2660–2664.
[143] T.J. Larson, M. Ehrmann, W. Boos, Periplasmic glycerophosphodiester phosphodiesterase of *Escherichia coli*, a new enzyme of the glp regulon, J. Biol. Chem. 258 (1983) 5428–5432.
[144] M.V. Airola, W.J. Allen, M.J. Pulkoski-Gross, L.M. Obeid, R. C. Rizzo, Y.A. Hannun, Structural basis for ceramide recognition and hydrolysis by human neutral ceramidase, Structure 23 (2015) 1482–1491.
[145] M.A. Wouters, I. Rigoutsos, C.K. Chu, L.L. Feng, D.B. Sparrow, S.L. Dunwoodie, Evolution of distinct EGF domains with specific functions, Protein Sci. 14 (2005) 1091–1103.
[146] A. Shmueli, O. Cohen-Gazala, F.S. Neuman-Silberberg, Gurken, a TGF-alpha-like protein involved in axis determination in *Drosophila*, directly binds to the EGF-receptor homolog Egfr, Biochem. Biophys. Res. Commun. 291 (2002) 732–737.
[147] C. Ghiglione, E.A. Bach, Y. Paraiso, K.L. Carraway III, S. Noselli, N. Perrimon, Mechanism of activation of the *Drosophila* EGF Receptor by the TGFalpha ligand Gurken during oogenesis, Development 129 (2002) 175–186.
[148] K. Sakamoto, W.S. Chao, K. Katsube, A. Yamaguchi, Distinct roles of EGF repeats for the Notch signaling system, Exp. Cell Res. 302 (2005) 281–291.
[149] K. Somogyi, B. Sipos, Z. Penzes, I. Ando, A conserved gene cluster as a putative functional unit in insect innate immunity, FEBS Lett. 584 (2010) 4375–4378.
[150] E. Kurucz, R. Markus, J. Zsamboki, K. Folkl-Medzihradszky, Z. Darula, P. Vilmos, A. Udvardy, I. Krausz, T. Lukacsovich, E. Gateff, C.J. Zettervall, D. Hultmark, I. Ando, Nimrod, a putative phagocytosis receptor with EGF repeats in *Drosophila* plasmatocytes, Curr. Biol. 17 (2007) 649–654.
[151] C. Kocks, J.H. Cho, N. Nehme, J. Ulvila, A.M. Pearson, M. Meister, C. Strom, S.L. Conto, C. Hetru, L.M. Stuart, T. Stehle, J.A. Hoffmann, J.M. Reichhart, D. Ferrandon, M. Ramet, R.A. Ezekowitz, Eater, a transmembrane protein mediating phagocytosis of bacterial pathogens in *Drosophila*, Cell 123 (2005) 335–346.
[152] J.F. Fullard, N.E. Baker, Signaling by the engulfment receptor draper: a screen in *Drosophila melanogaster* implicates cytoskeletal regulators, Jun N-terminal Kinase, and Yorkie, Genetics 199 (2015) 117–134.
[153] J. Engel, EGF-like domains in extracellular matrix proteins: localized signals for growth and differentiation? FEBS Lett. 251 (1989) 1–7.
[154] J. Cordle, S. Johnson, J.Z. Tay, P. Roversi, M.B. Wilkin, B.H. de Madrid, H. Shimizu, S. Jensen, P. Whiteman, B. Jin, C. Redfield, M. Baron, S.M. Lea, P.A. Handford, A conserved face of the Jagged/Serrate DSL domain is involved in Notch trans-activation and cis-inhibition, Nat. Struct. Mol. Biol. 15 (2008) 849–857.
[155] T. Yamamoto, C.G. Davis, M.S. Brown, W.J. Schneider, M.L. Casey, J.L. Goldstein, D.W. Russell, The human LDL receptor: a cysteine-rich protein with multiple Alu sequences in its mRNA, Cell 39 (1984) 27–38.
[156] G. Rudenko, L. Henry, K. Henderson, K. Ichtchenko, M.S. Brown, J.L. Goldstein, J. Deisenhofer, Structure of the LDL receptor extracellular domain at endosomal pH, Science 298 (2002) 2353–2358.
[157] G. Xie, H. Zhang, G. Du, Q. Huang, X. Liang, J. Ma, R. Jiao, Uif, a large transmembrane protein with EGF-like repeats, can antagonize Notch signaling in *Drosophila*, PLoS One 7 (2012), e36362.
[158] S. Loubery, C. Seum, A. Moraleda, A. Daeden, M. Furthauer, M. Gonzalez-Gaitan, Uninflatable and Notch control the targeting of Sara endosomes during asymmetric division, Curr. Biol. 24 (2014) 2142–2148.
[159] Y. Xu, T. Wang, CULD is required for rhodopsin and TRPL channel endocytic trafficking and survival of photoreceptor cells, J. Cell Sci. 129 (2016) 394–405.
[160] Y.J. Kim, H. Bao, L. Bonanno, B. Zhang, M. Serpe, *Drosophila* Neto is essential for clustering glutamate receptors at the neuromuscular junction, Genes Dev. 26 (2012) 974–987.
[161] K.B. Reid, A.J. Day, Structure-function relationships of the complement components, Immunol. Today 10 (1989) 177–180.
[162] B.P. Lazzaro, Elevated polymorphism and divergence in the class C scavenger receptors of *Drosophila melanogaster* and D. simulans, Genetics 169 (2005) 2023–2034.
[163] D.A. Yadin, I.B. Robertson, J. McNaught-Davis, P. Evans, D. Stoddart, P.A. Handford, S.A. Jensen, C. Redfield, Structure of the fibrillin-1 N-terminal domains suggests that heparan sulfate regulates the early stages of microfibril assembly, Structure 21 (2013) 1743–1756.

[164] R. Garcia-Castellanos, N.S. Nielsen, K. Runager, I.B. Thogersen, M.V. Lukassen, E.T. Poulsen, T. Goulas, J.J. Enghild, F.X. Gomis-Ruth, Structural and functional implications of human transforming growth factor beta-induced protein, TGFBIp, in corneal dystrophies, Structure 25 (2017) 1740–1750.
[165] C.A. Innis, M. Hyvonen, Crystal structures of the heparan sulfate-binding domain of follistatin. Insights into ligand binding, J. Biol. Chem. 278 (2003) 39969–39977.
[166] D. Bickel, R. Shah, S.C. Gesualdi, T.E. Haerry, *Drosophila* Follistatin exhibits unique structural modifications and interacts with several TGF-beta family members, Mech. Dev. 125 (2008) 117–129.
[167] J. Shahab, C. Baratta, B. Scuric, D. Godt, K.J. Venken, M.J. Ringuette, Loss of SPARC dysregulates basal lamina assembly to disrupt larval fat body homeostasis in *Drosophila melanogaster*, Dev. Dyn. 244 (2015) 540–552.
[168] B. Kobe, J. Deisenhofer, The leucine-rich repeat: a versatile binding motif, Trends Biochem. Sci. 19 (1994) 415–421.
[169] H. Bilak, S. Tauszig-Delamasure, J.L. Imler, Toll and Toll-like receptors in *Drosophila*, Biochem. Soc. Trans. 31 (2003) 648–651.
[170] S. Valanne, J.H. Wang, M. Ramet, The *Drosophila* Toll signaling pathway, J. Immunol. 186 (2011) 649–656.
[171] K.V. Anderson, G. Jurgens, C. Nusslein-Volhard, Establishment of dorsal-ventral polarity in the *Drosophila* embryo: genetic studies on the role of the Toll gene product, Cell 42 (1985) 779–789.
[172] P. Qiu, P.C. Pan, S. Govind, A role for the *Drosophila* Toll/Cactus pathway in larval hematopoiesis, Development 125 (1998) 1909–1920.
[173] Y. Yagi, Y. Nishida, Y.T. Ip, Functional analysis of Toll-related genes in *Drosophila*, Develop. Growth Differ. 52 (2010) 771–783.
[174] C. Luo, B. Shen, J.L. Manley, L. Zheng, Tehao functions in the Toll pathway in *Drosophila melanogaster*: possible roles in development and innate immunity, Insect Mol. Biol. 10 (2001) 457–464.
[175] J.Y. Ooi, Y. Yagi, X. Hu, Y.T. Ip, The *Drosophila* Toll-9 activates a constitutive antimicrobial defense, EMBO Rep. 3 (2002) 82–87.
[176] M.J. Williams, A. Rodriguez, D.A. Kimbrell, E.D. Eldon, The 18-wheeler mutation reveals complex antibacterial gene regulation in *Drosophila* host defense, EMBO J. 16 (1997) 6120–6130.
[177] S. Tauszig, E. Jouanguy, J.A. Hoffmann, J.L. Imler, Toll-related receptors and the control of antimicrobial peptide expression in *Drosophila*, Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 10520–10525.
[178] D.M. Vallejo, S. Juarez-Carreno, J. Bolivar, J. Morante, M. Dominguez, A brain circuit that synchronizes growth and maturation revealed through Dilp8 binding to Lgr3, Science 350 (2015), aac6767.
[179] C.W. Ward, T.P. Garrett, The relationship between the L1 and L2 domains of the insulin and epidermal growth factor receptors and leucine-rich repeat modules, BMC Bioinformatics 2 (2001) 4.
[180] A. Nose, V.B. Mahajan, C.S. Goodman, Connectin: a homophilic cell adhesion molecule expressed on a subset of muscles and the motoneurons that innervate them in *Drosophila*, Cell 70 (1992) 553–567.
[181] S. Raghavan, R.A. White, Connectin mediates adhesion in *Drosophila*, Neuron 18 (1997) 873–880.
[182] Y. Mao, M. Kerr, M. Freeman, Modulation of *Drosophila* retinal epithelial integrity by the adhesion proteins capricious and tartan, PLoS One 3 (2008), e1827.
[183] C. Krause, C. Wolf, J. Hemphala, C. Samakovlis, R. Schuh, Distinct functions of the leucine-rich repeat transmembrane proteins capricious and tartan in the *Drosophila* tracheal morphogenesis, Dev. Biol. 296 (2006) 253–264.
[184] M. Milan, L. Perez, S.M. Cohen, Boundary formation in the *Drosophila* wing: functional dissection of Capricious and Tartan, Dev. Dyn. 233 (2005) 804–810.
[185] M. Milan, L. Perez, S.M. Cohen, Short-range cell interactions and cell survival in the *Drosophila* wing, Dev. Cell 2 (2002) 797–805.
[186] T. Adachi-Yamada, T. Harumoto, K. Sakurai, R. Ueda, K. Saigo, M.B. O'Connor, H. Nakato, Wing-to-Leg homeosis by spineless causes apoptosis regulated by Fish-lips, a novel leucine-rich repeat transmembrane protein, Mol. Cell. Biol. 25 (2005) 3140–3150.
[187] A. Nose, Generation of neuromuscular specificity in *Drosophila*: novel mechanisms revealed by new technologies, Front. Mol. Neurosci. 5 (2012) 62.
[188] M. Kurusu, A. Cording, M. Taniguchi, K. Menon, E. Suzuki, K. Zinn, A screen of cell-surface molecules identifies leucine-rich repeat proteins as key mediators of synaptic target selection, Neuron 59 (2008) 972–985.
[189] L.E. Swanson, M. Yu, K.S. Nelson, P. Laprise, U. Tepass, G.J. Beitel, *Drosophila* convoluted/dALS is an essential gene required for tracheal tube morphogenesis and apical matrix organization, Genetics 181 (2009) 1281–1290.
[190] D.E. Krantz, S.L. Zipursky, *Drosophila* chaoptin, a member of the leucine-rich repeat family, is a photoreceptor cell-specific adhesion molecule, EMBO J. 9 (1990) 1969–1977.
[191] B. Wayburn, T. Volk, LRT, a tendon-specific leucine-rich repeat protein, promotes muscle-tendon targeting through its interaction with Robo, Development 136 (2009) 3607–3615.
