Mass spectrometric characterization of the major peptides of the male ejaculatory duct, including a glycopeptide with an unusual zwitterionic glycosylation

Sebastian Sturm[a], Adam Dowle[b] , Neil Audsley[c], R. Elwyn Isaac[a*]

[a] School of Biology, University of Leeds, Leeds LS2 9JT, UK.

[b] Bioscience Technology Facility, Department of Biology, University of York, Wentworth Way, York YO10 5DD, UK. Email:adam.dowle@york.ac.uk

[c] Institute for Agri-Food Research and Innovation, Newcastle University, Newcastle Upon-Tyne, NE1 7RU, UK. Email:neil.audsley@newcastle.ac.uk

*Corresponding author.

Email address: r.e.isaac@leeds.ac.uk

Declarations of interest: none.

ABSTRACT

Peptides present in the seminal fluid of *Drosophila melanogaster* can function as antimicrobial agents, enzyme inhibitors and as pheromones that elicit physiological and behavioural responses in the post-mated female. Understanding the molecular interactions by which these peptides influence reproduction requires detailed knowledge of their molecular structures. However, this information is often lacking and cannot be gleaned from just gene sequences and standard proteomic data. We now report the native structures of four seminal fluid peptides (andropin, CG42782, Met75C and Acp54A1) from the ejaculatory duct of male *D. melanogaster.* The mature CG42782, Met75C and Acp54A1 peptides each have a cyclic structure formed by a disulfide bond, which will reduce conformational freedom and enhance metabolic stability. In addition, the presence of a penultimate Pro in CG42782 and Met75C will help prevent degradation by carboxypeptidases. Met75C has undergone more extensive post-translational modifications with the formation of an N-terminal pyroglutamyl residue and the attachment of a mucin-like O-glycan to the side chain of Thr_4. Both of these modifications are expected to further enhance the stability of the secreted peptide. The glycan has a rare zwitterionic structure comprising an O-linked N-acetyl hexosamine, a hexose and, unusually, phosphoethanolamine. A survey of various genomes showed that *andropin, CG42782,* and *Acp54A1* are relatively recent genes and are restricted to the *melanogaster* subgroup. *Met75C,* however, was also found in members of the *obscura* species groups and in *Scaptodrosophila lebanonensis*. *Andropin* is related to the cecropin gene family and probably arose by tandem gene duplication, whereas *CG42782, Met75C* and *Acp54A1* possibly emerged *de novo*. We speculate that the post-translational modifications that we report for these gene products will be important not only for a biological function, but also for metabolic stability and might also facilitate transport across tissue barriers, such as the blood-brain barrier of the female insect.

Keywords: Seminal fluid peptides, novel genes, andropin, CG42782, Met75C, Acp54A1, male accessory organs, ejaculatory duct, *Drosophila melanogaster*, glycopeptide, phosphoethanolamine

1. Introduction

For several decades, the polyandrous *Drosophila melanogaster* with its amenable genetics and readily accessible bioinformation, has proven to be an excellent animal model to study the physiological role and evolutionary significance of seminal fluid (SF) proteins/peptides (SFPs)[1, 2]. Such studies have provided genetic and molecular insights into the role of SFPs in areas such as securing paternity, manipulation of female behaviour, sexual conflict, sperm storage, sperm competition and postcopulatory sexual selection [2-4]. The male *Drosophila* accessory sex glands, responsible for the synthesis and secretion of most of the SFPs, include the paired male accessory glands (MAGs) or paragonia, the ejaculatory duct (ED) and the ejaculatory bulb (EB) [5, 6]. The MAGs are relatively large elongated sacs that open below the vasa deferentia into the distended section of the ED that connects to the seminal vesicles. The wall of the MAG includes two types of secretory cells: a single layer of cuboidal binucleated cells with larger vacuolated secondary cells interspersed towards the apical region. The ED, consisting of a single layer of large glandular polygonal cells, tapers as it approaches and traverses the EB and discharges in the aedeagus (equivalent to penis). The EB is a thick-walled muscular organ that contains fluid of a composition different to the ED[5]. The ED is capable of strong peristaltic contractions, which presumably forces the MAG and ED gland secretions together with sperm into the EB.

Several studies have used mass spectrometry-based proteomic analysis and transcriptomic data of the *Drosophila* MAGs, ED and EB to identify candidate SFPs [7-10]. Confirmation that many of these are actually transferred to the female on mating was obtained by using isotope labelling of proteins to distinguish between male and female derived molecules in the female reproductive tract after mating [8]. In addition, indirect evidence of transfer has been obtained from quantitative proteomics of male reproductive glands before and after mating [10]. We can conclude from the proteomic studies and also from high-throughput expression data that many of the prominent SFPs are expressed either exclusively in the male reproductive tract or are at least highly enriched in these tissues. Proteomics of soluble proteins from separated male tissues, together with cell localisation data from *in situ* hybridisation, tagged protein expression in transgenic flies and immuno-histochemical studies, indicate that the MAGs, ED and EB all contribute to the synthesis of SFPs [9, 10].

Only a few of the SFPs of *D. melanogaster* have been biochemically characterised (e.g. [7]). For other SFPs, a biochemical function can sometimes be inferred from sequence homology to well characterised proteins. [1]. A number of SFPs can be classified as enzymes,

(e.g. proteases and lipases), binding proteins (e.g. lectins) or mating plug components [2, 10]. Another significant subset of SFPs are smaller proteins or peptides (molecular mass of <15 kDa) that include pheromones, antimicrobials, protease inhibitors as well as odorant binding proteins [1, 2]. The best known of these molecules is the sex peptide (SP), a 36-mer peptide pheromone that is responsible for eliciting multiple physiological and behavioural responses in the post-mated female, including a reluctance to re-mate for several days and increased oviposition [11]. In contrast, other *Drosophila* SF peptides are poorly characterised, both structurally and physiologically.

Proteomics, involving analysing the masses of tryptic peptide fragments, has revolutionised the way we identify and, in some instances, quantify protein products of the male reproductive glands [12]. These approaches, however, are primarily used to identify proteins and are not suited for the structural characterisation of native SFPs. The reliance on convenient tryptic cleavage sites and the fact that many SFPs are post-translationally modified, conspire to hinder the acquisition of structural information [13]. Furthermore, some proteomic methods involve the fractionation of proteins prior to proteolysis, which runs the risk of omitting small proteins and peptides from the analysis.

In this paper we report the application of peptidomics, a widely used method for the analysis of regulatory peptides in insect nervous and endocrine tissues, to investigate the structure of SF peptides originating from the ED of male *D. melanogaster*. This approach preferentially extracts peptides in their native form, whilst inhibiting endogenous protease activity during the extraction procedure and uses trypsin only selectively after peptide identification [14-16]. We have structurally characterised four major ED peptides with a molecular mass of < 4 kDa that are destined for the SF. All four are taxon-restricted and products of novel genes. One prominent ED peptide is a highly post-translationally modified product of two genes *Met75Ca* and *Met75Cb*. The mature form of Met75C is a 30-mer peptide with a rare zwitterionic O-glycan. We speculate that the post-translational modifications we describe are likely to be important not only for a biological function, but also for extracellular stability and might also facilitate transport across tissue barriers, such as the blood-brain barrier of the female insect.

2. Methods

2.1. Insects

The Dahomey strain of *D. melanogaster* was provided by Professor T. Chapman (University of East Anglia, U.K.). The wild-type Canton-S and w^{1118} strains were from Professor Y-J Kim (Gwangju Institute of Science and Technology, Gwangju 500-712, Republic of Korea). Insects were maintained on oatmeal/molasses/agar medium at 25ºC in a 12:12 light-dark cycle.

2.2. Preparation of tissue samples for mass spectrometry

Tissues were dissected from 3-8-day-old unmated males in cold insect saline containing 150mM NaCl, 10mM KCl, 3.9mM $NaHCO_3$, 3.5mM $MgCl_2$, 1.3mM $CaCl_2$ adjusted to pH 7.2. Individual tissue samples were rinsed in purified water and subsequently placed on a MALDI sample plate for direct tissue profiling. Alternatively, 10 organs were transferred into a vial containing 10 µl of extraction solution (50% aqueous methanol containing 0.1% formic acid (FA)). Tissue collections were sonicated in a water bath and subsequently centrifuged for 15 min at 13,000 rpm. Supernatants were transferred into clean vials and stored at -20ºC until required for analysis.

2.3. Reduction, alkylation, tryptic digestion, sulfonation and purification of peptides

Peptides were subjected to cystine reduction by dithiothreitol (Sigma-Aldrich Company Ltd. Gillingham, Dorset, U.K.) and alkylation by iodoacetic acid (Sigma-Aldrich Company Ltd.) followed by enzymatic digestion using trypsin (Sequencing Grade Modified Trypsin, Promega U.K. Ltd., Southampton, U.K.) as described previously [14]. Briefly, tissue supernatants were concentrated by using a SpeedVac, and adjusted to a volume of 25 µl with 50 mM ammonium bicarbonate (ABC) buffer, pH 8.2. Disulfide reduction was performed by adding 200 mM dithiothreitol (DTT) in ABC buffer to a final concentration of 10 mM DTT at 37 °C for 1 h. Carbamidomethylation of reduced cysteines was performed by adding 200 mM iodoacetamide (IAA) in ABC buffer to a final concentration of 40 mM IAA at room temperature for 1 h. Unreacted IAA was inactivated/precipitated by adding DTT at room temperature to a final concentration of 40 mM DTT for 15 min. Extracts with reduced and carbamidomethylated peptides were adjusted to 100 µl of 0.5 % aqueous FA prior to desalting and purification. For digestion of proteins, 0.1 µl of a 1 µg/µl solution of trypsin in 50 mM acetic acid (Sequencing Grade Modified Trypsin, Promega) was added to the reduced and alkylated extract and incubated at 37 °C for 16 h.

.

For *de novo* sequencing, peptides were sulfonated using 4-sulfophenyl isothiocyanate (SPITC; Sigma-Aldrich Company Ltd.) following the protocol of Sturm *et al*. [15]. Briefly, 25 µl of a 10 mg/ml SPITC solution in 20 mM $NaHCO_3$ (pH 9.0) was added to 25 µl of a digested extract. The reaction mixture was incubated in a thermo-mixer (Eppendorf, Hamburg, Germany) at 55 °C and 300 rpm for 1 h. After incubation the reaction was terminated by adding 10 µl of 1% aqueous FA.

The reduced and alkylated, or additionally digested and SPITC derivatized peptides were purified and desalted using custom made C18 spin columns using cut-outs of C18 Empore 3M extraction discs (3M, USA; gifted by IVA Analysentechnik e.K., Meerbusch, Germany) packed into 200 µl pipette tips (Eppendorf, Germany) as described in [17]. The sorbent was activated with 100 µl of 80% aqueous acetonitrile (ACN) containing 0.1% FA and equilibrated with 100 µl 0.1% aqueous FA by centrifuging the spin columns at 2000 rpm. Samples were diluted with 100 µl of 0.1% aqueous FA, loaded on the spin column and washed with 200 µl of 0.1% aqueous FA. Peptides were eluted with 5 µl of 80% aqueous ACN containing 0.1% FA step-by-step directly onto the MALDI sample plate using a modified 20 ml plastic syringe.

2.5. Matrix-assisted laser desorption ionisation mass spectrometry (MALDI-TOF MS)

Mass spectra were acquired in positive reflector mode using Bruker ultrafleX III or ultrafleXtreme mass spectrometers (Bruker Daltonics, Bremen, Germany). Samples were mixed with 0.5 µl 2,5-Dihydroxybenzoic acid (DHB, 10 mg/ml in 20% acetonitrile, 1% FA) on the MALDI sample plate and dried using a gentle stream of hot air. External calibration was conducted using calibration mixtures containing bradykinin$^{1\text{-}7}$, angiotensin I, angiotensin II, substance P, bombesin, ACTH clip$^{1–17}$, ACTH clip$^{18–39}$, somatostatin 28, insulin and ubiquitin I (Bruker Daltonics).

2.6. Acetylation of amines

The reaction mixture was freshly prepared by mixing acetic anhydride (Sigma-Aldrich Company Ltd.) and methanol (Fisher Chemicals U.K., Loughborough, U.K.) in a ratio of 1:3. A small droplet of the solution was added to cover the surface of the dried disrupted tissue on the MALDI sample plate and rapidly evaporated with a hair dryer after 30 seconds. Subsequently 0.5 µl of DHB matrix solution was added on the heated sample and left to dry.

