Correlated Gene Copy Number Changes in a Seminal Fluid Protein Network in *Drosophila*

J.A. Carlisle*[,1], B.K. McCormick*[,1], L. Verbakel[2], A.C. von Philipsborn[2], M.F. Wolfner[+,1], A.G. Clark[+,1]

*co-first authors
[+]corresponding authors: mfw5@cornell.edu, ac347@cornell.edu

Affiliations
1. Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853
2. Department of Neuroscience and Movement Science, Medicine Section, University of Fribourg, Switzerland

Keywords: Seminal Fluid Proteins, Copy Number Variation, Coevolution, Drosophila, gene network, gene duplication, gene loss

Abstract

Reproductive proteins often diverge rapidly between species, yet network function must be maintained. Shared selective pressures on network members and compensatory changes between members can drive their parallel evolutionary trajectories. Indeed, correlated evolutionary rates of amino acid sequence change have been observed for interacting reproductive proteins. But whether gene copy number changes also correlate has not been widely studied. Here, we investigated copy number variation (CNV) of genes in the *Drosophila* Sex Peptide Seminal Fluid Protein (Sfp) network. Previous research analyzed CNV of the Sfp Sex Peptide (SP) in *Drosophila* species. We focus on 9 other Sfps whose function is required to mediate the binding of SP to sperm in *D. melanogaster* which is required for persistence of female post-mating responses. To exhaustively annotate CNV of genes, we developed a computational pipeline pairing iterative protein queries to genome sequence searches with phylogenetic clustering to resolve homology relationships. We observed that the Sfp network's genes are ancestral to *Drosophila* and that there were repeated duplications and losses of network members across the genus. We detect statistically significant correlations in gene duplication or loss events among network proteins, and show this can be used to identify new members of the network. We also investigated CNV of female-derived proteins that act downstream of the SP sperm-binding network to modulate SP function, these proteins showed no significant correlation of gene turnover events with SP or its network. Our results provide insight into how evolving reproductive genes tolerate duplication and loss, and how network relationships could constrain reproductive protein evolution.

Significance Statement

Reproductive proteins often diverge rapidly between species, yet network function must be maintained. Shared selective pressures on network members and compensatory changes between members can drive their parallel evolutionary trajectories. We report correlated gene duplication and loss among members of the *Drosophila* Sex Peptide seminal fluid protein network, suggesting that duplication or loss events may drive corresponding events in other network genes. This work is a natural extension of the idea of evolutionary rate covariation, but instead of scoring rates of substitution it tracks correlated duplication and loss events on the phylogeny. Applied to the Sex Peptide network, the method reveals striking patterns, especially for coordinated loss, and identifies a new network gene that is experimentally confirmed.

Introduction

Reproductive proteins frequently show remarkable divergence between even closely related species, a phenomenon observed from mammals to invertebrates to plants and to sexually reproducing microbes[1–5]. Sexual selection and sexual conflict are frequently cited as explanations, although in some cases rapid change may instead reflect relaxed selective constraints [1,6–9]. Shared selective pressures on members of gene networks, paired with compensatory changes of network members to preserve function, can drive parallel evolutionary trajectories of network members [10–13]. Indeed, this coordinated evolution of functionally related genes has been documented for gene expression and sequence variation [10,14]. Yet, whether such coordination extends to gene copy number is largely unknown. We hypothesize that co-divergence in copy number among functionally related genes can provide insight into how genes operate together to mediate reproduction. Correlated patterns of gene copy number could even be used to identify new network members, in a guilt by association approach. Here, we test this hypothesis by investigating patterns of coordinated gene copy number variation (CNV) of members of a seminal fluid protein network that controls the long-term post-mating response in *Drosophila melanogaster*.

Seminal fluid proteins (Sfps) are key mediators of reproductive success, influencing sperm storage, fertilization, and female post-mating physiology [15,16]. *D. melanogaster* is a premier genetic model for studying

these functions, and the evolution of *Drosophila* Sfps is often cited as a prime example illustrating the evolution and rapid diversification of reproductive genes [8,17–20]. As phylogenetic distance increases, the number of recognizable *D. melanogaster* Sfps declines sharply, with only ~15% detectable in the accessory gland proteome *D. virilis* [21,22]. This loss of orthologs likely reflects a combination of rapid sequence divergence, gene turnover, and paralogy that obscures orthology. The recent availability of high-quality genome assemblies across the genus *Drosophila*, along with improved phylogenetic methods, provides an opportunity to improve the resolution of homologous relationships across evolutionary history and map the evolutionary trajectories of copy number changes of Sfps [23–28]. On average, Sfps show higher sequence divergence than most other non-reproductive proteins [8,18]. Beyond the evolutionary pressures that promote diversification of reproductive proteins, Sfps may also evolve more rapidly than intracellular proteins because, as secreted proteins, they experience fewer structural constraints and are exposed to more dynamic extracellular interactions [29–31]. Further, many Sfps also show expression restricted to the reproductive tract, reducing pleiotropic constraint on their evolution [31]. These factors may explain why, for some Sfps, copy number has also been observed to vary dramatically between species, making Sfps an exceptional system for probing the evolutionary forces driving CNV in reproductive genes and for testing how covariation in copy number may manifest within reproductive networks.

In *D. melanogaster*, the Sfp Sex Peptide (SP) regulates varied and persistent post-mating changes in female physiology and behavior, both reproductive and non-reproductive, through interaction with its female receptor SPR (Sex Peptide Receptor) [32–34]. These effects persist long-term and requires SP binding to sperm [35]. A network of Sfps is necessary for this sperm binding and thus for the female long-term post-mating response (LTR) [13,34,35]. The LTR includes the female exhibiting a sustained reduction in receptivity and increased egg laying [35]. The Sex Peptide Network (SPN) consists of Sfps that act after sperm enter the female and before they reach the sperm storage organs to enable stabilization of SP bound to sperm and localization of sperm to storage organs (see also Figure 1) [35]. Known SPN members belong to protein families (serine proteases and homologs, CRISPs/CAP-domain proteins, lectins) that are also found in the seminal fluid of species across the tree of life, including humans [15,36–43]. Successful storage of sperm with bound SP, followed by gradual release of SP's active region from sperm, is essential to sustain the female's long-term post-mating response [35]. Additional female-derived factors further modulate SP function and SP's binding to sperm; however, the identities of the relevant female reproductive tract molecules remain unknown [44].