[192] M. Musacchio, N. Perrimon, The *Drosophila* kekkon genes: novel members of both the leucine-rich repeat and immunoglobulin superfamilies expressed in the CNS, Dev. Biol. 178 (1996) 63–76.
[193] C. Ghiglione, K.L. Carraway III, L.T. Amundadottir, R.E. Boswell, N. Perrimon, J.B. Duffy, The transmembrane molecule kekkon 1 acts in a feedback loop to negatively regulate the activity of the *Drosophila* EGF receptor during oogenesis, Cell 96 (1999) 847–856.
[194] T.A. Evans, H. Haridas, J.B. Duffy, Kekkon5 is an extracellular regulator of BMP signaling, Dev. Biol. 326 (2009) 36–46.
[195] M. Szuperak, S. Salah, E.J. Meyer, U. Nagarajan, A. Ikmi, M.C. Gibson, Feedback regulation of *Drosophila* BMP signaling by the novel extracellular protein larval translucida, Development 138 (2011) 715–724.
[196] W. Ren, Y. Zhang, M. Li, L. Wu, G. Wang, G.H. Baeg, J. You, Z. Li, X. Lin, Windpipe controls *Drosophila* intestinal homeostasis by regulating JAK/STAT pathway via promoting receptor endocytosis and lysosomal degradation, PLoS Genet. 11 (2015), e1005180. .
[197] J. Bella, K.L. Hindle, P.A. McEwan, S.C. Lovell, The leucine-rich repeat structure, Cell. Mol. Life Sci. 65 (2008) 2307–2333.
[198] L.C. Ferguson, J. Green, A. Surridge, C.D. Jiggins, Evolution of the insect yellow gene family, Mol. Biol. Evol. 28 (2011) 257–272.
[199] J. Schmitzova, J. Klaudiny, S. Albert, W. Schroder, W. Schreckengost, J. Hanes, J. Judova, J. Simuth, A family of

major royal jelly proteins of the honeybee Apis mellifera L, Cell. Mol. Life Sci. 54 (1998) 1020–1030.
[200] P.J. Wittkopp, J.R. True, S.B. Carroll, Reciprocal functions of the *Drosophila* yellow and ebony proteins in the development and evolution of pigment patterns, Development 129 (2002) 1849–1858.
[201] M.D. Drapeau, S.A. Cyran, M.M. Viering, P.K. Geyer, A.D. Long, A cis-regulatory sequence within the yellow locus of *Drosophila melanogaster* required for normal male mating success, Genetics 172 (2006) 1009–1030.
[202] K. Broadie, S. Baumgartner, A. Prokop, Extracellular matrix and its receptors in *Drosophila* neural development, Dev. Neurobiol. 71 (2011) 1102–1130.
[203] M. Rodriguez-Vazquez, D. Vaquero, E. Parra-Peralbo, J.E. Mejia-Morales, J. Culi, *Drosophila* lipophorin receptors recruit the lipoprotein LTP to the plasma membrane to mediate lipid uptake, PLoS Genet. 11 (2015), e1005356. .
[204] C.P. Schonbaum, S. Lee, A.P. Mahowald, The *Drosophila* yolkless gene encodes a vitellogenin receptor belonging to the low density lipoprotein receptor superfamily, Proc. Natl. Acad. Sci. U. S. A. 92 (1995) 1485–1489.
[205] L. Schweizer, H. Varmus, Wnt/Wingless signaling through beta-catenin requires the function of both LRP/Arrow and frizzled classes of receptors, BMC Cell Biol. 4 (2003) 4.
[206] F. Riedel, D. Vorkel, S. Eaton, Megalin-dependent yellow endocytosis restricts melanization in the *Drosophila* cuticle, Development 138 (2011) 149–158.
[207] T.A. Springer, Folding of the N-terminal, ligand-binding region of integrin alpha-subunits into a beta-propeller domain, Proc. Natl. Acad. Sci. U. S. A. 94 (1997) 65–72.
[208] U. Yazdani, J.R. Terman, The semaphorins, Genome Biol. 7 (2006) 211.
[209] P. Cafferty, L. Yu, H. Long, Y. Rao, Semaphorin-1a functions as a guidance receptor in the *Drosophila* visual system, J. Neurosci. 26 (2006) 3999–4003.
[210] M. Hernandez-Fleming, E.W. Rohrbach, G.J. Bashaw, Sema-1a reverse signaling promotes midline crossing in response to secreted semaphorins, Cell Rep. 18 (2017) 174–184.
[211] L. Yu, Y. Zhou, S. Cheng, Y. Rao, Plexin a-semaphorin-1a reverse signaling regulates photoreceptor axon guidance in *Drosophila*, J. Neurosci. 30 (2010) 12151–12156.
[212] G. Ventura, M. Furriols, N. Martin, V. Barbosa, J. Casanova, closca, a new gene required for both Torso RTK activation and vitelline membrane integrity. Germline proteins contribute to *Drosophila* eggshell composition, Dev. Biol. 344 (2010) 224–232.
[213] A. Degelmann, P.A. Hardy, A.P. Mahowald, Genetic analysis of two female-sterile loci affecting eggshell integrity and embryonic pattern formation in *Drosophila melanogaster*, Genetics 126 (1990) 427–434.
[214] E. Staub, J. Perez-Tur, R. Siebert, C. Nobile, N.K. Moschonas, P. Deloukas, B. Hinzmann, The novel EPTP repeat defines a superfamily of proteins implicated in epileptic disorders, Trends Biochem. Sci. 27 (2002) 441–444.
[215] J. Adams, R. Kelso, L. Cooley, The kelch repeat superfamily of proteins: propellers of cell function, Trends Cell Biol. 10 (2000) 17–24.
[216] T.J. Mosca, On the Teneurin track: a new synaptic organization molecule emerges, Front. Cell. Neurosci. 9 (2015) 204.
[217] C.L. Loughner, E.A. Bruford, M.S. McAndrews, E.E. Delp, S. Swamynathan, S.K. Swamynathan, Organization, evolution and functions of the human and mouse Ly6/uPAR family genes, Hum. Genomics 10 (2016) 10.
[218] A. Hijazi, W. Masson, B. Auge, L. Waltzer, M. Haenlin, F. Roch, boudin is required for septate junction organisation in *Drosophila* and codes for a diffusible protein of the Ly6 superfamily, Development 136 (2009) 2199–2209.
[219] A. Nilton, K. Oshima, F. Zare, S. Byri, U. Nannmark, K.G. Nyberg, R.G. Fehon, A.E. Uv, Crooked, coiled and crimpled are three Ly6-like proteins required for proper localization of septate junction components, Development 137 (2010) 2427–2437.
[220] K. Koh, W.J. Joiner, M.N. Wu, Z. Yue, C.J. Smith, A. Sehgal, Identification of SLEEPLESS, a sleep-promoting factor, Science 321 (2008) 372–376.
[221] M. Shi, Z. Yue, A. Kuryatov, J.M. Lindstrom, A. Sehgal, Identification of Redeye, a new sleep-regulating protein whose expression is modulated by sleep amount, elife 3 (2014), e01473. .
[222] B. Moussian, J. Soding, H. Schwarz, C. Nusslein-Volhard, Retroactive, a membrane-anchored extracellular protein related to vertebrate snake neurotoxin-like proteins, is required for cuticle organization in the larva of *Drosophila melanogaster*, Dev. Dyn. 233 (2005) 1056–1063.
[223] B. Moussian, E. Tang, A. Tonning, S. Helms, H. Schwarz, C. Nusslein-Volhard, A.E. Uv, *Drosophila* Knickkopf and Retroactive are needed for epithelial tube growth and cuticle differentiation through their specific requirement for chitin filament organization, Development 133 (2006) 163–171.
[224] M.E. Hemler, Tetraspanin functions and associated microdomains, Nat. Rev. Mol. Cell Biol. 6 (2005) 801–811.
[225] C.M. Termini, J.M. Gillette, Tetraspanins function as regulators of cellular signaling, Front. Cell Dev. Biol. 5 (2017) 34.
[226] X. Jia, L. Schulte, A. Loukas, D. Pickering, M. Pearson, M. Mobli, A. Jones, K.J. Rosengren, N.L. Daly, G.N. Gobert, M.K. Jones, D.J. Craik, J. Mulvenna, Solution structure, membrane interactions, and protein binding partners of the tetraspanin Sm-TSP-2, a vaccine antigen from the human blood fluke *Schistosoma mansoni*, J. Biol. Chem. 289 (2014) 7151–7163.
[227] C.S. Stipp, T.V. Kolesnikova, M.E. Hemler, Functional domains in tetraspanin proteins, Trends Biochem. Sci. 28 (2003) 106–112.
[228] C.C. Kopczynski, G.W. Davis, C.S. Goodman, A neural tetraspanin, encoded by late bloomer, that facilitates synapse formation, Science 271 (1996) 1867–1870.
[229] L.G. Fradkin, J.T. Kamphorst, A. DiAntonio, C.S. Goodman, J.N. Noordermeer, Genomewide analysis of the *Drosophila* tetraspanins reveals a subset with similar function in the formation of the embryonic synapse, Proc. Natl. Acad. Sci. U. S. A. 99 (2002) 13663–13668.
[230] Y. Izumi, M. Motoishi, K. Furuse, M. Furuse, A tetraspanin regulates septate junction formation in *Drosophila* midgut, J. Cell Sci. 129 (2016) 1155–1164.
[231] H. Xu, S.J. Lee, E. Suzuki, K.D. Dugan, A. Stoddard, H.S. Li, L.A. Chodosh, C. Montell, A lysosomal tetraspanin associated with retinal degeneration identified via a genome-wide screen, EMBO J. 23 (2004) 811–822.
[232] E. Dornier, F. Coumailleau, J.F. Ottavi, J. Moretti, C. Boucheix, P. Mauduit, F. Schweisguth, E. Rubinstein, TspanC8 tetraspanins regulate ADAM10/Kuzbanian trafficking and promote Notch activation in flies and mammals, J. Cell Biol. 199 (2012) 481–496.
[233] S. Grunder, X. Chen, Structure, function, and pharmacology of acid-sensing ion channels (ASICs): focus on ASIC1a, Int. J. Physiol. Pathophysiol. Pharmacol. 2 (2010) 73–94.

[234] K.M. Zelle, B. Lu, S.C. Pyfrom, Y. Ben-Shahar, The genetic architecture of degenerin/epithelial sodium channels in *Drosophila*, G3 (Bethesda) 3 (2013) 441–450.
[235] Y.J. Lin, L. Seroude, S. Benzer, Extended life-span and stress resistance in the *Drosophila* mutant methuselah, Science 282 (1998) 943–946.
[236] A. Petrosyan, O.F. Goncalves, I.H. Hsieh, K. Saberi, Improved functional abilities of the life-extended *Drosophila* mutant Methuselah are reversed at old age to below control levels, Age (Dordr.) 36 (2014) 213–221.
[237] A. de Mendoza, J.W. Jones, M. Friedrich, Methuselah/Methuselah-like G protein-coupled receptors constitute an ancient metazoan gene family, Sci. Rep. 6 (2016) 21801.
[238] Y.K. Xu, R. Nusse, The Frizzled CRD domain is conserved in diverse proteins including several receptor tyrosine kinases, Curr. Biol. 8 (1998) R405–R406.
[239] J. Pei, N.V. Grishin, Cysteine-rich domains related to Frizzled receptors and Hedgehog-interacting proteins, Protein Sci. 21 (2012) 1172–1184.
[240] J. Filmus, M. Capurro, J. Rast, Glypicans, Genome Biol. 9 (2008) 224.
[241] A. Ghezzi, B.J. Liebeskind, A. Thompson, N.S. Atkinson, H. H. Zakon, Ancient association between cation leak channels and Mid1 proteins is conserved in fungi and animals, Front. Mol. Neurosci. 7 (2014) 15.
[242] P. Bork, G. Beckmann, The CUB domain. A widespread module in developmentally regulated proteins, J. Mol. Biol. 231 (1993) 539–545.