2.7. Fourier Transform Ion Cyclotron Resonance mass spectrometry (FT-ICR MS)

A FT-ICR MS analysis was performed using a solariX XR FT MS (Bruker Daltonics) with a 9.4 T superconducting magnet. Peptide solutions were diluted 1:20 into 50% aqueous acetonitrile containing 1% FA before introduction by TriVersa NanoMate (Advion BioSciences, Ithaca, NY) in positive-ion mode. The applied voltage was adjusted between 1.4-1.7 kV to achieve a stable ion current. A 120°C nitrogen dry gas was supplied at 1.3 L/min to aid desolvation. Instrument control and data acquisition used Compass 1.4 (Bruker Daltonics). Spectra were generated by the accumulation of 20 scans with 0.2 s ion cooling time and 0.5 s scan time with 400K data points recorded. Peptide precursors were manually selected for isolation and subsequent fragmentation by collision induced dissociation in the hexapole (Q-CID) with argon as the collision gas. Collision energies were optimized for each peptide. Spectra were processed using DataAnalysis version 4.0 (Bruker Daltonics). Mass deconvolution was performed using version 2.0 of the SNAP averaging algorithm (C 4.9384 %, N 1.3577 %, O 1.4773 %, S 0.0417 %, H 7.7583 %).

2.8 Bioinformatics

Sequence databases were searched using BLASTP and TBLASTN algorithms [18] (http://blast.ncbi.nlm.nih.gov/Blast.cgi). Nucleotide sequences were translated using ExPASy Translate (Swiss Institute of Bioinformatics; http://web.expasy.org/translate/). Retrieved protein sequences were submitted to the SignalP-5.0 server (http://www.cbs.dtu.dk/services/SignalP/) to establish the occurrence of a signal peptide for secretion and the probable cleavage site for signal peptidase [19]. Peptide sequences were aligned using ClustalOmega [20] (https://www.ebi.ac.uk/Tools/msa/clustalo/). The predicted cleavage sites of the signal peptidase differed between the Met75C orthologs resulting in varying lengths of the signal peptide and the propeptide. As a result of this, homologous sequence position may be part of the signal peptide in one species, but of the propeptide in another. Therefore, we manually adjusted and optimized the alignment for Met75C sequence logos to provide a consensus signal and propeptide. Sequence logos were created with Weblogo Version 3.7.4 (http://weblogo.threeplusone.com) using equiprobable composition [21].

A maximum likelihood phylogeny was calculated with MEGA X[22] using the LG model with gamma distribution and five categories [23]. Percentage similarities and identities of

sequences were calculated using MatGAT 2.01 [24] applying the in-build alignment algorithm using default settings and the BLOSUM50 substitution matrix.

3. Results

3.1 Peptidomics of the ejaculatory duct

We initially used MALDI-TOF MS to investigate the peptide inventory of tissues associated with the male reproductive tract in three different strains of *D. melanogaster*. The male accessory glands (MAG), dilated and narrow regions of the anterior ejaculatory duct (ED), ejaculatory bulb (EB), posterior ED, testis, and seminal vesicle (SV) of individual flies were analysed. The spectra for the different regions of the ED (Fig. 1) feature a common pattern of ion signals that are absent in tissue samples from the MAG, the EB, SV and testis (data not shown). We found highly similar and reproducible patterns, in the laboratory strains w^{1118}, Dahomey and Canton-S, only varying in the relative abundance of the ion signals between different individuals. To identify the structures of the detected ion signals, individual tissue sections and a peptide extract of 10 EDs that had been treated with DDT and IAA to reduce and alkylate cysteines, were subjected to MALDI-TOF MSMS. The same peptide extract was also analysed by FT-ICR MS. Ion signals in the MALDI-TOF MS spectra of the ED were identified as peptide products of the following genes: *CG34098* (*Acp4A1*; Supplementary Fig. 1), *CG42782* (Supplementary Fig. 2), *CG1361* (*andropin*; Supplementary Fig. 3), *CG32197* and *CG18064* (*Met75Ca* and *Met75Cb*; Fig. 2). All five genes code for proteins that are predicted to have a secretory signal peptide and a signal peptidase cleavage site. The ion signal of andropin was found to be consistently lower than the other three gene products. Surprisingly, three distinct ion signals at *m/z* 3451.3, *m/z* 3816.4 and *m/z* 3939.7 (Fig. 1b) were all identified as peptide forms encoded by *Met75Ca* and *Met75Cb* (Table 1). The sequences of the predicted prepropeptide products of *Met75Ca* and *Met75Cb* are identical and, hereafter, are referred to as Met75C. A literature search indicated that the mass differences between Met75C-derived ion signals and the fragmentation pattern of the peptide ion at *m/z* 3939.7 (Fig. 2) is likely due to the presence of a post-translational modification comprising an N-acetyl hexosamine (HexNAc), a hexose (Hex) and a phosphoethanolamine (PEA) [25, 26]. The ED peptide Dup99B reported by Saudan *et al.*[27] was not detected in any of the MALDI-TOF MS analyses. However, a mass match corresponding to the glycosylated form of Dup99B was found in the FT-ICR MS analysis. Table 1 summarises the detected ion signals and the deduced sequences for the native and alkylated peptides found in the ED by MALDI-TOF MSMS and FT-ICR MS.

3.2. Glycosylation of Met75C

In order to determine the glycosylated residue, the reduced and alkylated extract of 10 EDs was digested with trypsin and subsequently derivatized with 4-sulfophenyl isothiocyanate (SPITC), a procedure that sulfonates N-termini and facilitates sequencing due to the generation of abundant fragments of the *y*-ion series in MALDI-TOF MSMS. Figure 3 shows a MALDI-TOF mass spectrum of the derivatised ED tryptic peptides and the sequences of selected peptides confirmed by MSMS. Two ion signals at *m/z* 774.3 and *m/z* 936.4 were identified as glycosylated tryptic peptides derived from the N-terminus of Met75C (Fig. 3). Fragmentation of the ion signal at *m/z* 774.3 (data not shown) resulted in a complex pattern due to the co-fragmentation of the HexNAc-modified N-terminal tryptic peptide (Fig. 3 a_1) with the C-terminal tryptic peptide of Met75C (Fig. 3 a_2). Fragmentation of the ion signal at *m/z* 936.4 (Fig. 3), however, confirmed the glycosylation comprising a distal hexose and a proximal N-acetyl hexose (Fig. 4) attached to a residue in the N-terminal pentapeptide sequence (pQIATR). Thr_4 and Arg_5 are possible sites of attachment for an O-linked and N-linked glycan, respectively. However, glycosylation of an N of Arg is unlikely since this post-translational modification has been confirmed only for bacteria and the presence of the glycan structure at this position would likely interfere with tryptic cleavage of the Arg-Gln peptide bond [28]. We therefore conclude that Thr_4 is the glycosylation site.

3.3. Glycan substitution with phosphoethanolamine (PEA)

Phosphoethanolamine (PEA) is a zwitterionic structure exhibiting a negative charge at the glycosidic-attached phosphate group and a distal positive charge at the amine. In order to confirm this structure, we performed on-plate derivatisation of single ED tissues with a solution of acetic anhydride. Acetic anhydride acetylates primary amines resulting in a mass shift of +42 Da for free N-termini, each lysine and PEA. Acetylation resulted in the expected number of mass shifts, namely, three times +42 Da for the Acp54A1 and two times +42 Da for CG42782 (Fig. 5). The glycosylated Met75C, which has one Lys at position 12, but no free amino group at the N-terminus because of the cyclic N-terminal pyroglutamic acid, featured two mass shifts of +42 Da after derivatisation, which indicates the presence of an additional primary amine within the glycan structure.

3.4. Proposed glycan structure

We sought further clarification of the glycan structure using high-resolution FT-ICR MS, but the fragmentation of the Met75C parent ion using electron-capture dissociation or collision induced dissociation did not provide evidence of the location of the glycan. However, low energy qCID fragmentation (12 V) resulted in the multi-charged fragments *m/z* 779.54434 (z=5) and *m/z* 974.17824 (z=4).These ions correspond to Met75C with the remaining HexNAc-PEA and a loss of a hexose (deconvulated z=1; 3893.69259 and 3893.69113) respectively. This suggests that both the hexose (Hex) and the PEA are attached to the N-acetylhexosamine (HexNAc) rather than in a linear manner (Fig. 6).

3.5. Evolutionary history of the ejaculatory duct peptides

To investigate the evolutionary history of the ED peptides, we screened publicly available genome and transcriptome databases for sequences related to *Acp51A1 (CG34098)*, *CG42782*, *andropin* and *Met75C*. Our analysis shows that *andropin* orthologs are present in the *melanogaster* subgroup as well as the *eugracilis*, *takahashii* and *suzukii* subgroups (Fig. 7) and, as noted previously, the predicted mature peptides have diverged with non-identical residues replaced by conservative substitutions [29] (Supplementary material: Andropin_alignment.fas).

CG42782 orthologues were only found in members of the *melanogaster* subgroup (Fig. 7). These orthologues have the two absolutely conserved Cys residues that form the disulfide bridge, as well as four other conserved residues within the ring structure (Supplementary material: CG42782_alignment.fas). In total, seven out of ten residues of the peptide ring are either conserved or are substituted by amino acids with similar side chain properties. The predicted C-termini of the peptides, however, show structural divergence with four (*D. erecta, D. simulans, D. sechellia* and *D. mauritania*) out of six predicted to have a C-terminal Gly, which we expect to be utilised to convert the carboxyl group of the penultimate residue to an amide.

Orthologs of *Acp51A1* (*CG34098*) have been found in *D. simulans*, *D. sechellia* and *D. yakuba* of the *melanogaster* subgroup, but not in *D. erecta*. Additional *Acp51A1* orthologs could also be assigned to two species outside of the *melanogaster* subgroup, namely *D. biarmipes and D. takahashi* (Fig. 7). The ring forming Cys residues are absolutely conserved as are two Gly residues within the cyclic structure (Supplementary material: Acp54A1_alignment.fas).

In our database screening for *Met75C*-related genes, we found numerous orthologs in members of the subgenus *Sophophora* , as well as multiple paralogs in some of these species. However, we could not find *Met75C*-related genes in *D. willistoni* and the subgenus *Drosophila* (e.g., *D. albomicans, D. grimshawi, D. mojavensis* and *D. virilis*; Fig. 7). Nevertheless, we detected an ortholog in *Scaptodrosophila lebanonensis* that had 51.0% amino acid similarity and 42.6% identity with *D. melanogaster* Met75C. *D. melanogaste*r has two paralogous genes (*CG18064, Met75Cb* and *CG32197*, *Met75Ca*), encoding identical Met75C prepropeptides, that are adjacent to each other and separated by 2611 bp on the left arm of chromosome 3. Further species in which we found paralogs are *D. suzukii, D. subpulchrella, D. takahashii, D. rhopaloa, D. leontia, D. bocki, D. kikkawai, D. nikananu, D. bipectinata, D. azteca* and *D. athabasca,* each with two paralogs as well as *D. sechellia and D. subobscura* containing three. Phylogenetic analysis of the sequences suggests that most of these paralogs arose from independent gene duplication events with the exception of a common duplication in the ancestral line of *D. leontia*, *D. bocki*, and *D. kikkawai*, as well as a common evolutionary event shared by *D. azteca* and *D. athabasca* (Supplementary Fig.4). Interestingly, duplication in these clades leads to a diversification of one daughter gene while the other one remains more ancestral (see corresponding percent similarity/identity Supplementary Fig.4.). The average distance between paralogs is around 1800 bp with the exception of *D. azteca* and *D. athabasca*, where the paralogs are located on different chromosomes, namely chromosome 2 and XR as a result of chromosomal rearrangement [30]. There is a high degree of sequence conservation amongst the Met75C orthologues at the N-terminal glutamine, the two cysteines, a Gly and several Arg and Tyr residues within the ring structure (Fig. 8). Notably, there is a common C-terminal sequence comprising conservative hydrophobic residues on either side of a Pro. The Thr in *D. melanogaster* Met75C to which a glycan is attached, is also conserved in the majority of the predicted peptides.