SP originated within *Drosophillidae* in the ancestor of the subfamily *Drosophilinae*, which includes the genus *Drosophila* [12,32]. SP has no known homologs outside *Drosophilinae* subfamily, however its C-terminus, which is responsible for stimulating the LTR in *D. melanogaster,* is highly conserved[12,45,46]. SP shows extensive variation in copy number across the genus *Drosophila*, with some species having up to 7 copies[45]. In particular, higher copy number is observed within the clade where SPR acquired expression in the female reproductive tract[47]. Given that in *D. melanogaster* the SPN is critical to SP's long-term function in the female LTR, it may be expected that network Sfps may share similar selective pressures across the genus *Drosophila*, inducing signatures of coevolution between them[32,35]. Indeed, SP shows correlated evolutionary rates in its protein sequence with some, though not all, members of its network[13]. Such correlated patterns of sequence evolution have been used to identify several SPN members through Evolutionary Rate Covariation (ERC) with SP or with other previously identified network proteins[13]. Therefore, it seemed that the SPN would be an ideal testing ground to investigate correlations in copy number changes between functionally related reproductive proteins. We traced the origin of members of the SPN and their CNV across the genus Drosophila. We detected correlated duplication events or "gains" of SP with one member of its network. Interestingly, while loss events were strongly correlated between SPN proteins, no clear association was observed between SP loss and loss of network members.

Materials and Methods

Construction of phylogenies of *D. melanogaster* protein families

Paralogous genes within the species *D. melanogaster* were identified using Flybase annotations[48,49]. Translated protein sequences were acquired from FlyBase Batch Download using Fbgn identifiers[48]. Protein sequences were aligned with PROMALS3D[50]. The Interpro plug-in on Geneious (2024 versions) was used to annotate protein domains on sequences from the protein alignment[51]. The alignment was trimmed to the protein domain of interest as defined by Interpro, either a lectin domain (IPR001304), serine protease domain (IPR001254) or CAP domain (IPR014044)[52]. Trimmed protein alignments were used to construct a maximum likelihood phylogeny using RAxML-NG[53] with 1000 non-parametric bootstraps with resampling or until convergence. The best scoring topology of 20 starting trees (10 random and 10 parsimony-based) was used. Homologous protein relationships established through this analysis of *D. melanogaster* protein family members were validated by phylogenetic clustering of homologs collected from other species. Proteins were determined to be Sfps by following Wigby et al. 2020[15].

Annotation of orthologous and paralogous sequences in 135 Drosophilid species to assess copy number variation

Annotation of genes belonging to each gene family of interest and resolution of orthology and paralogy relationships between them was carried out via iterative blast and phylogenetic clustering (Supplementary Figure 1)[54,55]. The initial tblastn query for each gene family was constructed as follows: for single-exon families, we obtained the full-length amino acid sequences of all genes nested within the family of interest present in *D. melanogaster* and we used tblastn to collect the five best hits (sorted by e-value, with cutoff: 1e-5) for each of these genes in the 35 Drosophila species for which NCBI predicted transcriptomes were available (Supplemental File 1, Table 1). For multi-exon families, the initial query consisted of only the amino acid sequences of the longest exon containing a conserved functional domain for each focal *D. melanogaster* gene.

The resultant query files were then used to identify the coordinates of the five best hits for each query gene in 135 Drosophilid genomes (Supplemental File 1, Table 2) using tblastn (sorted by e-value, with cutoff: 1e-5). Using bedtools, we subsequently merged overlapping hits and obtained the corresponding genomic DNA sequences[56]. Open reading frames were predicted using the EMBOSS getorf tool (see Supplemental File 1, Table 3 for gene family-specific parameters for the above operations)[57]. The resultant ORFs were combined with the initial query file as well as a file containing the protein sequences of the ~10 closest paralogs to the gene family of interest. This combined fasta was then aligned using MAFFT and a protein phylogeny inferred using FastTree[58,59]. Strongly supported and well-separated clades of orthologous protein sequences were then manually annotated on the tree using FigTree and the corresponding sequences obtained using TREE2FASTA[60,61]. Only those sequences determined to be nested within the gene family of interest were retained.

The above steps, beginning with the tblastn search of the genomes, were repeated in a second and final round of BLAST and phylogenetic clustering[55]. In this round, the sequences obtained from the first round served as queries. Additional rounds would have been performed if new hits had been detected. Orthology and paralogy relationships were again inferred through manual annotation of well-supported clades on the resulting phylogenetic tree. To confirm the robustness of the phylogenetic clustering underlying our annotations, five additional combinations of aligners and tree-building programs were applied to the dataset, generally yielding highly concordant results (Supplemental File 4). For visualizing copy number, we separated gene families into orthogroups likely to have been single copy in the ancestor of the genus *Drosophila*. Finally, we confirmed orthogroup assignments by reciprocal BLAST against the *D. melanogaster* transcriptome, ensuring each sequence matched the expected gene(s). Associated code for the described pipeline for ortholog and paralog identification can be found in Supplementary File 2, a more detailed guide for this methodology and instructions on usage can be found in Supplementary File 3.

Inference of gene duplication and loss events via gene tree-species tree reconciliation

To determine whether network genes have undergone correlated changes in copy number across lineages, we sought to infer where on the species tree gene gain and loss events likely occurred. To this end, we applied a method called gene tree-species tree reconciliation using the package GeneRax[62]. The basis of this method is that the relationships among members of a gene family are necessarily linked to the relationships among the species that carry these genes. This is because changes to these genes in copy number and presence/absence represent events that occurred along the branches of the species tree. This means that the pattern of duplication and loss events underlying the complement of genes present in each taxon can be inferred along lineages by "reconciling" the differences between the gene tree and species tree (i.e. making it so that every split in the gene tree corresponds to either a speciation or gene duplication event on the species tree). We applied this method to each orthogroup, here defined as all descendants of genes inferred to be single copy in the common ancestor of *Drosophila*.