[243] B.A. Appleton, P. Wu, J. Maloney, J. Yin, W.C. Liang, S. Stawicki, K. Mortara, K.K. Bowman, J.M. Elliott, W. Desmarais, J.F. Bazan, A. Bagri, M. Tessier-Lavigne, A.W. Koch, Y. Wu, R.J. Watts, C. Wiesmann, Structural studies of neuropilin/antibody complexes provide insights into semaphorin and VEGF binding, EMBO J. 26 (2007) 4902–4912.
[244] A.F. Seasholtz, R.A. Valverde, R.J. Denver, Corticotropin-releasing hormone-binding protein: biochemistry and function from fishes to mammals, J. Endocrinol. 175 (2002) 89–97.
[245] G.M. Gibbs, K. Roelants, M.K. O'Bryan, The CAP superfamily: cysteine-rich secretory proteins, antigen 5, and pathogenesis-related 1 proteins—roles in reproduction, cancer, and immune defense, Endocr. Rev. 29 (2008) 865–897.
[246] X. Xu, I.M. Francischetti, R. Lai, J.M. Ribeiro, J.F. Andersen, Structure of protein having inhibitory disintegrin and leukotriene scavenging functions contained in single domain, J. Biol. Chem. 287 (2012) 10967–10976.
[247] R. Darwiche, L. Mene-Saffrane, D. Gfeller, O.A. Asojo, R. Schneiter, The pathogen-related yeast protein Pry1, a member of the CAP protein superfamily, is a fatty acid-binding protein, J. Biol. Chem. 292 (2017) 8304–8314.
[248] G.E. Kovalick, D.L. Griffin, Characterization of the SCP/TAPS gene family in *Drosophila melanogaster*, Insect Biochem. Mol. Biol. 35 (2005) 825–835.
[249] T. Moran, Y. Gat, D. Fass, Laminin L4 domain structure resembles adhesion modules in ephrin receptor and other transmembrane glycoproteins, FEBS J. 282 (2015) 2746–2757.
[250] S.A. Hussain, F. Carafoli, E. Hohenester, Determinants of laminin polymerization revealed by the structure of the alpha5 chain amino-terminal region, EMBO Rep. 12 (2011) 276–282.
[251] R. Timpl, D. Tisi, J.F. Talts, Z. Andac, T. Sasaki, E. Hohenester, Structure and function of laminin LG modules, Matrix Biol. 19 (2000) 309–317.
[252] A.L. Fidler, C.E. Darris, S.V. Chetyrkin, V.K. Pedchenko, S.P. Boudko, K.L. Brown, W. Gray Jerome, J.K. Hudson, A. Rokas, B.G. Hudson, Collagen IV and basement membrane at the evolutionary dawn of metazoan tissues, elife 6 (2017).
[253] F. Meyer, B. Moussian, *Drosophila* multiplexin (Dmp) modulates motor axon pathfinding accuracy, Develop. Growth Differ. 51 (2009) 483–498.
[254] R. Momota, M. Narasaki, T. Komiyama, I. Naito, Y. Ninomiya, A. Ohtsuka, *Drosophila* type XV/XVIII collagen mutants manifest integrin mediated mitochondrial dysfunction, which is improved by cyclosporin A and losartan, Int. J. Biochem. Cell Biol. 45 (2013) 1003–1011.
[255] C.R. Warren, E. Kassir, J. Spurlin, J. Martinez, N.H. Putnam, M.C. Farach-Carson, Evolution of the perlecan/HSPG2 gene and its activation in regenerating Nematostella vectensis, PLoS One 10 (2015), e0124578. .
[256] X. Lin, Functions of heparan sulfate proteoglycans in cell signaling during development, Development 131 (2004) 6009–6021.
[257] L.H. Margaritis, F.C. Kafatos, W.H. Petri, The eggshell of *Drosophila melanogaster*. I. Fine structure of the layers and regions of the wild-type eggshell, J. Cell Sci. 43 (1980) 1–35.
[258] T. Pascucci, J. Perrino, A.P. Mahowald, G.L. Waring, Eggshell assembly in *Drosophila*: processing and localization of vitelline membrane and chorion proteins, Dev. Biol. 177 (1996) 590–598.
[259] G.L. Waring, Morphogenesis of the eggshell in *Drosophila*, Int. Rev. Cytol. 198 (2000) 67–108.
[260] T. Wu, A.L. Manogaran, J.M. Beauchamp, G.L. Waring, *Drosophila* vitelline membrane assembly: a critical role for an evolutionarily conserved cysteine in the “VM domain” of sV23, Dev. Biol. 347 (2010) 360–368.
[261] M. Elalayli, J.D. Hall, M. Fakhouri, H. Neiswender, T.T. Ellison, Z. Han, P. Roon, E.K. LeMosy, Palisade is required in the *Drosophila* ovary for assembly and function of the protective vitelline membrane, Dev. Biol. 319 (2008) 359–369.
[262] Y.N. Osheim, O.L. Miller Jr., A.L. Beyer, Two *Drosophila* chorion genes terminate transcription in discrete regions near their poly(A) sites, EMBO J. 5 (1986) 3591–3596.
[263] G.L. Waring, A.P. Mahowald, Identification and time of synthesis of chorion proteins in *Drosophila melanogaster*, Cell 16 (1979) 599–607.
[264] A. Parks, A. Spradling, Spatially regulated expression of chorion genes during *Drosophila* oogenesis, Genes Dev. 1 (1987) 497–509.
[265] G.L. Waring, R.J. Hawley, T. Schoenfeld, Multiple proteins are produced from the dec-1 eggshell gene in *Drosophila* by alternative RNA splicing and proteolytic cleavage events, Dev. Biol. 142 (1990) 1–12.
[266] M. Fakhouri, M. Elalayli, D. Sherling, J.D. Hall, E. Miller, X. Sun, L. Wells, E.K. LeMosy, Minor proteins and enzymes of the *Drosophila* eggshell matrix, Dev. Biol. 293 (2006) 127–141.
[267] S. Parks, B. Wakimoto, A. Spradling, Replication and expression of an X-linked cluster of *Drosophila* chorion genes, Dev. Biol. 117 (1986) 294–305.
[268] T.L. Tootle, D. Williams, A. Hubb, R. Frederick, A. Spradling, *Drosophila* eggshell production: identification of new genes and coordination by Pxt, PLoS One 6 (2011), e19943.
[269] D.E. Klein, S.E. Stayrook, F. Shi, K. Narayan, M.A. Lemmon, Structural basis for EGFR ligand sequestration by Argos, Nature 453 (2008) 1271–1275.
[270] M. Oelgeschlager, J. Larrain, D. Geissert, E.M. De Robertis, The evolutionarily conserved BMP-binding protein Twisted

gastrulation promotes BMP signalling, Nature 405 (2000) 757–763.
[271] T. Malinauskas, A.R. Aricescu, W. Lu, C. Siebold, E.Y. Jones, Modular mechanism of Wnt signaling inhibition by Wnt inhibitory factor 1, Nat. Struct. Mol. Biol. 18 (2011) 886–893.
[272] A. Avanesov, S.M. Honeyager, J. Malicki, S.S. Blair, The role of glypicans in Wnt inhibitory factor-1 activity and the structural basis of Wif1's effects on Wnt and Hedgehog signaling, PLoS Genet. 8 (2012), e1002503.
[273] J.P. Burbach, What are neuropeptides? Methods Mol. Biol. 789 (2011) 1–36.
[274] P.D. Sun, D.R. Davies, The cystine-knot growth-factor superfamily, Annu. Rev. Biophys. Biomol. Struct. 24 (1995) 269–291.
[275] N.L. Daly, D.J. Craik, Bioactive cystine knot proteins, Curr. Opin. Chem. Biol. 15 (2011) 362–368.
[276] A. Upadhyay, L. Moss-Taylor, M.J. Kim, A.C. Ghosh, M.B. O'Connor, TGF-beta family signaling in *Drosophila*, Cold Spring Harb. Perspect. Biol. 9 (2017).
[277] A.J. Peterson, M.B. O'Connor, Strategies for exploring TGF-beta signaling in *Drosophila*, Methods 68 (2014) 183–193.
[278] K.E. Harris, N. Schnittke, S.K. Beckendorf, Two ligands signal through the *Drosophila* PDGF/VEGF receptor to ensure proper salivary gland positioning, Mech. Dev. 124 (2007) 441–448.
[279] J.A. McDonald, E.M. Pinheiro, D.J. Montell, PVF1, a PDGF/VEGF homolog, is sufficient to guide border cells and interacts genetically with Taiman, Development 130 (2003) 3469–3478.
[280] N.K. Cho, L. Keyes, E. Johnson, J. Heller, L. Ryner, F. Karim, M.A. Krasnow, Developmental control of blood cell migration by the *Drosophila* VEGF pathway, Cell 108 (2002) 865–876.
[281] K.F. Rewitz, N. Yamanaka, L.I. Gilbert, M.B. O'Connor, The insect neuropeptide PTTH activates receptor tyrosine kinase torso to initiate metamorphosis, Science 326 (2009) 1403–1405.
[282] Z. McBrayer, H. Ono, M. Shimell, J.P. Parvy, R.B. Beckstead, J.T. Warren, C.S. Thummel, C. Dauphin-Villemant, L.I. Gilbert, M.B. O'Connor, Prothoracicotropic hormone regulates developmental timing and body size in *Drosophila*, Dev. Cell 13 (2007) 857–871.
[283] S. Kadam, A. McMahon, P. Tzou, A. Stathopoulos, FGF ligands in *Drosophila* have distinct activities required to support cell migration and differentiation, Development 136 (2009) 739–747.
[284] J. Schlessinger, A.N. Plotnikov, O.A. Ibrahimi, A.V. Eliseenkova, B.K. Yeh, A. Yayon, R.J. Linhardt, M. Mohammadi, Crystal structure of a ternary FGF–FGFR–heparin complex reveals a dual role for heparin in FGFR binding and dimerization, Mol. Cell 6 (2000) 743–750.
[285] S. Swarup, E.M. Verheyen, Wnt/Wingless signaling in *Drosophila*, Cold Spring Harb. Perspect. Biol. 4 (2012).
[286] K.M. Kozopas, C.H. Samos, R. Nusse, DWnt-2, a *Drosophila* Wnt gene required for the development of the male reproductive tract, specifies a sexually dimorphic cell fate, Genes Dev. 12 (1998) 1155–1165.
[287] K.M. Kozopas, R. Nusse, Direct flight muscles in *Drosophila* develop from cells with characteristics of founders and depend on DWnt-2 for their correct patterning, Dev. Biol. 243 (2002) 312–325.
[288] E.D. Cohen, M.C. Mariol, R.M. Wallace, J. Weyers, Y.G. Kamberov, J. Pradel, E.L. Wilder, DWnt4 regulates cell movement and focal adhesion kinase during *Drosophila* ovarian morphogenesis, Dev. Cell 2 (2002) 437–448.
[289] M. Sato, D. Umetsu, S. Murakami, T. Yasugi, T. Tabata, DWnt4 regulates the dorsoventral specificity of retinal projections in the *Drosophila melanogaster* visual system, Nat. Neurosci. 9 (2006) 67–75.
[290] M. Inaki, S. Yoshikawa, J.B. Thomas, H. Aburatani, A. Nose, Wnt4 is a local repulsive cue that determines synaptic target specificity, Curr. Biol. 17 (2007) 1574–1579.
[291] S. Yoshikawa, R.D. McKinnon, M. Kokel, J.B. Thomas, Wnt-mediated axon guidance via the *Drosophila* Derailed receptor, Nature 422 (2003) 583–588.
[292] N. Doumpas, G. Jekely, A.A. Teleman, Wnt6 is required for maxillary palp formation in *Drosophila*, BMC Biol. 11 (2013) 104.
[293] K. Janson, E.D. Cohen, E.L. Wilder, Expression of DWnt6, DWnt10, and DFz4 during *Drosophila* development, Mech. Dev. 103 (2001) 117–120.