4. Discussion

Proteomic studies that exploit the prowess of *Drosophila melanogaster* as a model organism have, for some time, provided a wealth of information on the protein composition and chemical complexity of SF. Clearly, structural characterisation of individual SFPs is critical for advancing our understanding of the molecular mechanisms by which these molecules influence reproduction [7-9]. This aspect has often been neglected in proteomic studies of the reproductive tissues of *Drosophila*, which had the primary objective of identifying the presence or absence of a soluble protein in a tissue extract. By applying a

peptidomic strategy that has been used widely to characterise peptides of the insect nervous and endocrine systems, we have determined the native structures of four SF peptides extracted from the *D. melanogaster* ED. Of particular interest is the revelation that Met75C is O-glycosylated, a peptide modification that is not uncommon in *Drosophila.* It is found in several proline-rich antibacterial peptides (drosocin and maturated Pro-domain of attacin C) and the ED peptide, DUP99B [56, 27, 61, 62]. Drosocin occurs as both single and doubly glycosylated forms and glycosylation is essential for full antibacterial activity [56].

CG42782 encodes a 46-mer prepropeptide that is predicted to be processed by cleavage of the signal peptide at the Ser-Tyr bond to generate a 23-mer mature secreted peptide. The molecular ion $[M+H]^+$ at *m/z* 3034.0 and its fragmentation pattern confirms this prediction and establishes that the peptide forms a ring structure through the formation of a disulphide bond between Cys^7 and Cys^{18}. The presence of a penultimate Pro residue is expected to protect the peptide from general carboxypeptidase attack and the cyclic peptide is also expected to hinder degradation by proteolytic enzymes [31, 32]. We found no evidence that *CG42782* is present in the genomes outside of the *melanogaster* subgroup (Fig.7).

The mature Acp54A1 (CG34098) peptide found in the ED starts with a methionine, which is consistent with the predicted signal cleavage site at the Ala-Met bond. The peptide comprises 17 amino acids, 12 of which form a ring by virtue of a disulphide between Cys_3 and Cys_{12}, which will reduce vulnerability to peptidase attack [32]. Orthologs of *Acp54A1* were found in members of the *melanogaster* subgroup and also in *D. biarmipes* and *D. takahashii* (Fig. 7). The presence of *Acp54A1* in *D. yakuba* and the absence of the gene from *D. erecta* is discordant with a previous report that the gene could not be found in *D. yakuba* [33]. The failure to identify an Acp54A1 orthologue in *D. erecta* might be explained by gene loss in this species.

The andropin molecular ion $[M+H]^+$ at *m/z* 3749.0 (Supplementary Fig. 3) matched the predicted mass of the mature 34 residue peptide, generated after cleavage of the signal peptide at the Ala-Val bond and established the absence of any further post-translational modification. Prior studies have shown that *Andropin* expression is male specific and is restricted to the ED. *Andropin* is closely linked to the *Cecropin* gene family on the right arm of chromosome 3 and likely to have arisen from duplication of a common ancestral *Cecropin* gene with subsequent divergence in both sequence and tissue expression [34]. The only sequence similarity between the *Andropin* and *Cecropin* protein products lies in the signal

peptide of the preproprotein [29]. The synthetic andropin peptide has moderate anti-bacterial activity towards Gram-positive bacteria, but, unlike the cecropin peptides, is not active against Gram-negative bacteria [34]. It has been reported that the ED is the source of a second antibacterial compound in addition to andropin. It is conceivable that one or more of the other ED [35].

Met75C is a 30-mer peptide encoded by two intron-less *D. melanogaste*r genes (*CG18064, Met75Cb* and *CG32197*, *Met75Ca*) that are tandem duplicates. Met75C was identified previously as an ED peptide from tryptic fragment mass analysis and the peptide has been confirmed as a *bona fide* SF component that is passed to the female reproductive tract during copulation [8, 10]. We confirm the prediction that the signal peptide is cleaved at the Ala-Gln bond and reveal that the glutamine is cyclised to pyroglutamate in the mature peptide. The C-terminal half of the peptide, like both CG42782 and Acp51A1, is a ring structure formed by a disulphide bond between Cys_{14} and Cys_{27}. Of special interest, is the discovery that Met75C is decorated with an O-glycan attached to Thr_4 in the N-terminal sequence pQIATR. Our MS analysis predicts that the glycan comprises an N-acetylhexosamine and that this sugar is extended by a single hexose and the zwitterionic phosphoethanolamine. This type of modification, generating a positively and negatively charged glycan, is rare, but has been described previously for O-linked glycosylations in the nest material of the wasp, *Vespula germanica* and the glycopeptide, noctilisin, in the venom gland of *Sirex noctilio* [36, 37] (Fig.6). Phosphoethanolamine has also been reported for N-linked glycans in extracts of the royal jelly, the venom gland and homogenates of larvae of the honey bee, *Apis mellifera* [38] as well as for glycosphingolipids of the pupae of the blow fly *Calliphora vicina* [39, 40].

The most common O-glycosylation of secreted proteins in *D. melanogaster* is of the mucin-like O-glycan type that is initiated by the attachment of N-acetylgalactosamine (GalNAc) to the side chain of either Thr or Ser [41, 42, 62]. This reaction is catalysed by polypeptide N-acetylgalactosaminyltransferases (PGANTs), of which there are 14 in *D. melanogaster* [41, 42]. This family of type II membrane enzymes are resident in the Golgi and can display different substrate specificities and expression patterns. Transcriptomic data has shown that the MAG/ED tissues strongly express the genes encoding PGANT4 and PGANT9, suggesting that one or both enzymes are involved in initiating O-glycan synthesis in these tissues [43].

A common extension to the GalNAc is the addition of galactose in a β-1,3-linkage carried out by core β-1,3-galactosyltransferases (C1GalTs)[44]. In *Drosophila*, there are 11 genes for this family of transferases and of these, two are strongly and exclusively expressed in MAG/ED tissues [43]. Since it is estimated that around 90% of *D. melanogaster* O-glycans are of the mucin-type and given the strong expression of the aforementioned transferases in the male reproductive tract, it is very likely that this biosynthetic pathway is responsible for the O-glycosylation of Met75C.

A comparison of the predicted primary structures of the Met75C peptides from various *Drosophila* species in the subgenus *Sophophora* reveals that about half of the amino acids are highly conserved, whereas significant amino acid divergence has occurred in other positions both within and outside the peptide ring. Two plausible, and not necessarily mutually exclusive explanations for the conservation, are constraints imposed by structural and conformational requirements for a presumed physiological role in reproduction and/or the need to protect the integrity of the peptide from peptidases, not only in the seminal fluid but also in the female [45]. The reduced conformational freedom resulting from peptide cyclisation can confer specificity to molecular interactions, as well as enhanced stability [32]. An N-terminal glutamine that cyclises to pyroglutamate, the peptide ring and Pro residues are typical features that can provide protection from peptidases [31, 46]. The glycan with its zwitterionic phosphoethanolamine provides a charged carbohydrate shield, which is expected to provide protection of the N-terminal section of Met75C from proteolysis.

Met75C is not the only glycosylated peptide secreted by the ED of *D. melanogaster*. DUP99B is a 31-mer peptide that has a branched N-linked glycan, comprising two molecules of N-acetylglucosamine, three of mannose and two of fucose, attached at Asn_4. DUP99B is a paralogue of the MAG sex peptide (SP), and like SP, can elevate oviposition and inhibit male receptivity when injected into virgin females [27, 47]. A synthetic non-glycosylated form of DUP99B is also able to activate the SP receptor expressed in mammalian cells, albeit with a 10-fold reduction in potency [48]. Interestingly, the presence of the glycan in the natural DUP99B increases the potency of injected peptide in eliciting post-mating female responses[27], suggesting that the glycan influences receptor binding and/or has an *in vivo* role facilitating ligand access to the female SP receptor. Although there is no sequence homology between Met75C and DUP99B, there are compelling structural analogies in that both peptides have a pyroglutamyl N-terminus followed by a short peptide carrying a glycan, which is then separated by a linker sequence from a peptide ring of similar size that

dominates the C-terminal half of both peptides. It has been reported that DUP99B can attach to sperm, perhaps by virtue of the N-linked glycan [27]. Whether the zwitterionic glycan moiety of Met75C facilitates a similar adhesion to *D. melanogaster* sperm proteins or binding to other components of the ejaculate, such as the lectin-like SF proteins, is not known, but worthy of investigation. There is no functional information available for Met75C, but the aforementioned structural comparison with a biologically active DUP99B invites speculation that Met75C might serve as a pheromonic peptide that signals via the female central nervous system to influence the physiology or behaviour of the post-mated female. It is worth noting that glycosylation of mammalian neuroactive peptides is known to not only confer metabolic stability, but also increase penetration of the blood-brain barrier, which has encouraged the development of glycopeptide drugs that target the central nervous system [49, 50].

All five of the genes discussed here are taxon-restricted, without detectable homologs in non-Drosophilidae species and belong to a class commonly referred to as 'new' or 'young' protein-coding genes [51, 52]. Met75C is the most ancient of these five genes and was found in *Scaptodrosophila lebanonensis,* which diverged from *Drosophila melanogaster* about 70-74 Mya [53, 54]. The remaining genes appear to have originated in the *Sophophora* subgenus after split of the *willistoni* species group from the *obscura* and *melanogaster* lineages. The most recent genes are CG42782 and Acp54A1, estimated to have emerged about 8.8-12.8 Mya and 22.3 - 35.6 Mya, respectively[53, 55]. There are several known mechanisms for new-gene origination with gene duplication recognised as the dominant source [52]. *Andropin* is most probably a product of a duplication of an ancestral cecropin gene, but the origins of *CG42782, Acp54A1* and *Met75Ca/b* are unclear. It has been suggested that such SF peptide genes might have arisen *de novo* from non-coding DNA, since the intergenic and intronic DNA of *D. melanogaster* harbors thousands of small short open reading frames (ORF) that have the potential to generate secreted peptides [51]. If, during the course of evolution, these ORFs recruit appropriate regulatory elements to drive expression in the male reproductive tissues, then these *de novo* genes could rapidly evolve lineage-specific roles in reproductive physiology. In addition to the structural requirement of a signal peptide for translocation to the secretory pathway, the new peptide will also benefit functionally from features, like those described for Acp54A1, CG42782 and Met75C, that will reduce susceptibility to degradation by extracellular peptidases that the peptides will encounter in the ejaculate and in female reproductive tissues.

Credit authorship contribution statement

S. Sturm: Methodology, Investigation, Formal analysis, Writing- Original draft preparation, Writing- Reviewing and Editing. **A. Dowle**: Data collection, Formal analysis, Writing- Reviewing and Editing. **R.E. Isaac**: Conceptualization, Writing- Original draft preparation, Writing- Reviewing and Editing. **N. Audsley**: Conceptualization, Writing- Reviewing and Editing.

Acknowledgements

The York Centre of Excellence in Mass Spectrometry was created thanks to a major capital investment through Science City York, supported by Yorkshire Forward with funds from the Northern Way Initiative, and subsequent support from EPSRC (EP/K039660/1; EP/M028127/1). We thank Reinhard Predel (University of Cologne) for providing access to the MALDI-TOF MS. REI and NA thank the Leverhulme Trust for support (RPG-2020-368). SS and REI thank Carole Sowden for technical assistance and Divya Ramesh for critical reading of the manuscript.