We first constructed maximum-likelihood protein phylogenies for each orthogroup (for SP and SPR only, the relevant protein sequences were obtained from Hopkins *et al.* (2024) rather than from our annotations). Protein alignments were constructed using PROMALS3D, an alignment program that integrates structural information into its predicted alignments and has been used to investigate deep protein family relationships[50,63]. Sites with less than 50% occupancy for aligned proteins were then masked with Geneious. The alignment was subsequently manually checked for likely pseudogenes and technical duplicates due to assembly errors, as described below, and offending sequences were removed from the alignment. The starting gene tree was then inferred from the resulting alignment using RAxML-ng, and the starting species tree, taken from Kim et al. (2024), was trimmed to remove species not present in both our analysis and Hopkins *et al.* (2024) (leaving 128 spp.)[24,45,53]. These files, alongside the inferred protein alignment, were used as the inputs for GeneRax and will be made available in a data repository upon publication. In all cases, only duplications and losses (i.e., no horizontal transfers) were allowed (see Supplemental File 1, Table 10 for all relevant parameters).

Ruling out spurious duplication/loss events due to assembly errors

To rule out assembly errors as a major driver of the patterns observed, we manually validated a subset of the inferred gain and loss events. For gains, to rule out technical duplication due to (for example) incorrect haplotype phasing, we manually checked all inferred gene duplications on terminal branches (i.e., those observed only in a single genome). We flagged and removed any cases where the duplicates are both 1) identical in DNA and amino acid sequence, and 2) not convincingly shown to represent unique sites in the genome (e.g. the two copies are in tandem on single scaffold/contig). Across all gene families, we removed 3 probable technical duplicates from the dataset and adjusted relevant analyses accordingly. For losses, validation was performed only for those events associated with the four major concerted losses of sex peptide network members that were identified prior to the calculation of correlations. To do this, for long-read assemblies, we confirmed losses syntenically by checking the presence/order of conserved genes flanking the gene that was lost (Supplemental File 1, Table 12). To identify locations of flanking genes, we queried *D. melanogaster* protein sequences against other species' genomes using tblastn. Alternatively, for species with only short-read assemblies, as is the case for many *montium* group species, we searched for evidence of the lost genes in the raw sequence reads to rule out assembly artifacts. We used the amino acid sequences of each missing gene as a query in the tblastn searches against the raw reads associated with each assembly. We did this for the species with losses as well as for several close outgroups that did not experience the loss events of interest and compared the distribution of e-values obtained for each query between these groups (Supplemental File 1, Table 9). The expectation is that, in the case of a genuine loss (either pseudogenization and degeneration or total loss), the e-values of the hits obtained should decay much faster than for closely related species where the gene is intact. The only exception to this would be very recent pseudogenes that have degenerated minimally. In these cases, we used the assembly to confirm the presence of a frameshift

mutation or premature stop codon in the coding sequence. We also blasted the best hit obtained back to the *D. melanogaster* transcriptome to screen for the presence of pseudogenes. No spurious losses were identified.

Correlation of gene duplication and loss among orthogroups

Statistics were computed in R v. 4.5.1[64]. To test for correlated changes in copy number among gene families, we performed Spearman's rank correlations on the raw counts of events (i.e., gains or losses) inferred to have occurred on each branch of the species tree. This approach is conservative in that it identifies only cases where two gene families experience events of the same kind on the same branch, meaning that some types of genuine associations may be overlooked. For example, we do not test for associations between events of different types (e.g., when losses of orthogroup A are coupled to gains of orthogroup B) or across time (e.g., when ancestral gains of orthogroup A drive subsequent gains of orthogroup B on descendent branches). Also note that, for identifying correlated loss events, only complete losses of all members of orthogroups (as defined above) were considered. There are two reasons for this: 1) duplication followed quickly by sporadic loss of a single daughter copy in descendent lineages likely carries different effects than complete loss of an orthogroup that predates the genus, and 2) because we don't allow for gene transfer in our models, introgression or horizontal gene transfer could be misinterpreted as an ancient duplications followed by multiple subsequent loss events, thereby adding considerable noise to the analyses.

Testing whether a candidate gene functions in the Sex Peptide Network

Fly stocks used in this study are listed in Table 1 (in Supplementary File 5**)**. Control and experimental flies were maintained on a standard medium containing sugar, oatmeal, cornmeal, agar, and yeast, under a 12h light:12h dark cycle at 25°C. Flies were collected after eclosion and aged for 4-7 days in single-sex groups. Males were individually aged in a flat bottom 1.5 mL 96 well block filled with 0.5 mL food per well and covered with a PCR foil with air holes. Unmated females were aged in groups of 10-15 in food vials. All females used were from the wild type Canton-S strain. The detailed genotypes of experimental male flies are listed in Table 2 (in Supplementary File 5), but are briefly listed here as the two independent knockdown lines (*tub-GAL4 > RNAi BDSC_65910* and *tub-GAL4 > RNAi VDRC_14158*, with associated controls *RNAi BDSC_65910* control and *RNAi VDRC_14158* control, respectively; *tub-GAL4* flies were used as an additional control).

For remating assays, individual couples of 1 experimental or control male and 1 wild type Canton-S virgin female were observed in mating chambers (1 cm diameter, 4 mm height). Females that copulated within 1hr were recollected after the end of copulation and kept (without males) for 4 days. The mated females were then individually tested for remating with wild type Canton-S males within 1hr in a mating chamber.

Results and Discussion

Between species, changes in gene copy number of Sfps caused by duplication or loss events could arise through neutral or adaptive processes. Copy number of Sfp genes may increase due to strictly neutral processes, since Sfps, as secreted proteins, are likely to be dosage-tolerant, and paralogs may show functional redundancy which could buffer against gene loss[4,65–67]. Alternatively, copy number increases may be driven by natural selection to rapidly adjust dosage in response to selective pressures, and duplication can provide the evolutionary flexibility to allow a paralog to explore new functions without compromising the original gene copy[36,68–73]. Similarly, gene loss of Sfps may also be caused by a neutral process such as loss of selective constraint through redundancy or decreases in functional importance[36,69–71]. Gene loss may also occur as an adaptive process, such as when maintaining the gene imposes fitness costs (e.g. for reproductive proteins, through selection, or loss of benefit under changed mating systems or ecology)[74,75]. However, correlations in gene duplication or loss events between functionally related genes could point to whether directional selection informs CNV[76–78]. Here, we investigate the origin and CNV of a network of Sfps (and some functionally related female proteins) that work with the Sfp Sex Peptide (SP) to modulate the long-term post-

mating response in *Drosophila* females. To exhaustively identify all homologs and accurately distinguish between paralogy and orthology, we developed a pipeline pairing iterative queries of *Drosophila* genomes with phylogenetic clustering of hits to resolve accurate orthologous vs paralogous relationships. Statistical tests revealed covariation in gains and losses among certain functionally related Sfp proteins, suggesting that functional interactions between these proteins may underlie covariation, and that the observed CNV of Sfps is unlikely to be explained solely by neutral evolutionary processes.