[294] M.D. Gordon, M.S. Dionne, D.S. Schneider, R. Nusse, WntD is a feedback inhibitor of Dorsal/NF-kappaB in *Drosophila* development and immunity, Nature 437 (2005) 746–749.
[295] A. Ganguly, J. Jiang, Y.T. Ip, *Drosophila* WntD is a target and an inhibitor of the Dorsal/Twist/Snail network in the gastrulating embryo, Development 132 (2005) 3419–3429.
[296] O. Lamiable, C. Meignin, J.L. Imler, WntD and Diedel: two immunomodulatory cytokines in *Drosophila* immunity, Fly (Austin) 10 (2016) 187–194.
[297] T.R. Burglin, The Hedgehog protein family, Genome Biol. 9 (2008) 241.
[298] F. Balordi, G. Fishell, Hedgehog signaling in the subventricular zone is required for both the maintenance of stem cells and the migration of newborn neurons, J. Neurosci. 27 (2007) 5936–5947.
[299] X. Chen, H. Tukachinsky, C.H. Huang, C. Jao, Y.R. Chu, H.Y. Tang, B. Mueller, S. Schulman, T.A. Rapoport, A. Salic, Processing and turnover of the Hedgehog protein in the endoplasmic reticulum, J. Cell Biol. 192 (2011) 825–838.
[300] T.M. Hall, J.A. Porter, P.A. Beachy, D.J. Leahy, A potential catalytic site revealed by the 1.7-A crystal structure of the amino-terminal signalling domain of Sonic hedgehog, Nature 378 (1995) 212–216.
[301] M. Costa, E.T. Wilson, E. Wieschaus, A putative cell signal encoded by the folded gastrulation gene coordinates cell shape changes during *Drosophila* gastrulation, Cell 76 (1994) 1075–1089.
[302] A.J. Manning, K.A. Peters, M. Peifer, S.L. Rogers, Regulation of epithelial morphogenesis by the G protein-coupled receptor mist and its ligand fog. *Sci Signal* 6, ra98, 2013.
[303] A. Ratnaparkhi, K. Zinn, The secreted cell signal Folded Gastrulation regulates glial morphogenesis and axon guidance in *Drosophila*, Dev. Biol. 308 (2007) 158–168.
[304] E.R. Andersson, R. Sandberg, U. Lendahl, Notch signaling: simplicity in design, versatility in function, Development 138 (2011) 3593–3612.
[305] A. Kohyama-Koganeya, Y.J. Kim, M. Miura, Y. Hirabayashi, A *Drosophila* orphan G protein-coupled receptor BOSS functions as a glucose-responding receptor: loss of boss causes abnormal energy metabolism, Proc. Natl. Acad. Sci. U. S. A. 105 (2008) 15328–15333.
[306] M. Matis, J.D. Axelrod, Regulation of PCP by the Fat signaling pathway, Genes Dev. 27 (2013) 2207–2220.
[307] T. Bossing, A.H. Brand, Dephrin, a transmembrane ephrin with a unique structure, prevents interneuronal axons from exiting the *Drosophila* embryonic CNS, Development 129 (2002) 4205–4218.

[308] S. Chakrabarti, J.P. Dudzic, X. Li, E.J. Collas, J.P. Boquete, B. Lemaitre, Remote control of intestinal stem cell activity by haemocytes in *Drosophila*, PLoS Genet. 12 (2016), e1006089.
[309] V.M. Wright, K.L. Vogt, E. Smythe, M.P. Zeidler, Differential activities of the *Drosophila* JAK/STAT pathway ligands Upd, Upd2 and Upd3, Cell. Signal. 23 (2011) 920–927.
[310] D.A. Harrison, P.E. McCoon, R. Binari, M. Gilman, N. Perrimon, *Drosophila* unpaired encodes a secreted protein that activates the JAK signaling pathway, Genes Dev. 12 (1998) 3252–3263.
[311] T. Igaki, H. Kanda, Y. Yamamoto-Goto, H. Kanuka, E. Kuranaga, T. Aigaki, M. Miura, Eiger, a TNF superfamily ligand that triggers the *Drosophila* JNK pathway, EMBO J. 21 (2002) 3009–3018.
[312] S. Kauppila, W.S. Maaty, P. Chen, R.S. Tomar, M.T. Eby, J. Chapo, S. Chew, N. Rathore, S. Zachariah, S.K. Sinha, J.M. Abrams, P.M. Chaudhary, Eiger and its receptor, Wengen, comprise a TNF-like system in *Drosophila*, Oncogene 22 (2003) 4860–4867.
[313] D.S. Andersen, J. Colombani, V. Palmerini, K. Chakrabandhu, E. Boone, M. Rothlisberger, J. Toggweiler, K. Basler, M. Mapelli, A.O. Hueber, P. Leopold, The *Drosophila* TNF receptor Grindelwald couples loss of cell polarity and neoplastic growth, Nature 522 (2015) 482–486.
[314] N. Agrawal, R. Delanoue, A. Mauri, D. Basco, M. Pasco, B. Thorens, P. Leopold, The *Drosophila* TNF eiger is an adipokine that acts on insulin-producing cells to mediate nutrient response, Cell Metab. 23 (2016) 675–684.
[315] F. Coste, C. Kemp, V. Bobezeau, C. Hetru, C. Kellenberger, J.L. Imler, A. Roussel, Crystal structure of Diedel, a marker of the immune response of *Drosophila melanogaster*, PLoS One 7 (2012), e33416.
[316] M. Boutros, H. Agaisse, N. Perrimon, Sequential activation of signaling pathways during innate immune responses in *Drosophila*, Dev. Cell 3 (2002) 711–722.
[317] O. Lamiable, C. Kellenberger, C. Kemp, L. Troxler, N. Pelte, M. Boutros, J.T. Marques, L. Daeffler, J.A. Hoffmann, A. Roussel, J.L. Imler, Cytokine Diedel and a viral homologue suppress the IMD pathway in *Drosophila*, Proc. Natl. Acad. Sci. U. S. A. 113 (2016) 698–703.
[318] J.S. Parker, K. Mizuguchi, N.J. Gay, A family of proteins related to Spatzle, the toll receptor ligand, are encoded in the *Drosophila* genome, Proteins 45 (2001) 71–80.
[319] T.J. Sheldon, I. Miguel-Aliaga, A.P. Gould, W.R. Taylor, D. Conklin, A novel family of single VWC-domain proteins in invertebrates, FEBS Lett. 581 (2007) 5268–5274.
[320] S. Deddouche, N. Matt, A. Budd, S. Mueller, C. Kemp, D. Galiana-Arnoux, C. Dostert, C. Antoniewski, J.A. Hoffmann, J.L. Imler, The DExD/H-box helicase Dicer-2 mediates the induction of antiviral activity in *Drosophila*, Nat. Immunol. 9 (2008) 1425–1432.
[321] P.N. Paradkar, L. Trinidad, R. Voysey, J.B. Duchemin, P.J. Walker, Secreted Vago restricts West Nile virus infection in Culex mosquito cells by activating the Jak–STAT pathway, Proc. Natl. Acad. Sci. U. S. A. 109 (2012) 18915–18920.
[322] S. Tsuzuki, M. Ochiai, H. Matsumoto, S. Kurata, A. Ohnishi, Y. Hayakawa, *Drosophila* growth-blocking peptide-like factor mediates acute immune reactions during infectious and non-infectious stress, Sci. Rep. 2 (2012) 210.
[323] H. Matsumoto, S. Tsuzuki, A. Date-Ito, A. Ohnishi, Y. Hayakawa, Characteristics common to a cytokine family spanning five orders of insects, Insect Biochem. Mol. Biol. 42 (2012) 446–454.
[324] T. Koyama, C.K. Mirth, Growth-blocking peptides as nutrition-sensitive signals for insulin secretion and body size regulation, PLoS Biol. 14 (2016), e1002392.
[325] A. Ohnishi, Y. Oda, Y. Hayakawa, Characterization of receptors of insect cytokine, growth-blocking peptide, in human keratinocyte and insect Sf9 cells, J. Biol. Chem. 276 (2001) 37974–37979.
[326] E.J. Sung, M. Ryuda, H. Matsumoto, O. Uryu, M. Ochiai, M. E. Cook, N.Y. Yi, H. Wang, J.W. Putney, G.S. Bird, S.B. Shears, Y. Hayakawa, Cytokine signaling through *Drosophila* Mthl10 ties lifespan to environmental stress, Proc. Natl. Acad. Sci. U. S. A. 114 (2017) 13786–13791.
[327] R.S. Hewes, P.H. Taghert, Neuropeptides and neuropeptide receptors in the *Drosophila melanogaster* genome, Genome Res. 11 (2001) 1126–1142.
[328] K. Kannan, Y.W. Fridell, Functional implications of *Drosophila* insulin-like peptides in metabolism, aging, and dietary restriction, Front. Physiol. 4 (2013) 288.
[329] A. Garelli, A.M. Gontijo, V. Miguela, E. Caparros, M. Dominguez, Imaginal discs secrete insulin-like peptide 8 to mediate plasticity of growth and maturation, Science 336 (2012) 579–582.
[330] A. Garelli, F. Heredia, A.P. Casimiro, A. Macedo, C. Nunes, M. Garcez, A.R. Dias, Y.A. Volonte, T. Uhlmann, E. Caparros, T. Koyama, A.M. Gontijo, Dilp8 requires the neuronal relaxin receptor Lgr3 to couple growth to developmental timing, Nat. Commun. 6 (2015) 8732.
[331] C.W. Luo, E.M. Dewey, S. Sudo, J. Ewer, S.Y. Hsu, H.W. Honegger, A.J. Hsueh, Bursicon, the insect cuticle-hardening hormone, is a heterodimeric cystine knot protein that activates G protein-coupled receptor LGR2, Proc. Natl. Acad. Sci. U. S. A. 102 (2005) 2820–2825.
[332] A. Sellami, H.J. Agricola, J.A. Veenstra, Neuroendocrine cells in *Drosophila melanogaster* producing GPA2/GPB5, a hormone with homology to LH, FSH and TSH, Gen. Comp. Endocrinol. 170 (2011) 582–588.
[333] S. Sudo, Y. Kuwabara, J.I. Park, S.Y. Hsu, A.J. Hsueh, Heterodimeric fly glycoprotein hormone-alpha2 (GPA2) and glycoprotein hormone-beta5 (GPB5) activate fly leucine-rich repeat-containing G protein-coupled receptor-1 (DLGR1) and stimulation of human thyrotropin receptors by chimeric fly GPA2 and human GPB5, Endocrinology 146 (2005) 3596–3604.
[334] M. Galikova, M. Diesner, P. Klepsatel, P. Hehlert, Y. Xu, I. Bickmeyer, R. Predel, R.P. Kuhnlein, Energy homeostasis control in *Drosophila* adipokinetic hormone mutants, Genetics 201 (2015) 665–683.
[335] M.B. Blackburn, R.M. Wagner, J.P. Kochansky, D.J. Harrison, P. Thomas-Laemont, A.K. Raina, The identification of two myoinhibitory peptides, with sequence similarities to the galanins, isolated from the ventral nerve cord of *Manduca sexta*, Regul. Pept. 57 (1995) 213–219.
[336] N. Shibusawa, K. Hashimoto, M. Yamada, Thyrotropin-releasing hormone (TRH) in the cerebellum, Cerebellum 7 (2008) 84–95.
[337] F. Hauser, C.J. Grimmelikhuijzen, Evolution of the AKH/corazonin/ACP/GnRH receptor superfamily and their ligands in the *Protostomia*, Gen. Comp. Endocrinol. 209 (2014) 35–49.
[338] W. Liu, F. Guo, B. Lu, A. Guo, Amnesiac regulates sleep onset and maintenance in *Drosophila melanogaster*, Biochem. Biophys. Res. Commun. 372 (2008) 798–803.