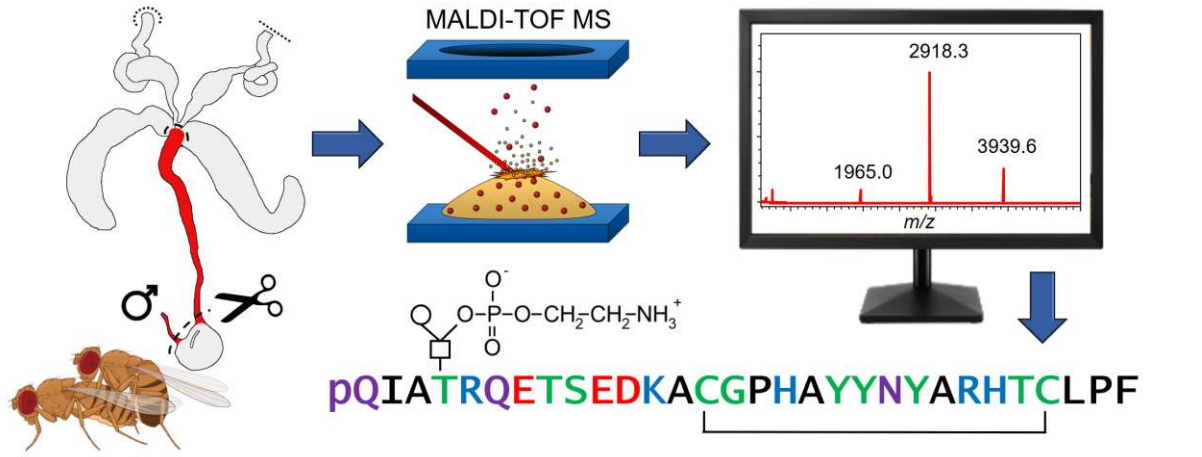
References

1. Avila, F.W., et al., *Insect seminal fluid proteins: identification and function.* Annu Rev Entomol, 2011. **56**: p. 21-40.
2. Wigby, S., et al., *The Drosophila seminal proteome and its role in postcopulatory sexual selection.* Philos Trans R Soc Lond B Biol Sci, 2020. **375**(1813): p. 20200072.
3. Sirot, L.K., et al., *Sexual Conflict and Seminal Fluid Proteins: A Dynamic Landscape of Sexual Interactions.* Cold Spring Harbor Perspectives in Biology, 2015. **7**(2).
4. Ramm, S.A., *Seminal fluid and accessory male investment in sperm competition.* Philos Trans R Soc Lond B Biol Sci, 2020. **375**(1813): p. 20200068.
5. Miller, A., *The internal anatomy and histology of the imago of Drosophila melanogaster*, in *Biology of Drosophila*, M. Demerec, Editor. 1950, John Wiley & Sons: London. p. 420-534.
6. Chen, P.S., *The Functional-Morphology and Biochemistry of Insect Male Accessory-Glands and Their Secretions.* Annual Review of Entomology, 1984. **29**: p. 233-255.
7. Walker, M.J., et al., *Proteomic identification of Drosophila melanogaster male accessory gland proteins, including a pro-cathepsin and a soluble gamma-glutamyl transpeptidase.* Proteome Sci, 2006. **4**: p. 9.
8. Findlay, G.D., et al., *Proteomics reveals novel Drosophila seminal fluid proteins transferred at mating.* PLoS Biol, 2008. **6**(7): p. e178.
9. Takemori, N. and M.T. Yamamoto, *Proteome mapping of the Drosophila melanogaster male reproductive system.* Proteomics, 2009. **9**(9): p. 2484-93.
10. Sepil, I., et al., *Quantitative Proteomics Identification of Seminal Fluid Proteins in Male Drosophila melanogaster.* Mol Cell Proteomics, 2019. **18**(Suppl 1): p. S46-S58.

11. Chen, P.S., et al., *A male accessory gland peptide that regulates reproductive behavior of female D. melanogaster.* Cell, 1988. **54**(3): p. 291-8.
12. Findlay, G.D. and W.J. Swanson, *Proteomics enhances evolutionary and functional analysis of reproductive proteins.* Bioessays, 2010. **32**(1): p. 26-36.
13. Fricker, L.D., et al., *Peptidomics: identification and quantification of endogenous peptides in neuroendocrine tissues.* Mass Spectrom Rev, 2006. **25**(2): p. 327-44.
14. Sturm, S. and R. Predel, *Mass spectrometric identification, sequence evolution, and intraspecific variability of dimeric peptides encoded by cockroach akh genes.* Anal Bioanal Chem, 2015. **407**(6): p. 1685-93.
15. Sturm, S., et al., *Agatoxin-like peptides in the neuroendocrine system of the honey bee and other insects.* J Proteomics, 2016. **132**: p. 77-84.
16. Sturm, S., et al., *The structure of the Drosophila melanogaster sex peptide: Identification of hydroxylated isoleucine and a strain variation in the pattern of amino acid hydroxylation.* Insect Biochem Mol Biol, 2020. **124**: p. 103414.
17. Rappsilber, J., M. Mann, and Y. Ishihama, *Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips.* Nat Protoc, 2007. **2**(8): p. 1896-906.
18. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
19. Almagro Armenteros, J.J., et al., *SignalP 5.0 improves signal peptide predictions using deep neural networks.* Nat Biotechnol, 2019. **37**(4): p. 420-423.
20. Sievers, F. and D.G. Higgins, *Clustal Omega for making accurate alignments of many protein sequences.* Protein Sci, 2018. **27**(1): p. 135-145.
21. Crooks, G.E., et al., *WebLogo: a sequence logo generator.* Genome Res, 2004. **14**(6): p. 1188-90.
22. Kumar, S., et al., *MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms.* Mol Biol Evol, 2018. **35**(6): p. 1547-1549.
23. Le, S.Q. and O. Gascuel, *An improved general amino acid replacement matrix.* Mol Biol Evol, 2008. **25**(7): p. 1307-20.
24. Campanella, J.J., L. Bitincka, and J. Smalley, *MatGAT: an application that generates similarity/identity matrices using protein or DNA sequences.* BMC Bioinformatics, 2003. **4**: p. 29.
25. Conboy, J.J. and J.D. Henion, *The determination of glycopeptides by liquid chromatography/mass spectrometry with collision-induced dissociation.* J Am Soc Mass Spectrom, 1992. **3**(8): p. 804-14.
26. Huddleston, M.J., M.F. Bean, and S.A. Carr, *Collisional fragmentation of glycopeptides by electrospray ionization LC/MS and LC/MS/MS: methods for selective detection of glycopeptides in protein digests.* Anal Chem, 1993. **65**(7): p. 877-84.
27. Saudan, P., et al., *Ductus ejaculatorius peptide 99B (DUP99B), a novel Drosophila melanogaster sex-peptide pheromone.* Eur J Biochem, 2002. **269**(3): p. 989-97.
28. Pan, X., J. Luo, and S. Li, *Bacteria-Catalyzed Arginine Glycosylation in Pathogens and Host.* Front Cell Infect Microbiol, 2020. **10**: p. 185.
29. Date-Ito, A., et al., *Rapid evolution of the male-specific antibacterial protein andropin gene in Drosophila.* J Mol Evol, 2002. **54**(5): p. 665-70.
30. Schaeffer, S.W., et al., *Polytene chromosomal maps of 11 Drosophila species: the order of genomic scaffolds inferred from genetic and physical maps.* Genetics, 2008. **179**(3): p. 1601-55.
31. Walter, R., W.H. Simmons, and T. Yoshimoto, *Proline specific endo- and exopeptidases.* Mol Cell Biochem, 1980. **30**(2): p. 111-27.

32. Dougherty, P.G., A. Sahni, and D. Pei, *Understanding Cell Penetration of Cyclic Peptides.* Chem Rev, 2019. **119**(17): p. 10241-10287.
33. Begun, D.J. and H.A. Lindfors, *Rapid evolution of genomic Acp complement in the melanogaster subgroup of Drosophila.* Mol Biol Evol, 2005. **22**(10): p. 2010-21.
34. Samakovlis, C., et al., *The andropin gene and its product, a male-specific antibacterial peptide in Drosophila melanogaster.* EMBO J, 1991. **10**(1): p. 163-9.
35. Lung, O., L. Kuo, and M.F. Wolfner, *Drosophila males transfer antibacterial proteins from their accessory gland and ejaculatory duct to their mates.* J Insect Physiol, 2001. **47**(6): p. 617-622.
36. Maes, E., et al., *Major O-glycans from the nest of Vespula germanica contain phospho-ethanolamine.* Carbohydr Res, 2005. **340**(11): p. 1852-8.
37. Bordeaux, J.M., et al., *Noctilisin, a Venom Glycopeptide of Sirex noctilio (Hymenoptera: Siricidae), Causes Needle Wilt and Defense Gene Responses in Pines.* J Econ Entomol, 2014. **107**(5): p. 1931-45.
38. Hykollari, A., et al., *Tissue-specific glycosylation in the honeybee: Analysis of the N-glycomes of Apis mellifera larvae and venom.* Biochim Biophys Acta Gen Subj, 2019. **1863**(11): p. 129409.
39. Helling, F., et al., *Glycosphingolipids in insects. The amphoteric moiety, N-acetylglucosamine-linked phosphoethanolamine, distinguishes a group of ceramide oligosaccharides from the pupae of Calliphora vicina (Insecta: Diptera).* Eur J Biochem, 1991. **200**(2): p. 409-21.
40. Weske, B., et al., *Glycosphingolipids in insects. Chemical structures of two variants of a glucuronic-acid-containing ceramide hexasaccharide from a pupae of Calliphora vicina (Insecta: Diptera), distinguished by a N-acetylglucosamine-bound phosphoethanolamine sidechain.* Eur J Biochem, 1990. **191**(2): p. 379-88.
41. Zhu, F., D. Li, and K. Chen, *Structures and functions of invertebrate glycosylation.* Open Biol, 2019. **9**(1): p. 180232.
42. Zhang, L. and K.G. Ten Hagen, *O-Linked glycosylation in Drosophila melanogaster.* Curr Opin Struct Biol, 2019. **56**: p. 139-145.
43. Leader, D.P., et al., *FlyAtlas 2: a new version of the Drosophila melanogaster expression atlas with RNA-Seq, miRNA-Seq and sex-specific data.* Nucleic Acids Res, 2018. **46**(D1): p. D809-D815.
44. Muller, R., et al., *Characterization of mucin-type core-1 beta1-3 galactosyltransferase homologous enzymes in Drosophila melanogaster.* FEBS J, 2005. **272**(17): p. 4295-305.
45. Laflamme, B.A. and M.F. Wolfner, *Identification and function of proteolysis regulators in seminal fluid.* Mol Reprod Dev, 2013. **80**(2): p. 80-101.
46. Isaac, R.E. and N. Audsley, *Insect Peptide Hormones*, in *Amino Acids, Peptides and Proteins in Organic Chemistry*, A.B. Hughes, Editor. 2010, Wiley-VCH Verlag GmbH & Co. p. 575-595.
47. Rexhepaj, A., et al., *The sex-peptide DUP99B is expressed in the male ejaculatory duct and in the cardia of both sexes.* Eur J Biochem, 2003. **270**(21): p. 4306-14.
48. Kim, Y.J., et al., *MIPs are ancestral ligands for the sex peptide receptor.* Proc Natl Acad Sci U S A, 2010. **107**(14): p. 6520-5.
49. Apostol, C.R., M. Hay, and R. Polt, *Glycopeptide drugs: A pharmacological dimension between "Small Molecules" and "Biologics".* Peptides, 2020. **131**: p. 170369.
50. Lefever, M., et al., *Structural Requirements for CNS Active Opioid Glycopeptides.* J Med Chem, 2015. **58**(15): p. 5728-41.

51. Begun, D.J., et al., *Recently evolved genes identified from Drosophila yakuba and D. erecta accessory gland expressed sequence tags.* Genetics, 2006. **172**(3): p. 1675-81.
52. Zhou, Q., et al., *On the origin of new genes in Drosophila.* Genome Res, 2008. **18**(9): p. 1446-55.
53. Russo, C.A., et al., *Phylogenetic analysis and a time tree for a large drosophilid data set (Diptera: Drosophilidae).* Zoological Journal of the Linnean Society, 2013. **169**(4): p. 765-775.
54. Gao, J.-j., et al., *Phylogenetic relationships between Sophophora and Lordiphosa, with proposition of a hypothesis on the vicariant divergences of tropical lineages between the Old and New Worlds in the family Drosophilidae.* Molecular phylogenetics and evolution, 2011. **60**(1): p. 98-107.
55. Tamura, K., S. Subramanian, and S. Kumar, *Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks.* Molecular biology and evolution, 2004. **21**(1): p. 36-44.
56. Bulet, P., et al., *A novel inducible antibacterial peptide of Drosophila carries an O-glycosylated substitution.* J Biol Chem, 1993. **268**(20): p. 14893-7.
57. van der Linde, K., et al., *A supermatrix-based molecular phylogeny of the family Drosophilidae.* Genet Res (Camb), 2010. **92**(1): p. 25-38.
58. Seetharam, A.S. and G.W. Stuart, *Whole genome phylogeny for 21 Drosophila species using predicted 2b-RAD fragments.* PeerJ, 2013. **1**: p. e226.
59. Suvorov, A., et al., *Widespread introgression across a phylogeny of 155 Drosophila genomes.* bioRxiv, 2020.
60. Kim, B.Y., et al., *Highly contiguous assemblies of 101 drosophilid genomes.* bioRxiv, 2020.
61 Nishihara, S. *Functional analysis of glycosylation using Drosophila melanogaster*. Glycoconjugate J, 2020. **37**: p.1–14.
62 Rabel, D., et al., *Primary Structure and in Vitro Antibacterial Properties of the Drosophila melanogaster Attacin C Pro-domain.* J Biol Chem, 2004. **279**(15): p. 14853-9.