The origin of Sex Peptide Network proteins largely predates Sex Peptide's origin

Of the eight known members of the SPN Sfps, four are predicted active or inactive serine proteases, two are lectins, and two are CAP domain proteins[35]. For protein families that contribute to the SPN, we constructed protein phylogenies of all *D. melanogaster* members. We determined that all SPN proteins belonging to the same protein family were closely paralogous to each other, and frequently have additional closely paralogous Sfps that have not yet been tested for network function[35] (Figure 1B). The close paralogy between many network members suggests that duplication followed by specialization of function may have contributed to the origin of the network[70]. We tested this by an iterative BLAST approach across 135 genomes, including species within the genus *Drosophila* as well as outgroup species beyond the genus[55]. This approach not only allowed for a detailed profiling of protein evolution over ~40 million years, but also increased the likelihood of identifying orthologous proteins that may exhibit elevated sequence divergence[24]. By iterating through progressively more distantly related species, the method enhanced sensitivity and helped avoid missing genes that could be obscured by piecemeal sampling within the genus. Notably, *seminase*'s frequent and complex pattern of duplication and loss between species led to many distinct lineages of *seminase* homologs across the genus. To trace changes in CNV across the genus, we combined all *seminase* paralogs into a single orthogroup, composed of the set of genes across species that are descended from a single ancestral gene (here, ancestral to the genus *Drosophila*).

We traced the origins of SPN genes (Figure 2, Supplementary File 1, Table 14). Our findings show that all members of the SPN are present in the common ancestor of the genus *Drosophila*. Additionally, all the network members are identifiable in the outgroup Drosophilinae fly species *Chymomyza costata* and *Scaptodrosophila lebanonensis*, suggesting that orthologs of all known network members existed in the ancestor of the Drosophilinae within the Drosophilidae family of flies. SP itself also originated in the ancestor of Drosophilinae[12,45]. However, many network members are detectable in the Drosophilidae subfamily Steganinae species *Leucophenga varia* (including *Sems*, *CG9997*, *Aqrs*, *CG17575*, *Antr*, and *CG1652*). These members represent all the paralogous groups of network proteins described in Figure 1A and include all network members required for transfer of other network proteins. This suggests that, except for duplication of *CG9997* and *CG1656* (to give rise to *intr* and *CG1652* respectively), the network originated in the common ancestor of *Drosophilidae*, predating the origin of SP. Notably, no orthologs of these proteins—or their close paralogs, which are also seminal fluid proteins in *D. melanogaster*—can be detected outside of *Drosophilidae*. Taken together, these findings may suggest the network existed before the emergence of SP with its own independent function, and that SP only later co-opted the network for its own function.

Extensive copy number variation of Seminal Fluid Proteins in the Sex Peptide Network

The presence of all SPN orthogroups in the *Drosophila* genus ancestor enables a 40-million-year comparative analysis of gene duplication and loss. By comprehensively annotating orthologs of SPN proteins and their close Sfp paralogs, we uncovered substantial CNV among many Sfps, arising from both gene gain and loss events (Figure 3).

Against our expectations, aside from *seminase* and CG43129, no SPN genes and paralogous Sfp genes investigated exhibit the exaggerated CNV observed for SP, indicating that functional constraints limited the expansion of these SPN members. Still, small numbers of duplications (~1–8) occur sporadically for nearly every gene. Loss events are even more striking. In 9 species spanning the genus the entire SPN is not present,

with the exception of *seminase* which is never lost. This recurrent, wholesale loss is unexpected given the ancestral origins of SPN genes and their crucial role in long-term post-mating responses in *D. melanogaster*, although their function in other species is not known. The pattern suggests functional interdependence among these network members such that the loss of one renders others dispensable. Notably, *seminase* and usually *SP* persist in these lineages, implying that their retention may reflect functions independent of the rest of the network (e.g. *seminase*'s interaction with *Semp1*[79,80]).

The SPN gene *seminase* as well as *CG34129*, a Sfp paralog of the serine protease SPN genes show extensive CNV across the genus driven by duplication events. SP shows 34 independent duplication events, and *seminase* 28, generating complex patterns of lineage-specific homology. In *D. melanogaster*, *seminase* has two close tandem paralogs (*CG11037* and *CG10587*), neither of which has yet to be functionally tested for SPN functions. Duplications of *SP* and *seminase*, but not *CG34129*, are largely restricted to clades where *SPR* is expressed in the female reproductive tract (FRT)[45–47].

CNV patterns provide insights into functional relationships and selective pressures. Duplications may be retained neutrally, by selection for increased dosage, or through neofunctionalization. Comparative –omics analyses can help distinguish these scenarios: divergence in sequence or expression suggests gains in copy number are driven by selection for novel function or correlated dosage-dependency. However, most duplication events of investigated Sfp genes are recent and specific to a few species, potentially implying that duplication events are not rare but short-lived, indicative of neutral evolution or frequently fluctuating selective pressures between species. Identifying correlations of copy number changes with other functionally related genes would reject neutrality in favor of selection. However, only *SP*, *seminase*, and *CG34129* display sufficient CNV for such tests, as most SPN Sfps appear resistant to duplication. Meanwhile, loss events can be seen as an inevitable outcome when a gene's function is no longer needed or perhaps when the gene has become deleterious. The repeated, coordinated losses of SPN members highlight a paradox: genes deeply conserved and essential in some species can become dispensable in others. This turnover underscores the fluidity of reproductive networks and suggests that evolutionary pressures on mating systems can rapidly restructure even highly integrated molecular networks. The repeated occurrence of lineages where all SPN members except *seminase* are lost could suggest interdependency in function between most of the SPN genes such that loss of one leads to the loss of functionality of the others. Interestingly, in most of these lineages where most SPN proteins are lost, *SP* is still present, indicating that these shared losses may be independent of *SP* function.