[339] S. Waddell, J.D. Armstrong, T. Kitamoto, K. Kaiser, W.G. Quinn, The amnesiac gene product is expressed in two

neurons in the *Drosophila* brain that are critical for memory, Cell 103 (2000) 805–813.
[340] H. Hashimoto, N. Shintani, A. Baba, Higher brain functions of PACAP and a homologous *Drosophila* memory gene amnesiac: insights from knockouts and mutants, Biochem. Biophys. Res. Commun. 297 (2002) 427–431.
[341] J.L. Hentze, M.A. Carlsson, S. Kondo, D.R. Nassel, K.F. Rewitz, The neuropeptide allatostatin A regulates metabolism and feeding decisions in *Drosophila*, Sci. Rep. 5 (2015) 11680.
[342] J.A. Veenstra, Allatostatin C and its paralog allatostatin double C: the arthropod somatostatins, Insect Biochem. Mol. Biol. 39 (2009) 161–170.
[343] D.R. Nassel, A.M. Winther, *Drosophila* neuropeptides in regulation of physiology and behavior, Prog. Neurobiol. 92 (2010) 42–104.
[344] C. Wegener, Z. Herbert, M. Eckert, R. Predel, The periviscerokinin (PVK) peptide family in insects: evidence for the inclusion of CAP(2b) as a PVK family member, Peptides 23 (2002) 605–611.
[345] X. Meng, G. Wahlstrom, T. Immonen, M. Kolmer, M. Tirronen, R. Predel, N. Kalkkinen, T.I. Heino, H. Sariola, C. Roos, The *Drosophila* hugin gene codes for myostimulatory and ecdysis-modifying neuropeptides, Mech. Dev. 117 (2002) 5–13.
[346] G.E. Jackson, A.N. Mabula, S.R. Stone, G. Gade, K.E. Kover, L. Szilagyi, D. van der Spoel, Solution conformations of an insect neuropeptide: crustacean cardioactive peptide (CCAP), Peptides 30 (2009) 557–564.
[347] G.R. Ren, F. Hauser, K.F. Rewitz, S. Kondo, A.F. Engelbrecht, A.K. Didriksen, S.R. Schjott, F.E. Sembach, S. Li, K.C. Sogaard, L. Sondergaard, C.J. Grimmelikhuijzen, CCHamide-2 is an orexigenic brain–gut peptide in *Drosophila*, PLoS One 10 (2015), e0133017. .
[348] S.H. Jung, J.H. Lee, H.S. Chae, J.Y. Seong, Y. Park, Z.Y. Park, Y.J. Kim, Identification of a novel insect neuropeptide, CNMa and its receptor, FEBS Lett. 588 (2014) 2037–2041.
[349] K. Furuya, R.J. Milchak, K.M. Schegg, J. Zhang, S.S. Tobe, G.M. Coast, D.A. Schooley, Cockroach diuretic hormones: characterization of a calcitonin-like peptide in insects, Proc. Natl. Acad. Sci. U. S. A. 97 (2000) 6469–6474.
[350] F.M. Horodyski, J. Ewer, L.M. Riddiford, J.W. Truman, Isolation, characterization and expression of the eclosion hormone gene of *Drosophila melanogaster*, Eur. J. Biochem. 215 (1993) 221–228.
[351] J.C. Chang, R.B. Yang, M.E. Adams, K.H. Lu, Receptor guanylyl cyclases in Inka cells targeted by eclosion hormone, Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 13371–13376.
[352] H. Dircksen, Insect ion transport peptides are derived from alternatively spliced genes and differentially expressed in the central and peripheral nervous system, J. Exp. Biol. 212 (2009) 401–412.
[353] H. Katayama, K. Nagata, T. Ohira, F. Yumoto, M. Tanokura, H. Nagasawa, The solution structure of molt-inhibiting hormone from the Kuruma prawn *Marsupenaeus japonicus*, J. Biol. Chem. 278 (2003) 9620–9623.
[354] N. Tsutsui, T. Sakamoto, F. Arisaka, M. Tanokura, H. Nagasawa, K. Nagata, Crystal structure of a crustacean hyperglycemic hormone (CHH) precursor suggests structural variety in the C-terminal regions of CHH superfamily members, FEBS J. 283 (2016) 4325–4339.
[355] C. Hermann-Luibl, T. Yoshii, P.R. Senthilan, H. Dircksen, C. Helfrich-Forster, The ion transport peptide is a new functional clock neuropeptide in the fruit fly *Drosophila melanogaster*, J. Neurosci. 34 (2014) 9522–9536.
[356] B. Al-Anzi, E. Armand, P. Nagamei, M. Olszewski, V. Sapin, C. Waters, K. Zinn, R.J. Wyman, S. Benzer, The leucokinin pathway and its neurons regulate meal size in *Drosophila*, Curr. Biol. 20 (2010) 969–978.
[357] T. Ida, T. Takahashi, H. Tominaga, T. Sato, K. Kume, M. Ozaki, T. Hiraguchi, T. Maeda, H. Shiotani, S. Terajima, H. Sano, K. Mori, M. Yoshida, M. Miyazato, J. Kato, N. Murakami, K. Kangawa, M. Kojima, Identification of the novel bioactive peptides dRYamide-1 and dRYamide-2, ligands for a neuropeptide Y-like receptor in *Drosophila*, Biochem. Biophys. Res. Commun. 410 (2011) 872–877.
[358] C. Collin, F. Hauser, P. Krogh-Meyer, K.K. Hansen, E. Gonzalez de Valdivia, M. Williamson, C.J. Grimmelikhuijzen, Identification of the *Drosophila* and *Tribolium* receptors for the recently discovered insect RYamide neuropeptides, Biochem. Biophys. Res. Commun. 412 (2011) 578–583.
[359] G. Baggerman, A. Cerstiaens, A. De Loof, L. Schoofs, Peptidomics of the larval *Drosophila melanogaster* central nervous system, J. Biol. Chem. 277 (2002) 40368–40374.
[360] G. Baggerman, K. Boonen, P. Verleyen, A. De Loof, L. Schoofs, Peptidomic analysis of the larval *Drosophila melanogaster* central nervous system by two-dimensional capillary liquid chromatography quadrupole time-of-flight mass spectrometry, J. Mass Spectrom. 40 (2005) 250–260.
[361] G. Overend, P. Cabrero, A.X. Guo, S. Sebastian, M. Cundall, H. Armstrong, I. Mertens, L. Schoofs, J.A. Dow, S.A. Davies, The receptor guanylate cyclase Gyc76C and a peptide ligand, NPLP1-VQQ, modulate the innate immune IMD pathway in response to salt stress, Peptides 34 (2012) 209–218.
[362] J. Chen, M.S. Choi, A. Mizoguchi, J.A. Veenstra, K. Kang, Y.J. Kim, J.Y. Kwon, Isoform-specific expression of the neuropeptide orcokinin in *Drosophila melanogaster*, Peptides 68 (2015) 50–57.
[363] A. Seluzicki, M. Flourakis, E. Kula-Eversole, L. Zhang, V. Kilman, R. Allada, Dual PDF signaling pathways reset clocks via TIMELESS and acutely excite target neurons to control circadian behavior, PLoS Biol. 12 (2014), e1001810.
[364] I. Janssen, L. Schoofs, K. Spittaels, H. Neven, J. Vanden Broeck, B. Devreese, J. Van Beeumen, J. Shabanowitz, D.F. Hunt, A. De Loof, Isolation of NEB-LFamide, a novel myotropic neuropeptide from the grey fleshfly, Mol. Cell. Endocrinol. 117 (1996) 157–165.
[365] S. Park, J.Y. Sonn, Y. Oh, C. Lim, J. Choe, SIFamide and SIFamide receptor defines a novel neuropeptide signaling to promote sleep in *Drosophila*, Mol. Cell 37 (2014) 295–301.
[366] K. Asahina, K. Watanabe, B.J. Duistermars, E. Hoopfer, C. R. Gonzalez, E.A. Eyjolfsdottir, P. Perona, D.J. Anderson, Tachykinin-expressing neurons control male-specific aggressive arousal in *Drosophila*, Cell 156 (2014) 221–235.
[367] H. Jiang, A. Lkhagva, I. Daubnerova, H.S. Chae, L. Simo, S.H. Jung, Y.K. Yoon, N.R. Lee, J.Y. Seong, D. Zitnan, Y. Park, Y.J. Kim, Natalisin, a tachykinin-like signaling system, regulates sexual activity and fecundity in insects, Proc. Natl. Acad. Sci. U. S. A. 110 (2013) E3526–E3534.
[368] T. Ida, T. Takahashi, H. Tominaga, T. Sato, K. Kume, K. Yoshizawa-Kumagaye, H. Nishio, J. Kato, N. Murakami, M. Miyazato, K. Kangawa, M. Kojima, Identification of the endogenous cysteine-rich peptide trissin, a ligand for an orphan G protein-coupled receptor in *Drosophila*, Biochem. Biophys. Res. Commun. 414 (2011) 44–48.

[369] E. Kubli, The sex-peptide, Bioessays 14 (1992) 779–784.
[370] K. Moehle, A. Freund, E. Kubli, J.A. Robinson, NMR studies of the solution conformation of the sex peptide from *Drosophila melanogaster*, FEBS Lett. 585 (2011) 1197–1202.
[371] J.C. Gelly, J. Gracy, Q. Kaas, D. Le-Nguyen, A. Heitz, L. Chiche, The KNOTTIN website and database: a new information system dedicated to the knottin scaffold, Nucleic Acids Res. 32 (2004) D156–D159.
[372] T. Chapman, L.F. Liddle, J.M. Kalb, M.F. Wolfner, L. Partridge, Cost of mating in *Drosophila melanogaster* females is mediated by male accessory gland products, Nature 373 (1995) 241–244.
[373] S.A. Monsma, M.F. Wolfner, Structure and expression of a *Drosophila* male accessory gland gene whose product resembles a peptide pheromone precursor, Genes Dev. 2 (1988) 1063–1073.
[374] J.F.V. Vincent, Arthropod cuticle: a natural composite shell system, Compos. A: Appl. Sci. Manuf. 33 (2015) 1311–1315.
[375] R.S. Cornman, Molecular evolution of *Drosophila* cuticular protein genes, PLoS One 4 (2009), e8345. .
[376] J.H. Willis, Structural cuticular proteins from arthropods: annotation, nomenclature, and sequence characteristics in the genomics era, Insect Biochem. Mol. Biol. 40 (2010) 189–204.
[377] J.E. Rebers, L.M. Riddiford, Structure and expression of a *Manduca sexta* larval cuticle gene homologous to *Drosophila* cuticle genes, J. Mol. Biol. 203 (1988) 411–423.
[378] N. Komori, J. Usukura, H. Matsumoto, Drosocrystallin, a major 52 kDa glycoprotein of the *Drosophila melanogaster* corneal lens. Purification, biochemical characterization, and subcellular localization, J. Cell Sci. 102 (Pt 2) (1992) 191–201.
[379] T. Kuraishi, O. Binggeli, O. Opota, N. Buchon, B. Lemaitre, Genetic evidence for a protective role of the peritrophic matrix against intestinal bacterial infection in *Drosophila melanogaster*, Proc. Natl. Acad. Sci. U. S. A. 108 (2011) 15966–15971.
[380] T. Suetake, S. Tsuda, S. Kawabata, K. Miura, S. Iwanaga, K. Hikichi, K. Nitta, K. Kawano, Chitin-binding proteins in invertebrates and plants comprise a common chitin-binding structural motif, J. Biol. Chem. 275 (2000) 17929–17932.
[381] S. Jasrapuria, C.A. Specht, K.J. Kramer, R.W. Beeman, S. Muthukrishnan, Gene families of cuticular proteins analogous to peritrophins (CPAPs) in *Tribolium castaneum* have diverse functions, PLoS One 7 (2012), e49844. .