Graphic abstract

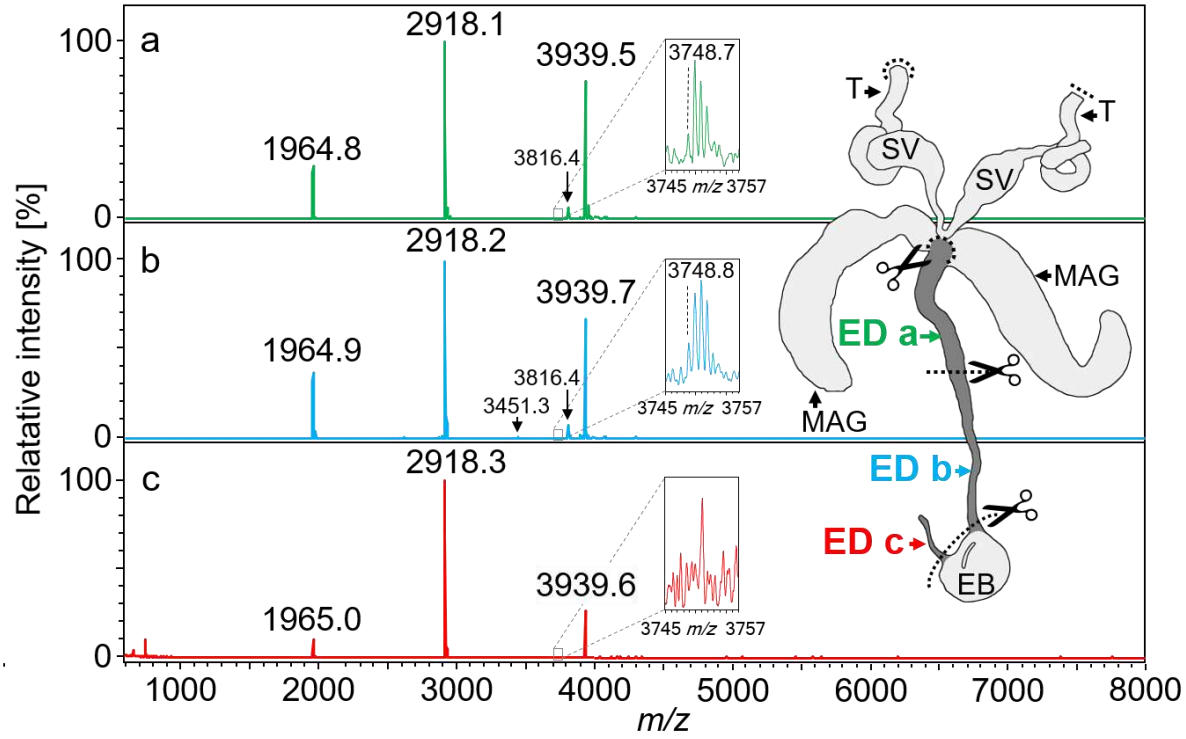


Fig. 1. **MALDI-TOF mass spectra of a single male ejaculatory duct (ED).** Spectra from (a) the dilated and (b) narrow regions of the anterior ejaculatory duct (ED), and (c) the posterior ED of a w^{1118} male. These regions are defined, relative to the other male reproductive tissues (T, testis; SV, seminal vesicle; MAG, male accessory gland) in the overlaid cartoon. Spectra were recorded in the mass ranges *m/z* 600-4000 and 3000-8000 and merged. All tissue sections gave prominent ion signals at *m/z* 1964.9, 2918.2 and 3939.6, corresponding to the molecular ions of Acp54A1 (CG34098), CG42782 and Met75Ca/b (CG18064 and CG32197), respectively. The minor ion signals at *m/z* 3451.3 and 3816.4 as

well as the abundant ion at *m/z* 3939.7 produced highly similar fragment patterns (see Fig. 2), hence considered to be peptide forms. A low abundant ion signal at *m/z* 3748.7 was noted in some spectra (see inset). This ion signal was more abundant in spectra of the ED extract and was assigned to andropin by MALDI-TOF MSMS.

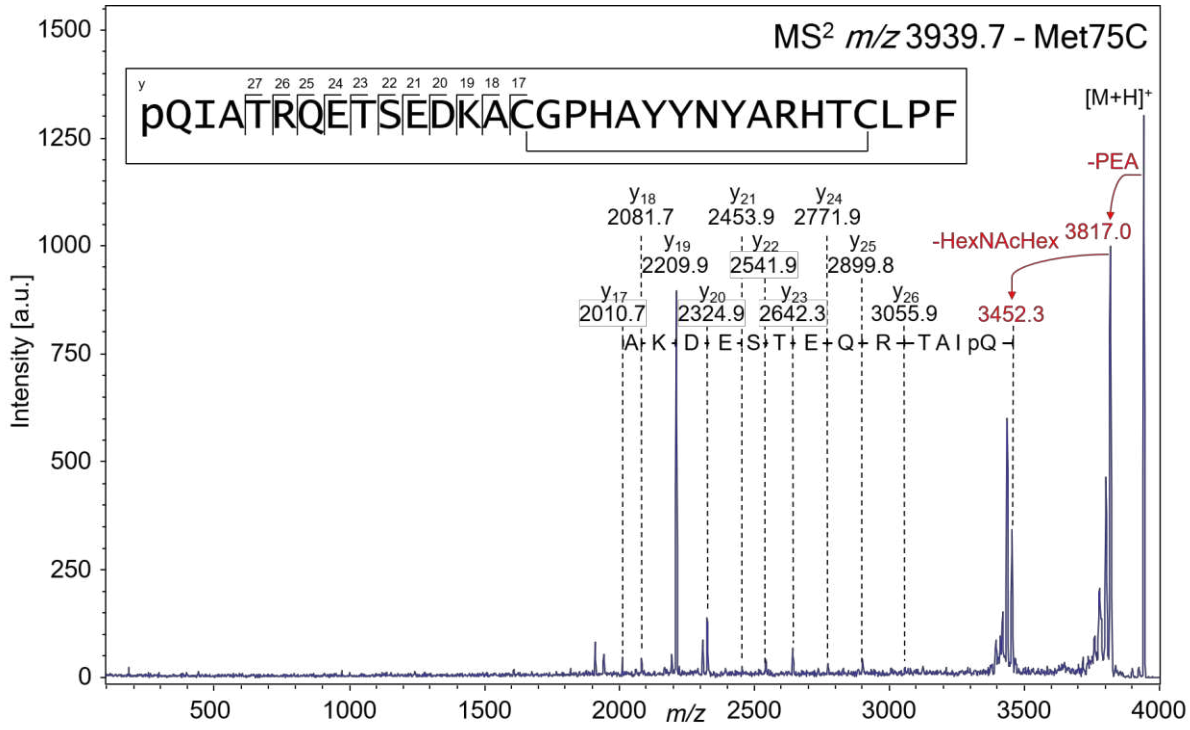


Fig. 2. **MALDI-TOF MS fragment spectrum of the Met75C precursor ion at *m/z* 3939.7 obtained from a single ED.** The spectrum of the non-reduced and non-derivatized peptide reveals the loss of phosphoethanolamine (PEA) and a disaccharide (HexNAcHex) as indicated. The fragment spectrum also comprises the *y*-ion series of the linear section of Met75C.

Table 1. Gene symbols, names, theoretical masses ($[M+H]^+$), deconvoluted experimental masses obtained from FT-ICR MS as well as the corresponding mass difference (Δ Mass) and sequences of ED peptides. Abbreviations: Carbamidomethylated cysteine, C*; pyroglutamic acid, pQ; N-acetylhexosamine, HexNAc; hexose, Hex; phosphoethanolamine, PEA; fucose, Fuc.

Gene symbol	Name	[M+H]$^+$	FT-ICR MS	Δ Mass	Sequence
CG34098	Acp54A1	1964.9768			MKCRLGFVKRGGQCTWP
		2081.0354	2081.0351	0.0003	MKC*RLGFVKRGGQC*TWP
CG42782		2918.3575			YEIIRQCPVHFVFKNNYCQYQPM
		3034.4161	3034.4153	0.0008	YEIIRQC*PVHFVFKNNYC*QYQPM
CG32197/CG18064	Met75Ca/b	3451.5583			pQIATRQETSEDKACGPHAYYNYARHTCLPF
		3567.6168	3567.6134	0.0034	pQIATRQETSEDKAC*GPHAYYNYARHTC*LPF
CG32197/CG18064	Met75Ca/b	3654.6376			pQIAT(HexNAc)RQETSEDKACGPHAYYNYARHTCLPF
		3770.6962	3770.6917	0.0045	pQIAT(HexNAc)RQETSEDKAC*GPHAYYNYARHTC*LPF
CG32197/CG18064	Met75Ca/b	3816.6905			pQIAT(HexNAcHex)RQETSEDKACGPHAYYNYARHTCLPF
		3932.7490	3932.7483	0.0007	pQIAT(HexNAcHex)RQETSEDKAC*GPHAYYNYARHTC*LPF
CG32197/CG18064	Met75Ca/b	3939.6990			pQIAT(HexNAcHexPEA)RQETSEDKACGPHAYYNYARHTCLPF
		4055.7576	4055.7571	0.0005	pQIAT(HexNAcHexPEA)RQETSEDKAC*GPHAYYNYARHTC*LPF
CG1361	Andropin	3749.0789	3749.0764	0.0025	VFIDILDKVENAIHNAAQVGIGFAKPFEKLINPK
CG33495	DUP99b	4930.2041			pQDRN(2HexNAc3Hex2Fuc)DTEWIQSQKDREKWCRLNLGPYLGGRC
		5046.2627	5046.2587	0.0040	pQDRN(2HexNAc3Hex2Fuc)DTEWIQSQKDREKWC*RLNLGPYLGGRC*

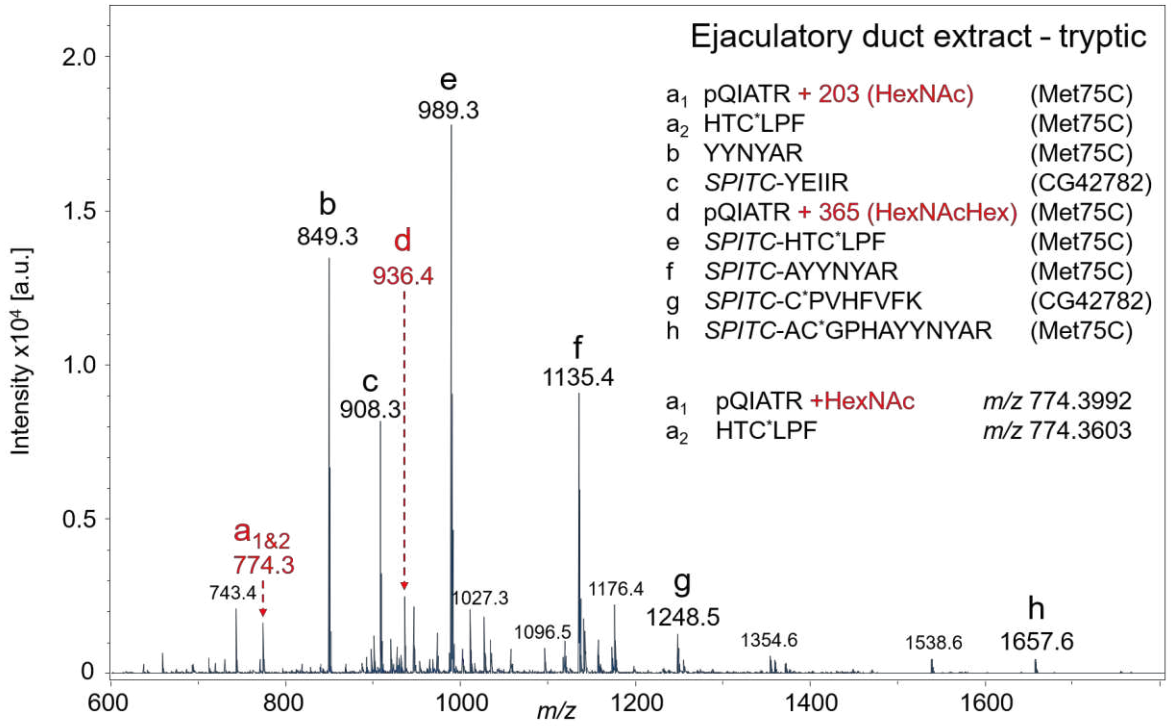


Fig. 3. **MALDI-TOF mass spectrum of the extracted ED peptides after reduction, alkylation, tryptic digestion and sulfonation .** The identity of the tryptic fragments (a-h) was confirmed by MSMS analysis.