Correlation of gene duplication events of *Sex Peptide* and *Seminase*

Frequent duplication of *SP*, the SPN gene *seminase*, and the Sfp *CG34129* across the genus *Drosophila* provided an opportunity to test whether duplication events among these genes are correlated. Rare duplications cause meaningful correlation analysis to be infeasible since technical noise could overshadow genuine signals. When a gene is duplicated and fixed in a species, that duplication will generally be inherited by descendant species, which may create correlations in copy number across lineages simply due to shared evolutionary history. To avoid the effects of phylogenetic nonindependence confounding our results, we tested for co-occurrence of discrete duplication events along branches of the species tree rather than raw copy number covariance (Supplementary File 1, Table 5 shows number of duplications along branches). Such correlations may indicate shared selective pressures driving increased copy number (e.g. pressures favoring increased dosage or evolutionary flexibility), or possibly some functional relationship between protein products where duplication of one gene promotes duplication of another[78,81–85].

Our analyses revealed a striking pattern: duplication events of *SP* and *seminase* are highly correlated across *Drosophila* lineages (Spearman's correlation, $p = 0.003$, $\rho = 0.19$), whereas no such correlation was observed between *SP* and *CG34129* ($p = 0.3$, $\rho = .064$), or between *seminase* and *CG34129* ($p = 0.2$, $\rho = .089$). Specificity in correlation between *SP* and the SPN member *seminase* but not the paralog *CG34129* suggests that frequent duplication of *SP* and *seminase* is evolutionary related and not likely driven by neutral evolution. Further, both *SP* and *seminase* show an elevated rate of duplication in lineages where *SPR* gained expression

in the FRT (*SP*: 4.0 vs. 0.22, 28 duplications over 7 units of branch length in the *SPR* in FRT clade vs 3 duplications in 13.5 units of branch length; *seminase*: 2.7 vs. 0.30, 19 duplications over 7 units of branch length in the *SPR* in FRT clade vs 4 duplications in 13.5 units of branch length duplications per unit branch length). We tested whether duplication events of *SP* and *seminase* were more common along branches in the clade where *SPR* has acquired expression in the FRT. We used a chi-squared test to compare the observed occurrences to expected values based on the total number of events observed and the total branch length for branches in the clade with SPR expression in the FRT or branches outside this clade (Supplemental File 1, Table 11). We observed that *SP* and *seminase* were more often duplicating in the clade where SPR gained FRT expression (*SP: p* <0.0001, *seminase: p* <0.001). No other genes with duplications shared this pattern. This indicates that after the origin of the *SPR* FRT expression, strong selective pressure acted on both *SP* and *seminase* to expand in copy number. In contrast, while Seminase acts in a cascade to cleave Semp1 (which then processes Acp36DE and ovulin), CNV of *Semp1* is not correlated with *seminase* or *SP* Supplemental Figure 2. This implies that *seminase*'s extensive duplication may be tied to its interaction with *SP* or to something else that also drives *SP*'s duplication rather than to its broader protease cascade function.

For Sfps with identifiable homologs across species, copy number can vary dramatically. The repeated co-duplication of *SP* and *seminase* cannot be explained by neutral processes alone, instead suggesting selective pressures or interdependence drive their CNV across species. Whether this correlation reflects unidirectional dependence (e.g., *SP* expansions occur following *seminase* duplication) or shared selective pressures acting on both genes remain unresolved, but either possibility points to a genuine functional connection. The increased rate of duplication of these orthogroups following SPR's acquisition of FRT expression further supports the idea that this transition amplified selective pressures on *SP* and *seminase*. Yet this raises a key question: why do *SP* and *seminase* display co-duplication, but not with the other identified SPN genes? One possibility is that *seminase* is unusual within the network: although required for *SP* and other Sfps to bind sperm, *seminase* itself is not detectably associated with sperm as has been seen for most of the other Sfps. This difference may make *seminase* more tolerant to duplication, and therefore more likely to co-expand with *SP*, whereas sperm-associated proteins may have to remain more constrained. In summary, the correlated duplication of *SP* and *seminase* suggests a potential selective and functional dependency, highlighting the value of phylogeny-based duplication analyses in uncovering unexpected gene interactions, and in generating specific, testable hypotheses about the molecular mechanisms driving the correlations.

Correlation of gene loss events between members of the Sex Peptide Network, but not with Sex Peptide

We also investigated correlations in gene loss events between SPN proteins across the genus *Drosophila* as an alternative signal of variation in natural selection acting across species. For our analysis we only considered complete losses of all copies within an orthogroup ancestral to the genus *Drosophila*, rather than individual loss events of a subset of paralogs (Supplemental File 1, Table 6 (orthogroup absence) & Table 7 (Spearman's rho).

We observed strongly correlated losses between all members of the network except *SP* and *seminase* (*SP* is lost 3 times, but the losses are uncorrelated). Indeed, signatures of correlated loss between SPN members appear to be stronger than the reported ERC between the same network member pairings. These effects are largely driven by branches upon which these SPN Sfps are lost at once. In three cases, all SPN genes (except *seminase*) disappear simultaneously along single branches (out of 258 total branches on the tree and less than 10 total losses per gene); in a fourth instance, *intr* is lost first, followed by the remainder of the network on a descending branch (lagging events would not increase statistical significance with our approach). We investigated the causal event of the losses of SPN proteins in the four lineages that lost all network members and, in many cases, decayed ORFs of pseudogenized genes were identifiable, indicating independent gene loss events rather than loss of multiple tandem genes via deletion of genomic regions, enabling resolution of coordinated loss events even between genomically adjacent paralogs.

Although pairwise comparisons between most network members showed statistically significant correlations among loss events, the biological meaning or strength of these correlations remains unclear. However, it is intriguing that the strongest correlation in loss events for *CG1652* and *CG1656* is to the three network Sfps that are required for their successful transfer (*Aqrs*, *Antr*, *CG9997*), suggesting functional interdependence driving their strongly correlated losses. Potentially, this could indicate physical interactions in a complex where loss of one protein leads to the destabilization of the complex and the loss of the other members, as was seen for correlated losses of PRDM9 complex members in mammals and for complexes in other organisms[78,81,82]. The disconnect in correlated loss events of *seminase* and *SP* to the rest of the SPN may be related to these genes' function. Since the SPN genes largely arose before *SP*, potentially these genes have an ancestral shared function independent of *SP*.