[382] C.M. Elvin, T. Vuocolo, R.D. Pearson, I.J. East, G.A. Riding, C.H. Eisemann, R.L. Tellam, Characterization of a major peritrophic membrane protein, peritrophin-44, from the larvae of *Lucilia cuprina*. cDNA and deduced amino acid sequences, J. Biol. Chem. 271 (1996) 8925–8935.
[383] Z. Shen, M. Jacobs-Lorena, A type I peritrophic matrix protein from the malaria vector *Anopheles gambiae* binds to chitin. Cloning, expression, and characterization, J. Biol. Chem. 273 (1998) 17665–17670.
[384] M. Behr, M. Hoch, Identification of the novel evolutionary conserved obstructor multigene family in invertebrates, FEBS Lett. 579 (2005) 6827–6833.
[385] M.K. Barry, A.A. Triplett, A.C. Christensen, A peritrophin-like protein expressed in the embryonic tracheae of *Drosophila melanogaster*, Insect Biochem. Mol. Biol. 29 (1999) 319–327.
[386] K. Tiklova, V. Tsarouhas, C. Samakovlis, Control of airway tube diameter and integrity by secreted chitin-binding proteins in *Drosophila*, PLoS One 8 (2013), e67415.
[387] Y.Y. Pesch, D. Riedel, M. Behr, Obstructor A organizes matrix assembly at the apical cell surface to promote enzymatic cuticle maturation in *Drosophila*, J. Biol. Chem. 290 (2015) 10071–10082.
[388] Z.A. Syed, T. Hard, A. Uv, I.F. van Dijk-Hard, A potential role for *Drosophila* mucins in development and physiology, PLoS One 3 (2008), e3041.
[389] S.W. Robinson, P. Herzyk, J.A. Dow, D.P. Leader, FlyAtlas: database of gene expression in the tissues of *Drosophila melanogaster*, Nucleic Acids Res. 41 (2013) D744–D750.
[390] X. Guan, B.W. Middlebrooks, S. Alexander, S.A. Wasserman, Mutation of TweedleD, a member of an unconventional cuticle protein family, alters body shape in *Drosophila*, Proc. Natl. Acad. Sci. U. S. A. 103 (2006) 16794–16799.
[391] L. Tang, J. Liang, Z. Zhan, Z. Xiang, N. He, Identification of the chitin-binding proteins from the larval proteins of silkworm, Bombyx mori, Insect Biochem. Mol. Biol. 40 (2010) 228–234.
[392] A.N. Zelensky, J.E. Gready, The C-type lectin-like domain superfamily, FEBS J. 272 (2005) 6179–6217.
[393] T. Tanji, A. Ohashi-Kobayashi, S. Natori, Participation of a galactose-specific C-type lectin in *Drosophila* immunity, Biochem. J. 396 (2006) 127–138.
[394] J. Ao, E. Ling, X.Q. Yu, *Drosophila* C-type lectins enhance cellular encapsulation, Mol. Immunol. 44 (2007) 2541–2548.
[395] S. Banerjee, A.M. Pillai, R. Paik, J. Li, M.A. Bhat, Axonal ensheathment and septate junction formation in the peripheral nervous system of *Drosophila*, J. Neurosci. 26 (2006) 3319–3329.
[396] A. Hildebrandt, R. Pflanz, M. Behr, T. Tarp, D. Riedel, R. Schuh, Bark beetle controls epithelial morphogenesis by septate junction maturation in *Drosophila*, Dev. Biol. 400 (2015) 237–247.
[397] Y. Li, P. Dharkar, T.H. Han, M. Serpe, C.H. Lee, M.L. Mayer, Novel functional properties of *Drosophila* CNS glutamate receptors, Neuron 92 (2016) 1036–1048.
[398] V. Croset, R. Rytz, S.F. Cummins, A. Budd, D. Brawand, H. Kaessmann, T.J. Gibson, R. Benton, Ancient protostome origin of chemosensory ionotropic glutamate receptors and the evolution of insect taste and olfaction, PLoS Genet. 6 (2010), e1001064.
[399] X. Tang, B. Zhou, Iron homeostasis in insects: insights from *Drosophila* studies, IUBMB Life 65 (2013) 863–872.
[400] A. Kuryatov, B. Laube, H. Betz, J. Kuhse, Mutational analysis of the glycine-binding site of the NMDA receptor: structural similarity with bacterial amino acid-binding proteins, Neuron 12 (1994) 1291–1300.
[401] C. Mitri, M.L. Parmentier, J.P. Pin, J. Bockaert, Y. Grau, Divergent evolution in metabotropic glutamate receptors. A new receptor activated by an endogenous ligand different from glutamate in insects, J. Biol. Chem. 279 (2004) 9313–9320.
[402] M. Mezler, T. Muller, K. Raming, Cloning and functional expression of GABA(B) receptors from *Drosophila*, Eur. J. Neurosci. 13 (2001) 477–486.
[403] K.J. Vogel, M.R. Brown, M.R. Strand, Phylogenetic investigation of Peptide hormone and growth factor receptors in five dipteran genomes, Front. Endocrinol. (Lausanne) 4 (2013) 193.
[404] R.M. Joseph, J.R. Carlson, *Drosophila* chemoreceptors: a molecular interface between the chemical world and the brain, Trends Genet. 31 (2015) 683–695.
[405] F.G. Vieira, J. Rozas, Comparative genomics of the odorant-binding and chemosensory protein gene families

across the Arthropoda: origin and evolutionary history of the chemosensory system, Genome Biol. Evol. 3 (2011) 476–490.
[406] R.G. Vogt, G.D. Prestwich, M.R. Lerner, Odorant-binding-protein subfamilies associate with distinct classes of olfactory receptor neurons in insects, J. Neurobiol. 22 (1991) 74–84.
[407] W.S. Leal, Odorant reception in insects: roles of receptors, binding proteins, and degrading enzymes, Annu. Rev. Entomol. 58 (2013) 373–391.
[408] L. Sabatier, E. Jouanguy, C. Dostert, D. Zachary, J.L. Dimarcq, P. Bulet, J.L. Imler, Pherokine-2 and -3, Eur. J. Biochem. 270 (2003) 3398–3407.
[409] Z. Nichols, R.G. Vogt, The SNMP/CD36 gene family in Diptera, Hymenoptera and Coleoptera: *Drosophila melanogaster*, *D. pseudoobscura*, *Anopheles gambiae*, *Aedes aegypti*, *Apis mellifera*, and *Tribolium castaneum*, Insect Biochem. Mol. Biol. 38 (2008) 398–415.
[410] R.L. Silverstein, M. Febbraio, CD36, a scavenger receptor involved in immunity, metabolism, angiogenesis, and behavior, Sci. Signal. 2 (2009), re3.
[411] C. Han, Y. Song, H. Xiao, D. Wang, N.C. Franc, L.Y. Jan, Y.N. Jan, Epidermal cells are the primary phagocytes in the fragmentation and clearance of degenerating dendrites in *Drosophila*, Neuron 81 (2014) 544–560.
[412] C. Kiefer, E. Sumser, M.F. Wernet, J. Von Lintig, A class B scavenger receptor mediates the cellular uptake of carotenoids in *Drosophila*, Proc. Natl. Acad. Sci. U. S. A. 99 (2002) 10581–10586.
[413] C. Gomez-Diaz, B. Bargeton, L. Abuin, N. Bukar, J.H. Reina, T. Bartoi, M. Graf, H. Ong, M.H. Ulbrich, J.F. Masson, R. Benton, A CD36 ectodomain mediates insect pheromone detection via a putative tunnelling mechanism, Nat. Commun. 7 (2016) 11866.
[414] D.R. Flower, A.C. North, C.E. Sansom, The lipocalin protein family: structural and sequence overview, Biochim. Biophys. Acta 1482 (2000) 9–24.
[415] D. Sanchez, B. Lopez-Arias, L. Torroja, I. Canal, X. Wang, M.J. Bastiani, M.D. Ganfornina, Loss of glial lazarillo, a homolog of apolipoprotein D, reduces lifespan and stress resistance in *Drosophila*, Curr. Biol. 16 (2006) 680–686.
[416] J. Hull-Thompson, J. Muffat, D. Sanchez, D.W. Walker, S. Benzer, M.D. Ganfornina, H. Jasper, Control of metabolic homeostasis by stress signaling is mediated by the lipocalin NLaz, PLoS Genet. 5 (2009), e1000460. .
[417] D.W. Walker, J. Muffat, C. Rundel, S. Benzer, Overexpression of a *Drosophila* homolog of apolipoprotein D leads to increased stress resistance and extended lifespan, Curr. Biol. 16 (2006) 674–679.
[418] J.R. Thompson, L.J. Banaszak, Lipid-protein interactions in lipovitellin, Biochemistry 41 (2002) 9398–9409.
[419] W. Palm, J.L. Sampaio, M. Brankatschk, M. Carvalho, A. Mahmoud, A. Shevchenko, S. Eaton, Lipoproteins in *Drosophila melanogaster*—assembly, function, and influence on tissue lipid composition, PLoS Genet. 8 (2012), e1002828.
[420] J. Chen, S.M. Honeyager, J. Schleede, A. Avanesov, A. Laughon, S.S. Blair, Crossveinless d is a vitellogenin-like lipoprotein that binds BMPs and HSPGs, and is required for normal BMP signaling in the *Drosophila* wing, Development 139 (2012) 2170–2176.
[421] M. Lamant, F. Smih, R. Harmancey, P. Philip-Couderc, A. Pathak, J. Roncalli, M. Galinier, X. Collet, P. Massabuau, J.M. Senard, P. Rouet, ApoO, a novel apolipoprotein, is an original glycoprotein up-regulated by diabetes in human heart, J. Biol. Chem. 281 (2006) 36289–36302.
[422] N.M. Wang, M.F. Lee, C.H. Wu, Immunologic characterization of a recombinant American cockroach (*Periplaneta americana*) Per a 1 (Cr-PII) allergen, Allergy 54 (1999) 119–127.
[423] G.A. Mueller, L.C. Pedersen, F.B. Lih, J. Glesner, A.F. Moon, M. D. Chapman, K.B. Tomer, R.E. London, A. Pomes, The novel structure of the cockroach allergen Bla g 1 has implications for allergenicity and exposure assessment, J. Allergy Clin. Immunol. 132 (2013) 1420–1426.
[424] K. Honjo, S.E. Mauthner, Y. Wang, J.H.P. Skene, W.D. Tracey Jr., Nociceptor-enriched genes required for normal thermal nociception, Cell Rep. 16 (2016) 295–303.
[425] B.A. Stokes, S. Yadav, U. Shokal, L.C. Smith, I. Eleftherianos, Bacterial and fungal pattern recognition receptors in homologous innate signaling pathways of insects and mammals, Front. Microbiol. 6 (2015) 19.
[426] J.L. Imler, P. Bulet, Antimicrobial peptides in *Drosophila*: structures, activities and gene regulation, Chem. Immunol. Allergy 86 (2005) 1–21.
[427] J.R. Carlson, D.S. Hogness, The Jonah genes: a new multigene family in *Drosophila melanogaster*, Dev. Biol. 108 (1985) 341–354.
[428] S. Wang, C. Magoulas, D. Hickey, Concerted evolution within a trypsin gene cluster in *Drosophila*, Mol. Biol. Evol. 16 (1999) 1117–1124.
[429] M. Dissing, H. Giordano, R. DeLotto, Autoproteolysis and feedback in a protease cascade directing *Drosophila* dorsal-ventral cell fate, EMBO J. 20 (2001) 2387–2393.
[430] G. Didelot, F. Molinari, P. Tchenio, D. Comas, E. Milhiet, A. Munnich, L. Colleaux, T. Preat, Tequila, a neurotrypsin ortholog, regulates long-term memory formation in *Drosophila*, Science 313 (2006) 851–853.
[431] J.L. Sitnik, C. Francis, K. Hens, R. Huybrechts, M.F. Wolfner, P. Callaerts, Neprilysins: an evolutionarily conserved family of metalloproteases that play important roles in reproduction in *Drosophila*, Genetics 196 (2014) 781–797.