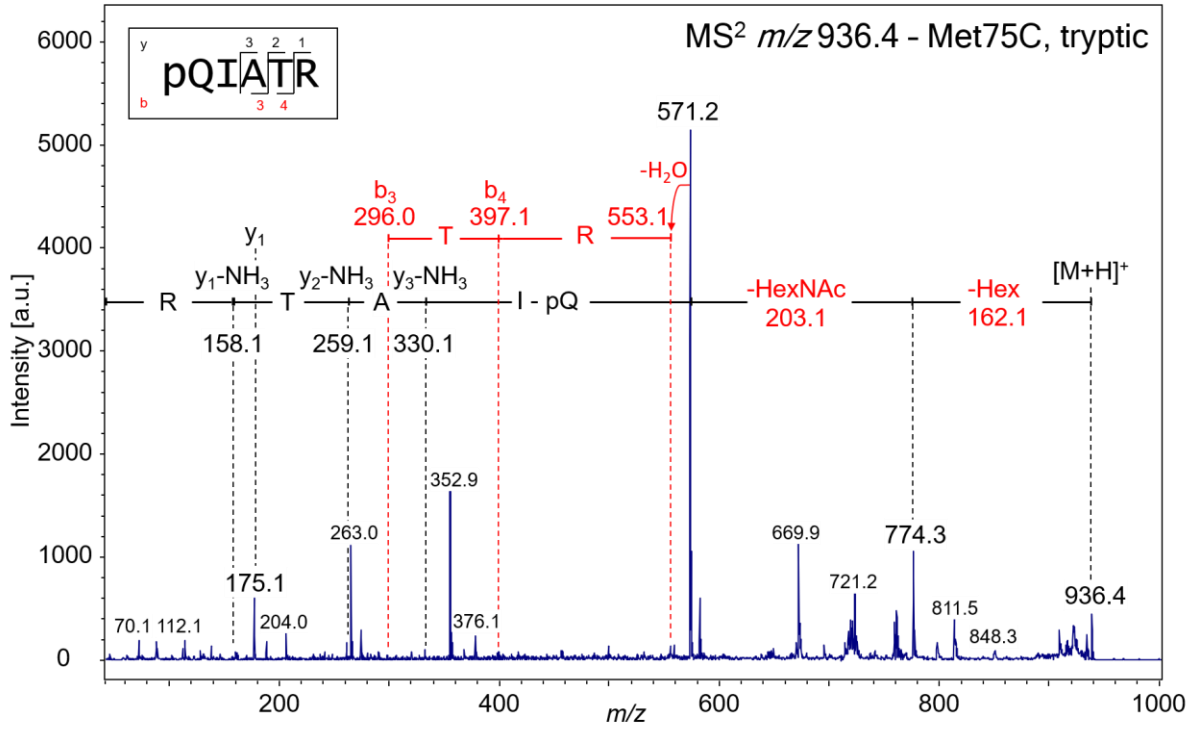


Fig. 4. **MALDI-TOF MS fragment spectrum of the Met75C tryptic peptide ion $[M+H]^+$, *m/z* 936.4.** The loss of a hexose and N-acetylhexosamine from the peptide and the *y*-ion series for the pentapeptide, pQIATR, are indicated.

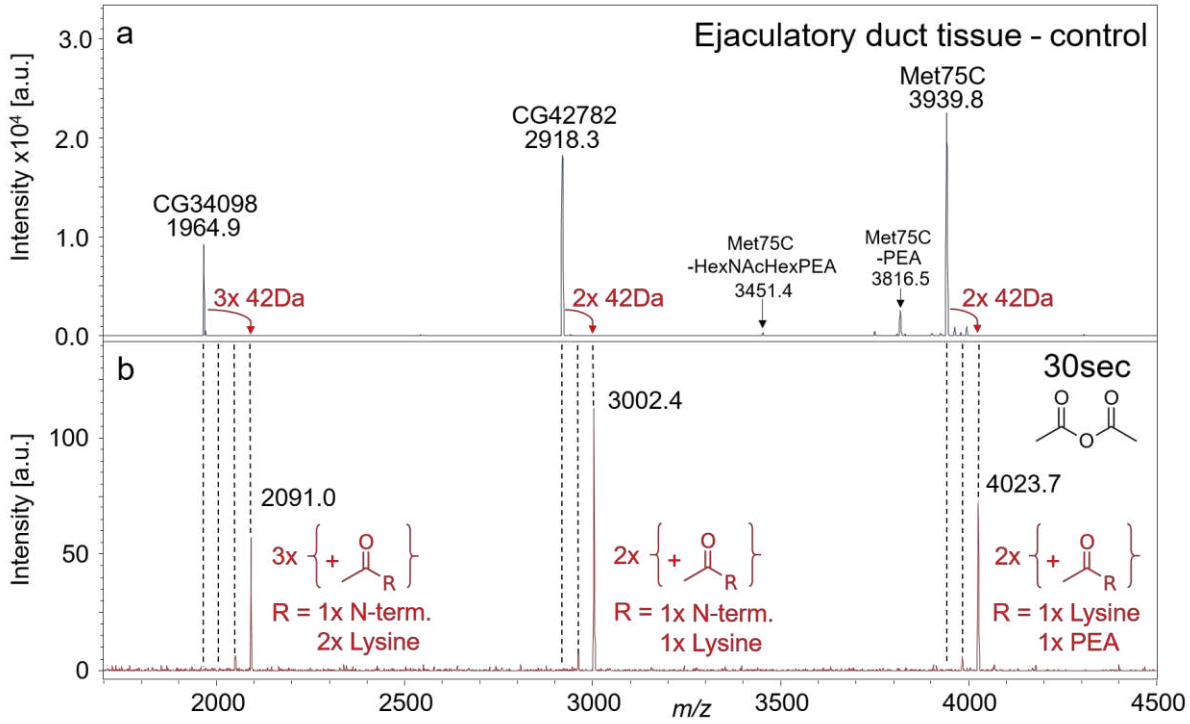


Fig. 5 **MALDI-TOF mass spectra by direct tissue analysis of a single male ejaculatory duct (ED) either untreated (a) or treated (b) for 30 seconds with acetic anhydride to acetylate primary amines.** The increase in mass of the molecular ions for Acp54A1, CG42782 and Met75C by the addition of two or three acetyl groups (42 Da) is labelled with red arrows. Dashed lines indicate the mass shifts between precursors, partially acetylated peptides and final products.

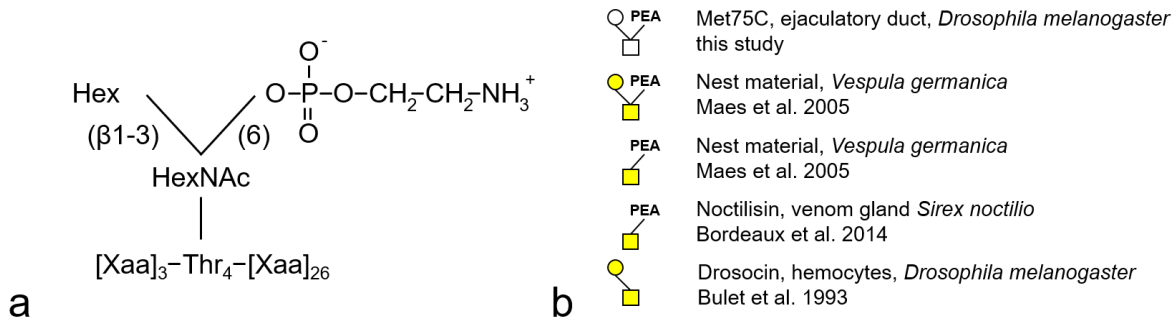
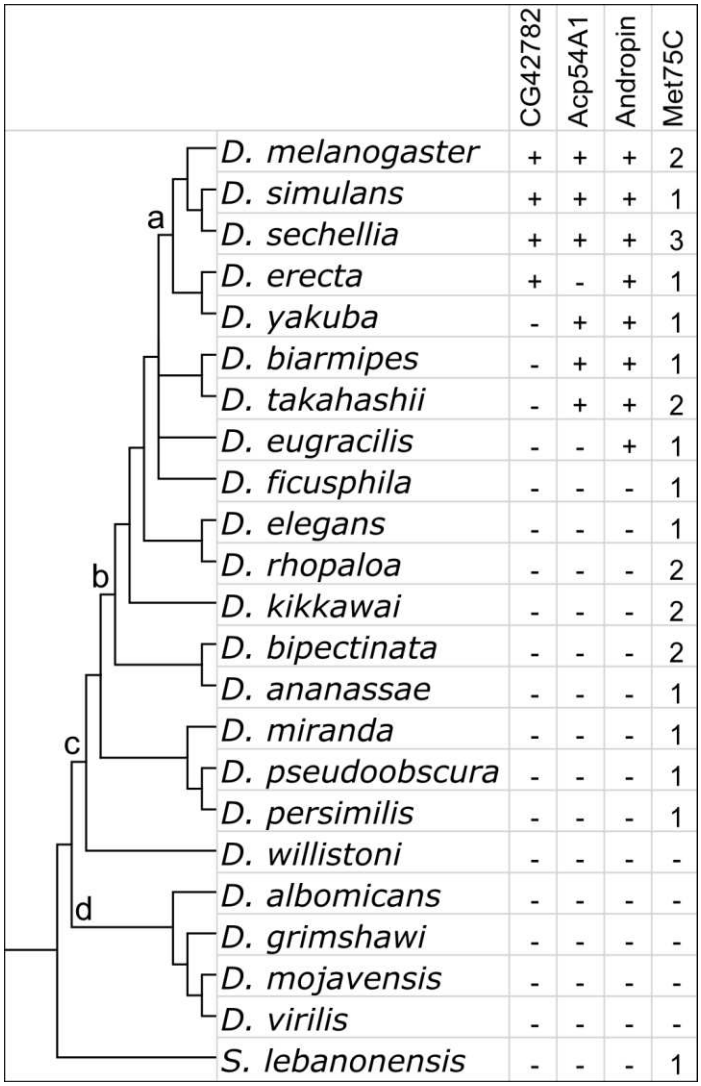


Fig. 6. **The structure of the glycan of Met75C and similar insect mucin-like O-glycans.** (a) The proposed structure of the zwitterionic Met75C glycan attached to Thr and (b) structures of selected insect O-glycans that have N-acetylgalactosamine (GalNAc, yellow

684 square) directly linked to Thr or Ser. GalNAc are extended with galactose (yellow circle) and
685 phosphoethanolamine (PEA). Maes *et al.* 2005[36]; Bulet *et al.* 1993[56]; Bordeaux *et al.*
686 1965[37].

687



688

689 Fig. 7. **Phylogram showing the distribution of orthologs in the genomes of traditional**
690 **representatives of various *Drosophila* groups**. Topology adapted from [57-60] with
691 disputed clades illustrated as polytomy. Names of clades referred to in the text are labelled at
692 the nodes (a) *melanogaster* subgroup (b) *melanogaster* group (c) subgenus *Sophophora* (d)
693 subgenus *Drosophila.* Corresponding accession numbers and sequences are given in the
694 supplementary material.