Sex Peptide network gene loss correlates primarily with loss of other known Sex Peptide Network genes

The coordinated losses of SPN Sfps observed is suggestive of selective forces that act upon the network as a whole. However, this pattern is also consistent with pressures acting on Sfps more broadly to cause loss events along specific lineages, leading to strong signals of coordinated CNV independent of pressures on the network *per se*. If coordinated losses reflect lineage-specific selective pressure on Sfps and are not specific to SPN members, then we would expect Sfps without association with the SPN to not show correlation of loss events with SPN members. Indeed, not all Sfps paralogous to SPN members showed correlated patterns of loss, indicating that this signature is more likely to be specific to network members. *CG9997-dup* (absent in *D. melanogaster* but ancestral to genus *Drosophila*) is a close tandem paralog of SPN members but does not show correlated losses with any known SPN members. Also, *CG30486*, a close tandem paralog of *antares*, shows no correlation of loss events nor ERC. Previous genetic investigation of this gene showed no role in the SPN.

To more strongly query our hypothesis that correlation in loss was specific to proteins that are part of the SPN, we additionally investigated 30 other Sfps for shared patterns of loss with SPN proteins (Supplementary File 1, Table 13). We chose Sfps belonging to the same protein families as known SPN members but which are not closely related to any of them, and Sfps representing other protein families. We investigated whether any of these Sfps show losses along the same four lineages where all known SPN members are lost. We used reciprocal BLAST to screen for the presence/absence of these Sfps within genomes of the species where all network members were lost as well as a few closely related taxa where they remained intact. None of the examined Sfps showed losses along lineages in which all network genes were lost. This result suggests that the detected correlation in loss events between members of the network reflects their shared roles, rather than overall pressure on a similar category of proteins. The specificity of correlation of losses to members of the SPN or their paralogs (with the exception of *SP* and *seminase*), and not Sfps more broadly points to the usefulness of this metric in identifying functionally related proteins in a guilt by association-type approach. However, since the method requires gene loss, it can only be applied to orthogroups that are dispensable in some lineages and ancestral enough for repeated loss events to occur.

New SPN members can be identified as genes whose losses correlated with Sex Peptide Network Genes'

Interestingly, correlations in loss events were also observed between SPN members and some paralogous proteins whose SPN functions were unknown. Further, these Sfps showed correlated losses with each other. A striking example of this is the Sfp *CG42564*, which shows correlated losses with the SPN members to which it is closely paralogous, as well as to other unrelated SPN members. This result either indicates correlation in loss events unrelated to SPN function, perhaps related to its close evolutionary relationship to the SPN proteins or shared selective pressure acting broadly on Sfps, or could suggest that the

CG42564 gene is an uncharacterized member of the SPN. Supporting the latter hypothesis, in all four lineages where all known SPN members are lost, *CG42564* is also absent. To test whether *CG42564* is, in fact, a new member of the SPN, we assayed a long-term post-mating response in mates of males knocked down for *CG42564*, using two independent RNAi lines, and comparison to controls. In both independent knockdowns tested, mates of knockdown males regained receptivity to mating, in contrast to mates of control males (remating at 4 days; for line BDSC 65910 – 36/49 females (73.5%) mated to knockdown males vs 0/42 females mated to control males, for line VDRC 14158 – 46/54 females (85.2%) mated to KD males vs 0/23 females mated to control males; similarly, 0/48 mates of *tub-GAL4* males re-mated; Fisher's exact test for all knockdown/control remating comparisons: $p<0.0001$), indicating that *CG42564* is indeed a member of the SPN. Functional investigation of other genes that show correlated loss with SPN genes may reveal additional new members of the SPN.

Female genes that modulate Sex Peptide function downstream of SPN action do not show extensive CNV or correlation of gene duplications or losses with Sex Peptide or its network

The SPN Sfps transferred by the male during copulation are essential to SP's binding to sperm and long-term effects on the female, however, female factors are involved in mediating SP's downstream effects on female post-mating physiology and behavior independent of the SPN[13,35,44]. Most famously in *Drosophila melanogaster*, SPR (Sex Peptide Receptor) mediates many of SP's effects on the female[86]. However, despite their crucial interaction, there is no correlation in gene CNV between *SP* and *SPR*, with *SPR* consistently found as a single-copy gene in all species where it has been identified, and no correlation in sequence evolution of these two genes is seen by ERC[45].

Here we investigated CNV of three additional female-expressed genes that mediate downstream effects of SP in females (*Frma*, *Hdly*, and *Esp)* which, unlike *SPR*, were identified via correlated sequence evolution with *SP* via ERC[13] (Supplementary Figure 3). *Frma* is a neprilysin-like metalloendopeptidase, a class of proteins common among seminal fluid proteins; *Esp* is a member of the SLC26 family of anion exchangers; *Hdly* is an orphan gene with no known structural domains[49]. *SPR*, *Hdly*, and *Esp* originated prior to the emergence of SP and its associated network, with clear one-to-one orthologs identified in *Aedes aegypti*, the most distantly related outgroup examined, *Frma*'s origin appears to predate the common ancestor of the examined taxa (Supplementary File 1, Table 14). This suggests that while their specific functions and expression in the FRT across species are unknown, most of these genes are ancestral to Diptera, predating SP and the SPN. No significant CNV was observed for any of these genes across the genus, although *Hdly* and *Esp* were occasionally lost in some lineages. Therefore, as was the case for SPR, there is no CNV correlation between these genes with each other, *SP*, or the SPN.

While it is noteworthy that female proteins that mediate SP's downstream effects in *D. melanogaster* show uncorrelated CNV with *SP* and its network, this pattern may not be unexpected due to broader functional constraints impacting these genes' evolution. SPR, for example, has an alternative ligand and was likely co-opted as the SP receptor within the genus, limiting the extent of shared evolutionary pressures with SP[87]. Other female modulators may be similarly constrained: protein properties such as expression specificity, molecular role, or network connectivity can strongly influence tolerance to CNV[77]. Meanwhile, secreted proteins such as Sfps may be less constrained in CNV due to greater functional flexibility and lower structural constraints compared to transmembrane or signaling proteins like SPR and Esp[29–31]. Although both male and female reproductive genes have been shown to diversify rapidly between species, sometimes driven by positive selection, this pattern is generally observed to be more pronounced in male-expressed genes[1,4,88]. However, differences in gene category and function are frequently overlooked as contributors in these comparisons.