[432] S. Matsuoka, S. Gupta, E. Suzuki, Y. Hiromi, M. Asaoka, gone early, a novel germline factor, ensures the proper size of the stem cell precursor pool in the *Drosophila* ovary, PLoS One 9 (2014), e113423.
[433] O. Shimmi, M.B. O'Connor, Physical properties of Tld, Sog, Tsg and Dpp protein interactions are predicted to help create a sharp boundary in Bmp signals during dorsoventral patterning of the *Drosophila* embryo, Development 130 (2003) 4673–4682.
[434] R.J. Siviter, C.A. Taylor, D.M. Cottam, A. Denton, M.P. Dani, M.J. Milner, A.D. Shirras, R.E. Isaac, Ance, a *Drosophila* angiotensin-converting enzyme homologue, is expressed in imaginal cells during metamorphosis and is regulated by the steroid, 20-hydroxyecdysone, Biochem. J. 367 (2002) 187–193.
[435] G. Sidyelyeva, L.D. Fricker, Characterization of *Drosophila* carboxypeptidase D, J. Biol. Chem. 277 (2002) 49613–49620.
[436] S. Schilling, C. Lindner, B. Koch, M. Wermann, J.U. Rahfeld, A. von Bohlen, T. Rudolph, G. Reuter, H.U. Demuth, Isolation and characterization of glutaminyl cyclases from *Drosophila*: evidence for enzyme forms with different subcellular localization, Biochemistry 46 (2007) 10921–10930.
[437] Y. Hu, Y. Ye, M.E. Fortini, Nicastrin is required for gamma-secretase cleavage of the *Drosophila* Notch receptor, Dev. Cell 2 (2002) 69–78.
[438] C. Haffner, U. Dettmer, T. Weiler, C. Haass, The Nicastrin-like protein Nicalin regulates assembly and stability of

the Nicalin-nodal modulator (NOMO) membrane protein complex, J. Biol. Chem. 282 (2007) 10632–10638.
[439] Y. Liao, J. Pei, H. Cheng, N.V. Grishin, An ancient autoproteolytic domain found in GAIN, ZU5 and Nucleoporin98, J. Mol. Biol. 426 (2014) 3935–3945.
[440] K. Mouri, Y. Nishino, M. Arata, D. Shi, S.Y. Horiuchi, T. Uemura, A novel planar polarity gene pepsinogen-like regulates wingless expression in a posttranscriptional manner, Dev. Dyn. 243 (2014) 791–799.
[441] T. Barnett, C. Pachl, J.P. Gergen, P.C. Wensink, The isolation and characterization of *Drosophila* yolk protein genes, Cell 21 (1980) 729–738.
[442] M. Bownes, Why is there sequence similarity between insect yolk proteins and vertebrate lipases? J. Lipid Res. 33 (1992) 777–790.
[443] A.M. Craig, Y. Kang, Neurexin-neuroligin signaling in synapse development, Curr. Opin. Neurobiol. 17 (2007) 43–52.
[444] R. Dixit, Y. Arakane, C.A. Specht, C. Richard, K.J. Kramer, R.W. Beeman, S. Muthukrishnan, Domain organization and phylogenetic analysis of proteins from the chitin deacetylase gene family of *Tribolium castaneum* and three other species of insects, Insect Biochem. Mol. Biol. 38 (2008) 440–451.
[445] D.E. Blair, O. Hekmat, A.W. Schuttelkopf, B. Shrestha, K. Tokuyasu, S.G. Withers, D.M. van Aalten, Structure and mechanism of chitin deacetylase from the fungal pathogen *Colletotrichum lindemuthianum*, Biochemistry 45 (2006) 9416–9426.
[446] S. Luschnig, T. Batz, K. Armbruster, M.A. Krasnow, Serpentine and vermiform encode matrix proteins with chitin binding and deacetylation domains that limit tracheal tube length in *Drosophila*, Curr. Biol. 16 (2006) 186–194.
[447] S. Chakraborti, B.J. Bahnson, Crystal structure of human senescence marker protein 30: insights linking structural, enzymatic, and physiological functions, Biochemistry 49 (2010) 3436–3444.
[448] M.A. Hicks, A.E. Barber II, L.A. Giddings, J. Caldwell, S.E. O'Connor, P.C. Babbitt, The evolution of function in strictosidine synthase-like proteins, Proteins 79 (2011) 3082–3098.
[449] C.C. Akoh, G.C. Lee, Y.C. Liaw, T.H. Huang, J.F. Shaw, GDSL family of serine esterases/lipases, Prog. Lipid Res. 43 (2004) 534–552.
[450] C.S. Seong, A. Varela-Ramirez, X. Tang, B. Anchondo, D. Magallanes, R.J. Aguilera, Cloning and characterization of a novel *Drosophila* stress induced DNase, PLoS One 9 (2014), e103564.
[451] A. Kleinschmit, M. Takemura, K. Dejima, P.Y. Choi, H. Nakato, *Drosophila* heparan sulfate 6-*O*-endosulfatase Sulf1 facilitates wingless (Wg) protein degradation, J. Biol. Chem. 288 (2013) 5081–5089.
[452] J. You, T. Belenkaya, X. Lin, Sulfated is a negative feedback regulator of wingless in *Drosophila*, Dev. Dyn. 240 (2011) 640–648.
[453] R. Bhadra, N. Srinivasan, S.B. Pandit, A new domain family in the superfamily of alkaline phosphatases, In Silico Biol. 5 (2005) 379–387.
[454] J.D. Funkhouser, N.N. Aronson Jr., Chitinase family GH18: evolutionary insights from the genomic history of a diverse protein family, BMC Evol. Biol. 7 (2007) 96.
[455] P.F. Varela, A.S. Llera, R.A. Mariuzza, J. Tormo, Crystal structure of imaginal disc growth factor-2. A member of a new family of growth-promoting glycoproteins from *Drosophila melanogaster*, J. Biol. Chem. 277 (2002) 13229–13236.
[456] Y.Y. Pesch, D. Riedel, K.R. Patil, G. Loch, M. Behr, Chitinases and Imaginal disc growth factors organize the extracellular matrix formation at barrier tissues in insects, Sci. Rep. 6 (2016) 18340.
[457] S. Bachali, M. Jager, A. Hassanin, F. Schoentgen, P. Jolles, A. Fiala-Medioni, J.S. Deutsch, Phylogenetic analysis of invertebrate lysozymes and the evolution of lysozyme function, J. Mol. Evol. 54 (2002) 652–664.
[458] L.L. Zavalova, I.P. Baskova, S.A. Lukyanov, A.V. Sass, E.V. Snezhkov, S.B. Akopov, I.I. Artamonova, V.S. Archipova, V. A. Nesmeyanov, D.G. Kozlov, S.V. Benevolensky, V.I. Kiseleva, A.M. Poverenny, E.D. Sverdlov, Destabilase from the medicinal leech is a representative of a novel family of lysozymes, Biochim. Biophys. Acta 1478 (2000) 69–77.
[459] L.L. Zavalova, N.V. Antipova, I. Fadeeva Iu, M.S. Pavliukov, N. V. Pletneva, V.Z. Pletnev, I.P. Baskova, Catalytic sites of medicinal leech enzyme destabilase-lysozyme (Mldl). Structure–functional correlation, Bioorg. Khim. 38 (2012) 229–233.
[460] R. Pfeiffer, G. Rossier, B. Spindler, C. Meier, L. Kuhn, F. Verrey, Amino acid transport of y + L-type by heterodimers of 4F2hc/CD98 and members of the glycoprotein-associated amino acid transporter family, EMBO J. 18 (1999) 49–57.
[461] R. Leonard, D. Rendic, C. Rabouille, I.B. Wilson, T. Preat, F. Altmann, The *Drosophila* fused lobes gene encodes an *N*-acetylglucosaminidase involved in *N*-glycan processing, J. Biol. Chem. 281 (2006) 4867–4875.
[462] M. Yoshida, H. Matsuda, H. Kubo, T. Nishimura, Molecular characterization of Tps1 and Treh genes in *Drosophila* and their role in body water homeostasis, Sci. Rep. 6 (2016) 30582.
[463] J. Royet, D. Gupta, R. Dziarski, Peptidoglycan recognition proteins: modulators of the microbiome and inflammation, Nat. Rev. Immunol. 11 (2011) 837–851.
[464] M. Zurovec, T. Dolezal, M. Gazi, E. Pavlova, P.J. Bryant, Adenosine deaminase-related growth factors stimulate cell proliferation in *Drosophila* by depleting extracellular adenosine, Proc. Natl. Acad. Sci. U. S. A. 99 (2002) 4403–4408.
[465] P. Bork, E.V. Koonin, A new family of carbon–nitrogen hydrolases, Protein Sci. 3 (1994) 1344–1346.
[466] G. Camporeale, E. Giordano, R. Rendina, J. Zempleni, J.C. Eissenberg, *Drosophila melanogaster* holocarboxylase synthetase is a chromosomal protein required for normal histone biotinylation, gene transcription patterns, lifespan, and heat tolerance, J. Nutr. 136 (2006) 2735–2742.
[467] T. Inoue, N. Okino, Y. Kakuta, A. Hijikata, H. Okano, H.M. Goda, M. Tani, N. Sueyoshi, K. Kambayashi, H. Matsumura, Y. Kai, M. Ito, Mechanistic insights into the hydrolysis and synthesis of ceramide by neutral ceramidase, J. Biol. Chem. 284 (2009) 9566–9577.
[468] E.M. Ha, C.T. Oh, J.H. Ryu, Y.S. Bae, S.W. Kang, I.H. Jang, P.T. Brey, W.J. Lee, An antioxidant system required for host protection against gut infection in *Drosophila*, Dev. Cell 8 (2005) 125–132.
[469] G.R. Hemsworth, B. Henrissat, G.J. Davies, P.H. Walton, Discovery and characterization of a new family of lytic polysaccharide monooxygenases, Nat. Chem. Biol. 10 (2014) 122–126.
[470] Z. Forsberg, A.K. Mackenzie, M. Sorlie, A.K. Rohr, R. Helland, A.S. Arvai, G. Vaaje-Kolstad, V.G. Eijsink, Structural and functional characterization of a conserved pair of bacterial cellulose-oxidizing lytic polysaccharide monooxygenases, Proc. Natl. Acad. Sci. U. S. A. 111 (2014) 8446–8451.
[471] D.J. Tan, H. Dvinge, A. Christoforou, P. Bertone, A. Martinez Arias, K.S. Lilley, Mapping organelle proteins

and protein complexes in *Drosophila melanogaster*, J. Proteome Res. 8 (2009) 2667–2678.
[472] K. Kongton, K. McCall, A. Phongdara, Identification of gamma-interferon-inducible lysosomal thiol reductase (GILT) homologues in the fruit fly *Drosophila melanogaster*, Dev. Comp. Immunol. 44 (2014) 389–396.
[473] A. Messerschmidt, R. Ladenstein, R. Huber, M. Bolognesi, L. Avigliano, R. Petruzzelli, A. Rossi, A. Finazzi-Agro, Refined crystal structure of ascorbate oxidase at 1.9 A resolution, J. Mol. Biol. 224 (1992) 179–205.
[474] S.T. Prigge, R.E. Mains, B.A. Eipper, L.M. Amzel, New insights into copper monooxygenases and peptide amidation: structure, mechanism and function, Cell. Mol. Life Sci. 57 (2000) 1236–1259.
[475] M. Monastirioti, C.E. Linn Jr., K. White, Characterization of *Drosophila* tyramine beta-hydroxylase gene and isolation of mutant flies lacking octopamine, J. Neurosci. 16 (1996) 3900–3911.
[476] A.S. Kolhekar, M.S. Roberts, N. Jiang, R.C. Johnson, R.E. Mains, B.A. Eipper, P.H. Taghert, Neuropeptide amidation in *Drosophila*: separate genes encode the two enzymes catalyzing amidation, J. Neurosci. 17 (1997) 1363–1376.