695

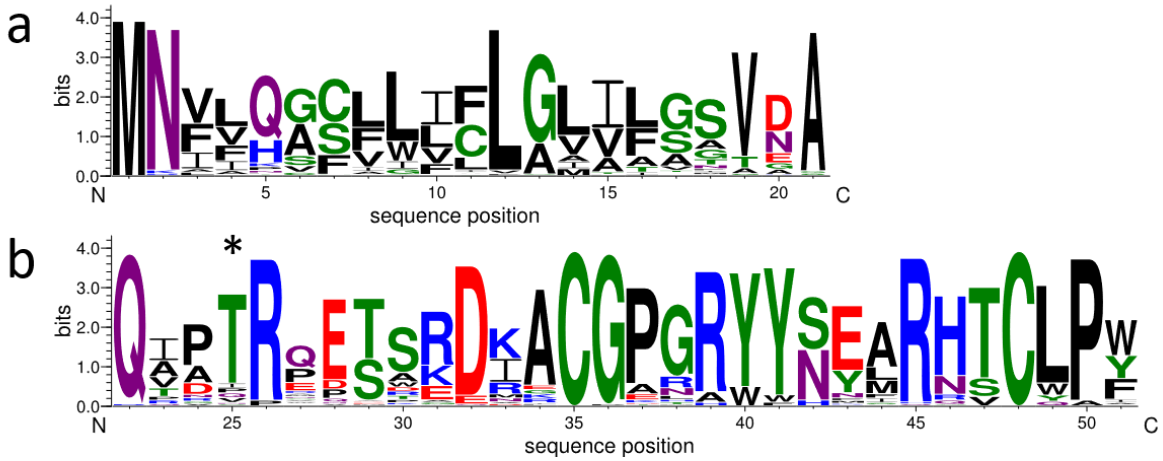
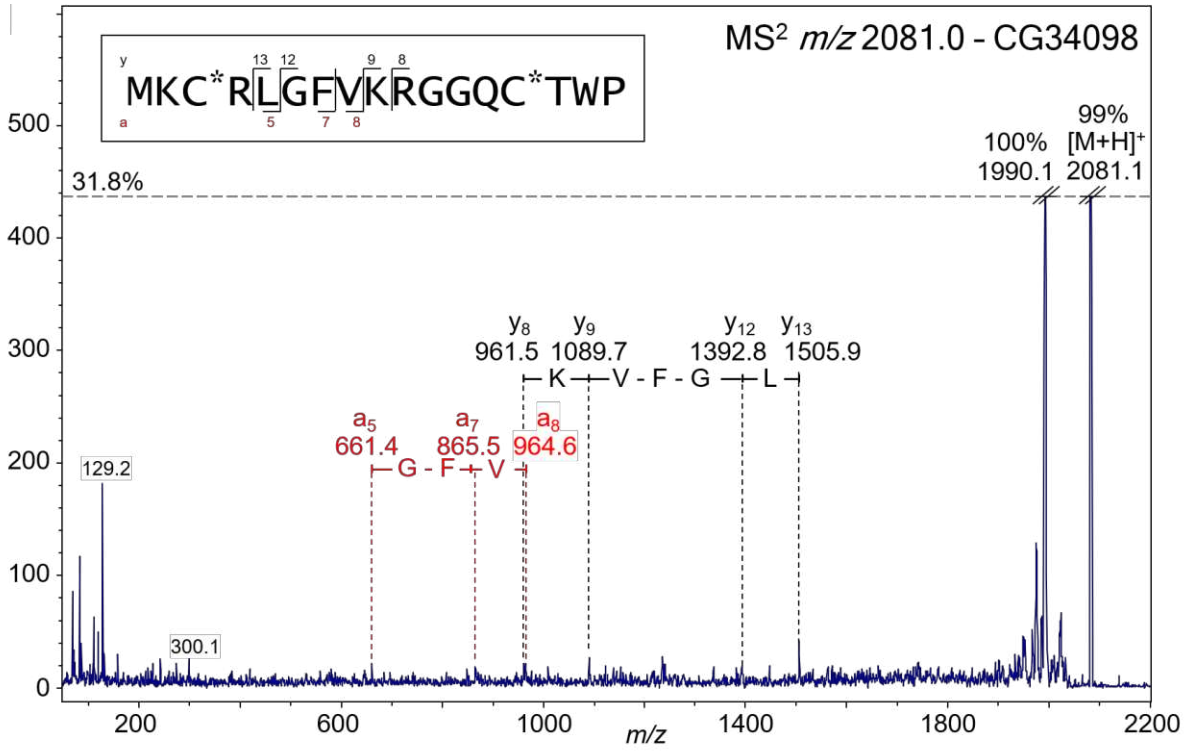
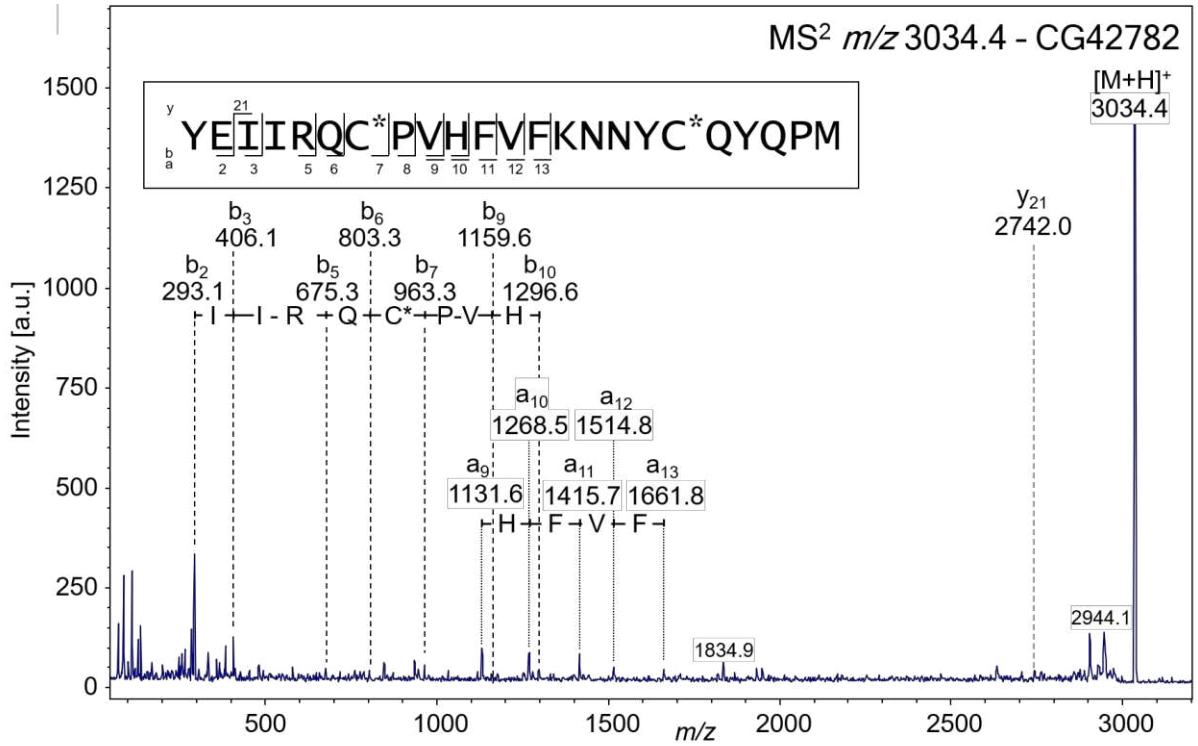


Fig. 8. **A graphical representation of the conservation of amino acids of 79 aligned sequences of Met75 ortho- and paralogs predicted from genome data of 51 *Drosophila* species and *Scaptodrosophila lebanonensis*.** The overall height of a stack indicates the sequence conservation at that position, while the relative height of letters indicates the frequency of occurrence of amino acids in one particular position. Consensus range of the predicted signal peptides (a) and the predicted mature peptides after signal peptide cleavage (b). Amino acid residues are colored according to their biochemical properties; polar (green), neutral (purple), basic (blue), acidic (red), hydrophobic (black). The proposed glycosylation site in *D. melanogaster* is indicated by an asterisk. The width of each stack is proportional to the fraction of valid symbols in that position. Rare sequence positions due to insertions or extension were excluded applying a 30%-cut-off. A complete sequence logo and the corresponding alignment are given in the supplementary material (Supplementary Fig.4; Met75C_alignment.fas).