Conclusion

Lineage-specific evolutionary pressures –whether positive, negative, or relaxed selection– affect genes in ways that shape the evolution of their sequence, expression levels, and copy numbers. These effects leave detectable signatures that can be analyzed to generate testable hypotheses about underlying molecular mechanisms. Phylogenetic profiling methods harness such patterns to infer functional linkages between traits or molecular partners subject to shared evolutionary pressures[78,89–91]. Most famously among genomic approaches, Evolutionary Rate Covariation (ERC) identifies functionally related genes through correlated sequence changes along phylogenetic branches and has been successfully applied to diverse biological processes, including in *Drosophila*[13,14,92–95]. Indeed, many SPN proteins in our study were first identified through ERC[13]. Recent increases in available genome sequences spanning deep evolutionary timescales have expanded opportunities to use patterns of gene loss and gain to identify conserved members of protein complexes[23,25]. Here, by combining newly available, high-quality *Drosophila* genome sequences with a high-throughput approach for orthology and paralogy annotation via phylogenetic clustering, we characterized copy number changes within a recently evolved protein network that shows recurrent duplication and positive selection. Our detection of correlated loss events among SPN members validates this approach for uncovering functionally related proteins, and reveals one candidate, *CG42564,* as a new SPN member. Correlated evolutionary patterns in sequence, expression, copy number, and other phylogenetic profiling metrics are likely complementary, producing both overlapping and distinct sets of candidate genes. The nature of a gene's functional relationships may influence which correlations are detectable by different methods. For example, ERC can reflect direct or indirect interactions, whereas shared loss correlations (such as those we observed among SPN members) may primarily indicate more direct interaction, perhaps as members of a complex or direct pathway. ERC also detected downstream female modulators of SP function that were missed by shared loss analysis, consistent with pleiotropy and indirect roles. On the other hand, ERC may struggle to identify correlated patterns of evolution when complex paralogy and orthology relationships (common among the proteins we investigated) are present, although new strategies may mitigate this limitation[96]. Thus, combining ERC with duplication/loss correlations improves candidate identification: overlaps may highlight strong partners, while ERC-only hits may reflect indirect interactions. Both approaches, however, depend on conserved relationships across deep time and are limited by tolerance of duplications, multifunctionality, and lineage-specific copy number dynamics.

Data Availability

All supplemental tables described in the text can be found in Supplemental File 1. This file contains, among other supporting information, lists of publicly available transcriptomes and genomes used in the analysis, parameters for the sequence collection pipeline, and data underlying figures and statistical analysis. Supplemental File 2 and 3 contain the code used for identifying orthologs and a more detailed description of its usage respectively. Supplemental Figures can be found in File 4. We have also uploaded intermediate analysis files to a data repository upon publication.

Funding

This work was supported by NIH grants R01-HD059060 awarded to MFW and AGC and R37-HD038921 awarded to MFW, and Swiss NSF project grant 310030-212222 awarded to AVP. JAC was supported by NIH postdoctoral fellowship F32-HD111231 and by the Center for Vertebrate Genomics Distinguished Scholars Program.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

We thank Dr. Elissa Cosgrove with assistance in the bioinformatic analysis. We thank Drs. Nathan Clark and Matt Pennell for helpful discussions. Finally, we thank members of the Wolfner and Clark labs for helpful discussions and suggestions.

Citations

1. Carlisle, J. A. & Swanson, W. J. Molecular mechanisms and evolution of fertilization proteins. *J. Exp. Zoolog. B Mol. Dev. Evol.* **336**, 652–665 (2021).
2. Xie, X., Qiu, W.-G. & Lipke, P. N. Accelerated and Adaptive Evolution of Yeast Sexual Adhesins. *Mol. Biol. Evol.* **28**, 3127–3137 (2011).
3. Clark, N. L., Aagaard, J. E. & Swanson, W. J. Evolution of reproductive proteins from animals and plants. https://doi.org/10.1530/rep.1.00357 (2006) doi:10.1530/rep.1.00357.
4. Swanson, W. J. & Vacquier, V. D. The rapid evolution of reproductive proteins. *Nat. Rev. Genet.* **3**, 137–144 (2002).
5. Moyle, L. C., Wu, M. & Gibson, M. J. S. Reproductive Proteins Evolve Faster Than Non-reproductive Proteins Among Solanum Species. *Front. Plant Sci.* **12**, (2021).
6. Dapper, A. L. & Wade, M. J. Relaxed Selection and the Rapid Evolution of Reproductive Genes. *Trends Genet.* **36**, 640–649 (2020).
7. Rooney, A. P. & Zhang, J. Rapid evolution of a primate sperm protein: relaxation of functional constraint or positive Darwinian selection? *Mol. Biol. Evol.* **16**, 706–710 (1999).
8. Patlar, B., Jayaswal, V., Ranz, J. M. & Civetta, A. Nonadaptive molecular evolution of seminal fluid proteins in Drosophila. *Evolution* **75**, 2102–2113 (2021).
9. Panhuis, T. M., Clark, N. L. & Swanson, W. J. Rapid evolution of reproductive proteins in abalone and Drosophila. *Philos. Trans. R. Soc. B Biol. Sci.* **361**, 261–268 (2006).
10. Correia, R. B. *et al.* The conserved genetic program of male germ cells uncovers ancient regulators of human spermatogenesis. *eLife* **13**, (2024).

11. Clark, N. L. *et al.* Coevolution of Interacting Fertilization Proteins. *PLOS Genet.* **5**, e1000570 (2009).

12. Peng, J., Svetec, N., Molina, H. & Zhao, L. The Origin and Evolution of Sex Peptide and Sex Peptide Receptor Interactions. *Mol. Biol. Evol.* **41**, msae065 (2024).

13. Findlay, G. D. *et al.* Evolutionary Rate Covariation Identifies New Members of a Protein Network Required for Drosophila melanogaster Female Post-Mating Responses. *PLOS Genet.* **10**, e1004108 (2014).