[477] X. Xin, R.E. Mains, B.A. Eipper, Monooxygenase X, a member of the copper-dependent monooxygenase family localized to the endoplasmic reticulum, J. Biol. Chem. 279 (2004) 48159–48167.
[478] J. Molnar, K.S. Fong, Q.P. He, K. Hayashi, Y. Kim, S.F. Fong, B. Fogelgren, K.M. Szauter, M. Mink, K. Csiszar, Structural and functional diversity of lysyl oxidase and the LOX-like proteins, Biochim. Biophys. Acta 1647 (2003) 220–224.
[479] X. Zhang, Q. Wang, J. Wu, J. Wang, Y. Shi, M. Liu, Crystal structure of human lysyl oxidase-like 2 (hLOXL2) in a precursor state, Proc. Natl. Acad. Sci. U. S. A. 115 (2018) 3828–3833.
[480] R.M. Cardoso, C.H. Silva, A.P. Ulian de Araujo, T. Tanaka, M. Tanaka, R.C. Garratt, Structure of the cytosolic Cu,Zn superoxide dismutase from *Schistosoma mansoni*, Acta Crystallogr. D Biol. Crystallogr. 60 (2004) 1569–1578.
[481] N.H. Haunerland, Insect storage proteins: gene families and receptors, Insect Biochem. Mol. Biol. 26 (1996) 755–765.
[482] T. Burmester, C. Antoniewski, J.A. Lepesant, Ecdysone-regulation of synthesis and processing of fat body protein 1, the larval serum protein receptor of *Drosophila melanogaster*, Eur. J. Biochem. 262 (1999) 49–55.
[483] C. Courtay, T. Oster, F. Michelet, A. Visvikis, M. Diederich, M. Wellman, G. Siest, Gamma-glutamyltransferase: nucleotide sequence of the human pancreatic cDNA. Evidence for a ubiquitous gamma-glutamyltransferase polypeptide in human tissues, Biochem. Pharmacol. 43 (1992) 2527–2533.
[484] J. Wang, A.K. Gebre, R.A. Anderson, J.S. Parks, Cloning and in vitro expression of rat lecithin:cholesterol acyltransferase, Biochim. Biophys. Acta 1346 (1997) 207–211.
[485] A.J. Oakley, M. Coggan, P.G. Board, Identification and characterization of gamma-glutamylamine cyclotransferase, an enzyme responsible for gamma-glutamyl-epsilon-lysine catabolism, J. Biol. Chem. 285 (2010) 9642–9648.
[486] J. Xiao, V.S. Tagliabracci, J. Wen, S.A. Kim, J.E. Dixon, Crystal structure of the Golgi casein kinase, Proc. Natl. Acad. Sci. U. S. A. 110 (2013) 10574–10579.
[487] K.S. Krishnan, R. Rikhy, S. Rao, M. Shivalkar, M. Mosko, R. Narayanan, P. Etter, P.S. Estes, M. Ramaswami, Nucleoside diphosphate kinase, a source of GTP, is required for dynamin-dependent synaptic vesicle recycling, Neuron 30 (2001) 197–210.
[488] C.U. Vieira, A.M. Bonetti, Z.L. Simoes, A.Q. Maranhao, C.S. Costa, M.C. Costa, A.C. Siquieroli, F.M. Nunes, Farnesoic acid *O*-methyl transferase (FAMeT) isoforms: conserved traits and gene expression patterns related to caste differentiation in the stingless bee, Melipona scutellaris, Arch. Insect Biochem. Physiol. 67 (2008) 97–106.
[489] S. Lindskog, Structure and mechanism of carbonic anhydrase, Pharmacol. Ther. 74 (1997) 1–20.
[490] M. Han, D. Park, P.J. Vanderzalm, R.E. Mains, B.A. Eipper, P.H. Taghert, *Drosophila* uses two distinct neuropeptide amidating enzymes, dPAL1 and dPAL2, J. Neurochem. 90 (2004) 129–141.
[491] P. Wang, J. Heitman, The cyclophilins, Genome Biol. 6 (2005) 226.
[492] F. Veillard, L. Troxler, J.M. Reichhart, *Drosophila melanogaster* clip-domain serine proteases: structure, function and regulation, Biochimie 122 (2016) 255–269.
[493] S. Ranasinghe, D.P. McManus, Structure and function of invertebrate Kunitz serine protease inhibitors, Dev. Comp. Immunol. 39 (2013) 219–227.
[494] V. Rimphanitchayakit, A. Tassanakajon, Structure and function of invertebrate Kazal-type serine proteinase inhibitors, Dev. Comp. Immunol. 34 (2010) 377–386.
[495] J.M. Reichhart, Tip of another iceberg: *Drosophila* serpins, Trends Cell Biol. 15 (2005) 659–665.
[496] R.S. Comman, The distribution of GYR- and YLP-like motifs in *Drosophila* suggests a general role in cuticle assembly and other protein–protein interactions, PLoS One 5 (2010).
[497] T. Kawano, M. Shimoda, H. Matsumoto, M. Ryuda, S. Tsuzuki, Y. Hayakawa, Identification of a gene, Desiccate, contributing to desiccation resistance in *Drosophila melanogaster*, J. Biol. Chem. 285 (2010) 38889–38897.
[498] T. Kawano, M. Ryuda, H. Matsumoto, M. Ochiai, Y. Oda, T. Tanimura, G. Csikos, M. Moriya, Y. Hayakawa, Function of desiccate in gustatory sensilla of *Drosophila melanogaster*, Sci. Rep. 5 (2015) 17195.
[499] T. Shiraiwa, E. Nitasaka, T. Yamazaki, Geko, a novel gene involved in olfaction in *Drosophila melanogaster*, J. Neurogenet. 14 (2000) 145–164.
[500] D.R. Dorer, J.A. Rudnick, E.N. Moriyama, A.C. Christensen, A family of genes clustered at the Triplo-lethal locus of *Drosophila melanogaster* has an unusual evolutionary history and significant synteny with *Anopheles gambiae*, Genetics 165 (2003) 613–621.
[501] J.M. Andrade Lopez, S.M. Lanno, J.M. Auerbach, E.C. Moskowitz, L.A. Sligar, P.J. Wittkopp, J.D. Coolon, Genetic basis of octanoic acid resistance in *Drosophila sechellia*: functional analysis of a fine-mapped region, Mol. Ecol. 26 (2017) 1148–1160.
[502] C.R. Smith, C. Morandin, M. Noureddine, S. Pant, Conserved roles of Osiris genes in insect development, polymorphism and protection, J. Evol. Biol. 31 (2018) 516–529.
[503] M. Trienens, K. Kraaijeveld, B. Wertheim, Defensive repertoire of *Drosophila* larvae in response to toxic fungi, Mol. Ecol. 26 (2017) 5043–5057.
[504] R.S. Comman, D. Lopez, J.D. Evans, Transcriptional response of honey bee larvae infected with the bacterial pathogen *Paenibacillus* larvae, PLoS One 8 (2013), e65424.
[505] S. Uttenweiler-Joseph, M. Moniatte, M. Lagueux, A. Van Dorsselaer, J.A. Hoffmann, P. Bulet, Differential display of peptides induced during the immune response of *Drosophila*: a matrix-assisted laser desorption ionization time-of-flight mass spectrometry study, Proc. Natl. Acad. Sci. U. S. A. 95 (1998) 11342–11347.

[506] A.W. Clemmons, S.A. Lindsay, S.A. Wasserman, An effector Peptide family required for *Drosophila* toll-mediated immunity, PLoS Pathog. 11 (2015), e1004876.
[507] A. Goto, T. Yano, J. Terashima, S. Iwashita, Y. Oshima, S. Kurata, Cooperative regulation of the induction of the novel antibacterial Listericin by peptidoglycan recognition protein LE and the JAK–STAT pathway, J. Biol. Chem. 285 (2010) 15731–15738.
[508] S. Ekengren, D. Hultmark, A family of Turandot-related genes in the humoral stress response of *Drosophila*, Biochem. Biophys. Res. Commun. 284 (2001) 998–1003.
[509] M. Chen, J.L. Manley, Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches, Nat. Rev. Mol. Cell Biol. 10 (2009) 741–754.
[510] A.R. Kornblihtt, Promoter usage and alternative splicing, Curr. Opin. Cell Biol. 17 (2005) 262–268.
[511] D.C. Di Giammartino, K. Nishida, J.L. Manley, Mechanisms and consequences of alternative polyadenylation, Mol. Cell 43 (2011) 853–866.
[512] M. Schaub, W. Keller, RNA editing by adenosine deaminases generates RNA and protein diversity, Biochimie 84 (2002) 791–803.
[513] A.V. Kochetov, Alternative translation start sites and hidden coding potential of eukaryotic mRNAs, Bioessays 30 (2008) 683–691.
[514] I. Jungreis, M.F. Lin, R. Spokony, C.S. Chan, N. Negre, A. Victorsen, K.P. White, M. Kellis, Evidence of abundant stop codon readthrough in *Drosophila* and other metazoa, Genome Res. 21 (2011) 2096–2113.
[515] G.A. Khoury, R.C. Baliban, C.A. Floudas, Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database, Sci. Rep. 1 (2011).
[516] E.E. Rosenbaum, K.S. Brehm, E. Vasiljevic, A. Gajeski, N.J. Colley, *Drosophila* GPI-mannosyltransferase 2 is required for GPI anchor attachment and surface expression of chaoptin, Vis. Neurosci. 29 (2012) 143–156.
[517] C. Faivre-Sarrailh, S. Banerjee, J. Li, M. Hortsch, M. Laval, M.A. Bhat, *Drosophila* contactin, a homolog of vertebrate contactin, is required for septate junction organization and paracellular barrier function, Development 131 (2004) 4931–4942.
[518] G. Gennarini, G. Cibelli, G. Rougon, M.G. Mattei, C. Goridis, The mouse neuronal cell surface protein F3: a phosphatidylinositol-anchored member of the immunoglobulin superfamily related to chicken contactin, J. Cell Biol. 109 (1989) 775–788.
[519] R.M. Waterhouse, F. Tegenfeldt, J. Li, E.M. Zdobnov, E.V. Kriventseva, OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs, Nucleic Acids Res. 41 (2013) D358–D365.
[520] K. Katoh, D.M. Standley, MAFFT multiple sequence alignment software version 7: improvements in performance and usability, Mol. Biol. Evol. 30 (2013) 772–780.
[521] T.N. Petersen, S. Brunak, G. von Heijne, H. Nielsen, SignalP 4.0: discriminating signal peptides from transmembrane regions, Nat. Methods 8 (2011) 785–786.
[522] L. Kall, A. Krogh, E.L. Sonnhammer, Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server, Nucleic Acids Res. 35 (2007) W429–W432.
[523] A. Krogh, B. Larsson, G. von Heijne, E.L. Sonnhammer, Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes, J. Mol. Biol. 305 (2001) 567–580.
[524] A. Pierleoni, P.L. Martelli, R. Casadio, PredGPI: a GPI-anchor predictor, BMC Bioinformatics 9 (2008) 392.
[525] G. Poisson, C. Chauve, X. Chen, A. Bergeron, FragAnchor: a large-scale predictor of glycosylphosphatidylinositol anchors in eukaryote protein sequences by qualitative scoring, Genomics Proteomics Bioinformatics 5 (2007) 121–130.
[526] R.D. Finn, J. Clements, S.R. Eddy, HMMER web server: interactive sequence similarity searching, Nucleic Acids Res. 39 (2011) W29–W37.
[527] A. Bateman, L. Coin, R. Durbin, R.D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E.L. Sonnhammer, D.J. Studholme, C. Yeats, S.R. Eddy, The Pfam protein families database, Nucleic Acids Res. 32 (2004) D138–D141.
[528] J. Soding, Protein homology detection by HMM-HMM comparison, Bioinformatics 21 (2005) 951–960.