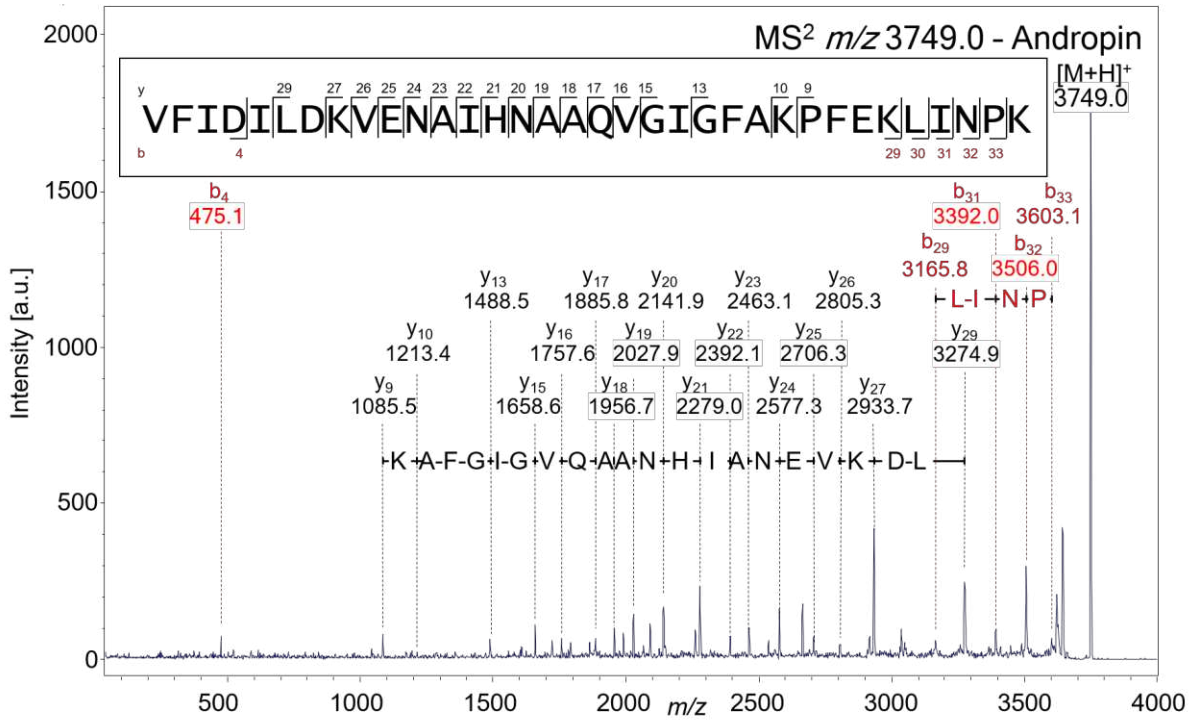
Supplementary material



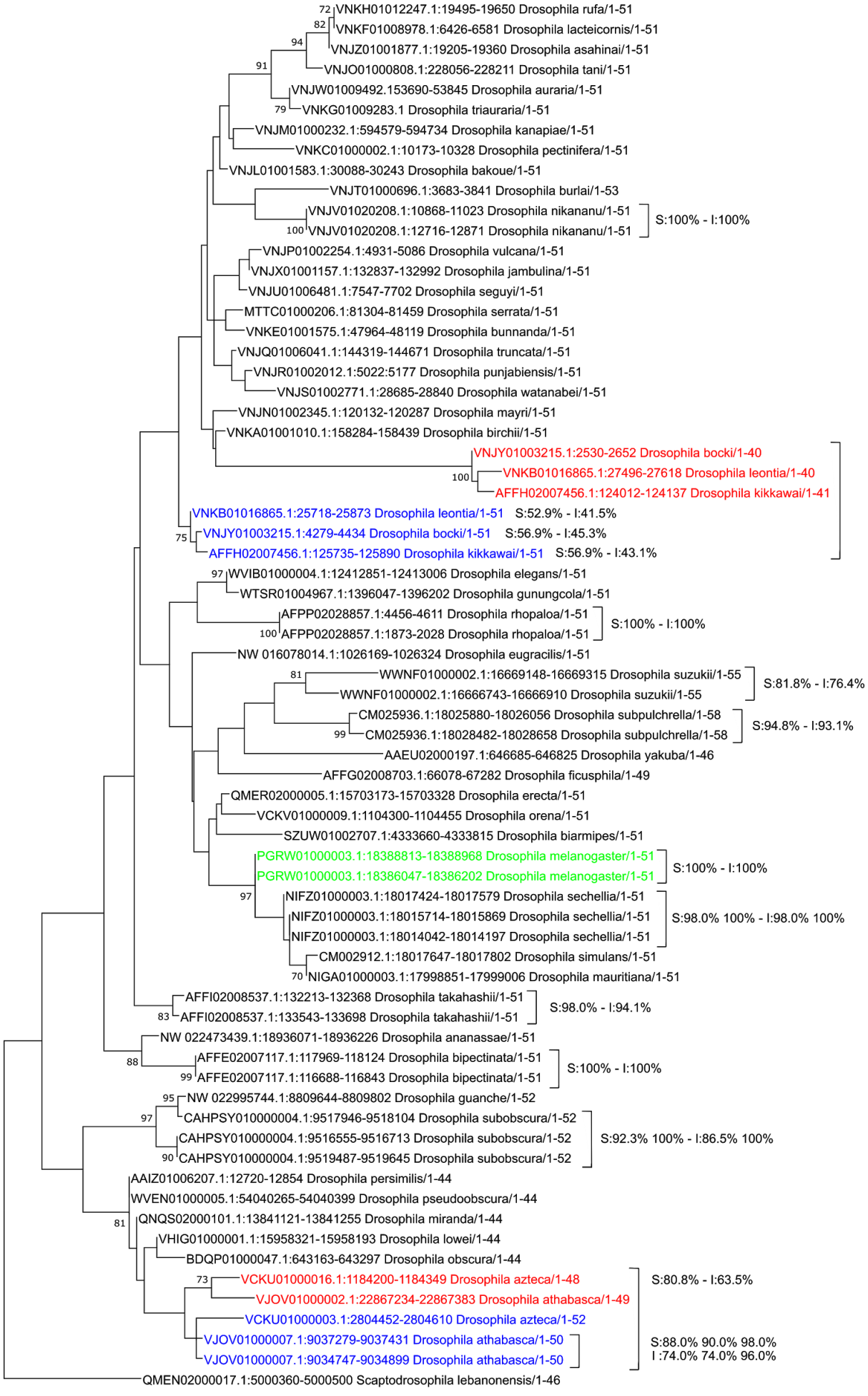
Supplementary Fig. 1. MALDI-TOF MSMS fragment spectrum of CG34098 from a reduced and alkylated ED extract. Product-ion spectrum resulting from fragmentation of the isolated molecular ion at *m/z* 2081.1, assigned to the CG34098 peptide. A selection of diagnostic fragments is labelled in the spectrum and the sequence coverage by all observed *a*- and *y*-series fragment ions is given in the schematic representation in the inset.



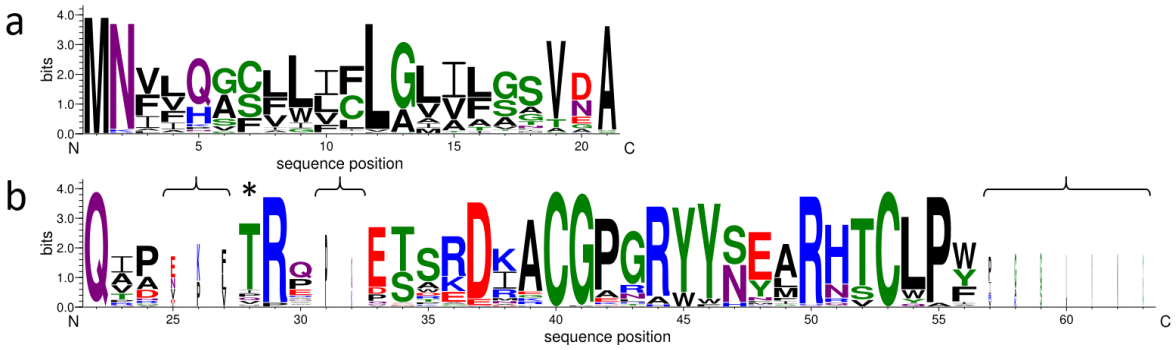
Supplementary Fig. 2. MALDI-TOF MSMS fragment spectrum of CG42782 from a reduced and alkylated ED extract. Product-ion spectrum resulting from fragmentation of the isolated molecular ion at *m/z* 3034.4, assigned to the CG42782 peptide. A selection of diagnostic fragments is labelled in the spectrum and the sequence coverage by all observed *a*-, *b*- and *y*-series fragment ions is given in the schematic representation in the inset.



Supplementary Fig. 3. MALDI-TOF MSMS fragment spectrum of andropin from a reduced and alkylated ED extract. Product-ion spectrum resulting from fragmentation of the isolated molecular ion at *m/z* 3749.0 assigned to the andropin peptide. A selection of diagnostic fragments is labelled in the spectrum and the sequence coverage by all observed *b*- and *y*-type fragment ions is given in the schematic representation in the inset.


72 VNKH01012247.1:19495-19650 Drosophila rufa/1-51
82 VNKF01008978.1:6426-6581 Drosophila lacteicornis/1-51
94 VNJZ01001877.1:19205-19360 Drosophila asahinai/1-51
91 VNJO01000808.1:228056-228211 Drosophila tani/1-51
VNJW01009492.153690-53845 Drosophila auraria/1-51
79 VNKG01009283.1 Drosophila triauraria/1-51
VNJM01000232.1:594579-594734 Drosophila kanapiae/1-51
VNKC01000002.1:10173-10328 Drosophila pectinifera/1-51
VNJL01001583.1:30088-30243 Drosophila bakoue/1-51
VNJT01000696.1:3683-3841 Drosophila burlai/1-53
VNJV01020208.1:10868-11023 Drosophila nikananu/1-51
100 VNJV01020208.1:12716-12871 Drosophila nikananu/1-51
S:100% - I:100%
VNJP01002254.1:4931-5086 Drosophila vulcana/1-51
VNJX01001157.1:132837-132992 Drosophila jambulina/1-51
VNJU01006481.1:7547-7702 Drosophila seguyi/1-51
MTTC01000206.1:81304-81459 Drosophila serrata/1-51
VNKE01001575.1:47964-48119 Drosophila bunnanda/1-51
VNJQ01006041.1:144319-144671 Drosophila truncata/1-51
VNJR01002012.1:5022:5177 Drosophila punjabiensis/1-51
VNJS01002771.1:28685-28840 Drosophila watanabei/1-51
VNJN01002345.1:120132-120287 Drosophila mayri/1-51
VNKA01001010.1:158284-158439 Drosophila birchii/1-51
VNJY01003215.1:2530-2652 Drosophila bocki/1-40
100 VNKB01016865.1:27496-27618 Drosophila leontia/1-40
AFFH02007456.1:124012-124137 Drosophila kikkawai/1-41
VNKB01016865.1:25718-25873 Drosophila leontia/1-51 S:52.9% - I:41.5%
75 VNJY01003215.1:4279-4434 Drosophila bocki/1-51 S:56.9% - I:45.3%
AFFH02007456.1:125735-125890 Drosophila kikkawai/1-51 S:56.9% - I:43.1%
97 WVIB01000004.1:12412851-12413006 Drosophila elegans/1-51
WTSR01004967.1:1396047-1396202 Drosophila gunungcola/1-51
AFPP02028857.1:4456-4611 Drosophila rhopaloa/1-51
100 AFPP02028857.1:1873-2028 Drosophila rhopaloa/1-51
S:100% - I:100%
NW 016078014.1:1026169-1026324 Drosophila eugracilis/1-51
81 WWNF01000002.1:16669148-16669315 Drosophila suzukii/1-55
WWNF01000002.1:16666743-16666910 Drosophila suzukii/1-55
S:81.8% - I:76.4%
CM025936.1:18025880-18026056 Drosophila subpulchrella/1-58
99 CM025936.1:18028482-18028658 Drosophila subpulchrella/1-58
S:94.8% - I:93.1%
AAEU02000197.1:646685-646825 Drosophila yakuba/1-46
AFFG02008703.1:66078-67282 Drosophila ficusphila/1-49
QMER02000005.1:15703173-15703328 Drosophila erecta/1-51
VCKV01000009.1:1104300-1104455 Drosophila orena/1-51
SZUW01002707.1:4333660-4333815 Drosophila biarmipes/1-51
PGRW01000003.1:18388813-18388968 Drosophila melanogaster/1-51
PGRW01000003.1:18386047-18386202 Drosophila melanogaster/1-51
S:100% - I:100%
97 NIFZ01000003.1:18017424-18017579 Drosophila sechellia/1-51
NIFZ01000003.1:18015714-18015869 Drosophila sechellia/1-51
NIFZ01000003.1:18014042-18014197 Drosophila sechellia/1-51
S:98.0% 100% - I:98.0% 100%
CM002912.1:18017647-18017802 Drosophila simulans/1-51
70 NIGA01000003.1:17998851-17999006 Drosophila mauritiana/1-51
AFFI02008537.1:132213-132368 Drosophila takahashii/1-51
83 AFFI02008537.1:133543-133698 Drosophila takahashii/1-51
S:98.0% - I:94.1%
NW 022473439.1:18936071-18936226 Drosophila ananassae/1-51
88 AFFE02007117.1:117969-118124 Drosophila bipectinata/1-51
99 AFFE02007117.1:116688-116843 Drosophila bipectinata/1-51
S:100% - I:100%
95 NW 022995744.1:8809644-8809802 Drosophila guanche/1-52
97 CAHPSY010000004.1:9517946-9518104 Drosophila subobscura/1-52
CAHPSY010000004.1:9516555-9516713 Drosophila subobscura/1-52
90 CAHPSY010000004.1:9519487-9519645 Drosophila subobscura/1-52
S:92.3% 100% - I:86.5% 100%
AAIZ01006207.1:12720-12854 Drosophila persimilis/1-44
WVEN01000005.1:54040265-54040399 Drosophila pseudoobscura/1-44
81 QNQS02000101.1:13841121-13841255 Drosophila miranda/1-44
VHIG01000001.1:15958321-15958193 Drosophila lowei/1-44
BDQP01000047.1:643163-643297 Drosophila obscura/1-44
73 VCKU01000016.1:1184200-1184349 Drosophila azteca/1-48
VJOV01000002.1:22867234-22867383 Drosophila athabasca/1-49
VCKU01000003.1:2804452-2804610 Drosophila azteca/1-52
VJOV01000007.1:9037279-9037431 Drosophila athabasca/1-50
VJOV01000007.1:9034747-9034899 Drosophila athabasca/1-50
S:80.8% - I:63.5%
S:88.0% 90.0% 98.0%
I :74.0% 74.0% 96.0%
QMEN02000017.1:5000360-5000500 Scaptodrosophila lebanonensis/1-46
0.20

Supplementary Fig. 4. Phylogenetic tree of Met75C orthologs and paralogs. The evolutionary history of 69 Met75C-related genes with 63 positions from 52 species was inferred using the Maximum Likelihood method and the Le Gascuel (2008) amino acid substitution model with five discrete gamma categories. The tree with the highest log likelihood (-2333.99) is shown. Sequences of a common species have been included if they were located in the same contig to exclude sequences which are no paralogs with the exception of *D. azteca* and *D. athabasca*, where paralogs are located at different chromosomes. Gene duplication events are highlighted with square brackets labeled with the pairwise percent similarity (S) / identity (I). Note the presence of multiple independent gene duplication events with the exception of a common duplication in the ancestral line of *D. leontia*, *D. bocki*, and *D. kikkawai*, as well as a common evolutionary event shared by *D. azteca* and *D. athabasca* (highlighted in red and blue). Nodes which occured in at least 70% out of 500 bootstrap replicates are labeled at the branches.



Supplementary Fig. 4. . The complete graphical representation of the conservation of amino acids of 79 aligned sequences of Met75C ortho- and paralogs predicted from genome data of 51 *Drosophila* species and *Scaptodrosophila lebanonensis*. The overall height of a stack indicates the sequence conservation at that position, while the relative height of letters indicates the frequency of occurrence of amino acids in one particular position. (a) the predicted signal peptide for the majority of sequences and (b) the predicted mature peptides after signal peptide cleavage. Amino acid residues are colored according to their biochemical properties; polar (green), neutral (purple), basic (blue), acidic (red), hydrophobic (black). The proposed glycosylation site in *D. melanogaster* is indicated by an asterisk. The

759 width of each stack is proportional to the fraction of valid symbols in that position. Rare
760 sequence positions due to insertions or extensions are marked with braces.

761

762

763

764