14. Clark, N. L., Alani, E. & Aquadro, C. F. Evolutionary rate covariation reveals shared functionality and coexpression of genes. *Genome Res.* **22**, 714–720 (2012).

15. Wigby, S. *et al.* The Drosophila seminal proteome and its role in postcopulatory sexual selection. *Philos. Trans. R. Soc. B Biol. Sci.* **375**, 20200072 (2020).

16. Hopkins, B. R., Allen, S. E., Avila, F. W. & Wolfner, M. F. Male reproductive glands and their Secretions in insects. *Ref. Module Biomed. Sci.* https://benrhopkins.com/wp-content/uploads/2024/07/3-s2.0-b9780443214776001346-main.pdf (2024).

17. Sirot, L. K., Wong, A., Chapman, T. & Wolfner, M. F. Sexual Conflict and Seminal Fluid Proteins: A Dynamic Landscape of Sexual Interactions. *Cold Spring Harb. Perspect. Biol.* **7**, a017533 (2015).

18. Swanson, W. J., Clark, A. G., Waldrip-Dail, H. M., Wolfner, M. F. & Aquadro, C. F. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila. *Proc. Natl. Acad. Sci.* **98**, 7375–7379 (2001).

19. Majane, A. C., Cridland, J. M., Blair, L. K. & Begun, D. J. Evolution and genetics of accessory gland transcriptome divergence between Drosophila melanogaster and D. simulans. *Genetics* **227**, iyae039 (2024).

20. Fan, T., Cridland, J. M. & Begun, D. J. Adaptive gene expression parallelism in the male reproductive tract of two Drosophila species. *Genetics* iyaf113 (2025).

21. Haerty, W. *et al.* Evolution in the Fast Lane: Rapidly Evolving Sex-Related Genes in Drosophila. *Genetics* **177**, 1321–1335 (2007).

22. Garlovsky, M. D. & Ahmed-Braimah, Y. H. Evolutionary Quantitative Proteomics of Reproductive Protein Divergence in Drosophila. *Mol. Cell. Proteomics* **22**, (2023).

23. Kim, B. Y. *et al.* Highly contiguous assemblies of 101 drosophilid genomes. *eLife* **10**, e66405 (2021).

24. Kim, B. Y. *et al.* Single-fly genome assemblies fill major phylogenomic gaps across the Drosophilidae Tree of Life. *PLOS Biol.* **22**, e3002697 (2024).

25. Suvorov, A. *et al.* Widespread introgression across a phylogeny of 155 Drosophila genomes. *Curr. Biol.* **32**, 111-123.e5 (2022).

26. Lemoine, F. *et al.* Renewing Felsenstein's phylogenetic bootstrap in the era of big data. *Nature* **556**, 452–456 (2018).

27. Lutteropp, S., Kozlov, A. M. & Stamatakis, A. A fast and memory-efficient implementation of the transfer bootstrap. *Bioinformatics* **36**, 2280–2281 (2020).

28. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).

29. Feyertag, F., Berninsone, P. M. & Alvarez-Ponce, D. Secreted Proteins Defy the Expression Level–Evolutionary Rate Anticorrelation. *Mol. Biol. Evol.* **34**, 692–706 (2017).

30. Liao, B.-Y., Weng, M.-P. & Zhang, J. Impact of extracellularity on the evolutionary rate of mammalian proteins. *Genome Biol. Evol.* **2**, 39–43 (2010).

31. Winter, E. E., Goodstadt, L. & Ponting, C. P. Elevated Rates of Protein Secretion, Evolution, and Disease Among Tissue-Specific Genes. *Genome Res.* **14**, 54–61 (2004).

32. Hopkins, B. R. & Perry, J. C. The evolution of sex peptide: sexual conflict, cooperation, and coevolution. *Biol. Rev.* **97**, 1426–1448 (2022).

33. Avila, F. W., Mattei, A. L. & Wolfner, M. F. Sex peptide receptor is required for the release of stored sperm by mated *Drosophila melanogaster* females. *J. Insect Physiol.* **76**, 1–6 (2015).

34. Ram, K. R. & Wolfner, M. F. A network of interactions among seminal proteins underlies the long-term postmating response in Drosophila. *Proc. Natl. Acad. Sci.* **106**, 15384–15389 (2009).

35. Singh, A. *et al.* Long-term interaction between *Drosophila* sperm and sex peptide is mediated by other seminal proteins that bind only transiently to sperm. *Insect Biochem. Mol. Biol.* **102**, 43–51 (2018).

36. Mueller, J. L., Ripoll, D. R., Aquadro, C. F. & Wolfner, M. F. Comparative structural modeling and inference of conserved protein classes in Drosophila seminal fluid. *Proc. Natl. Acad. Sci.* **101**, 13542–13547 (2004).

37. Koppers, A. J., Reddy, T. & O'Bryan, M. K. The role of cysteine-rich secretory proteins in male fertility. *Asian J. Androl.* **13**, 111–117 (2011).

38. Gaikwad, A. S., Hu, J., Chapple, D. G. & O'Bryan, M. K. The functions of CAP superfamily proteins in mammalian fertility and disease. *Hum. Reprod. Update* **26**, 689–723 (2020).

39. Hu, J. *et al.* Epididymal cysteine-rich secretory proteins are required for epididymal sperm maturation and optimal sperm function. *Mol. Hum. Reprod.* **24**, 111–122 (2018).

40. Słowińska, M. *et al.* Characterization and biological role of cysteine-rich venom protein belonging to CRISPs from turkey seminal plasma†. *Biol. Reprod.* **104**, 1302–1321 (2021).

41. LaFlamme, B. A. & Wolfner, M. F. Identification and function of proteolysis regulators in seminal fluid. *Mol. Reprod. Dev.* **80**, 80–101 (2013).

42. Fijak, M., Hasan, H. & Meinhardt, A. Galectin-1 and galectin-3 in male reproduction - impact in health and disease. *Semin. Immunopathol.* **47**, 6 (2025).

43. Lad, P. M., Cooper, J. F., Learn, D. B. & Olson, C. V. Identification of structural and secretory lectin-binding glycoproteins of normal and cancerous human prostate. *Biochim. Biophys. Acta* **791**, 186–197 (1984).

44. Misra, S., Buehner, N. A., Singh, A. & Wolfner, M. F. Female factors modulate Sex Peptide's association with sperm in Drosophila melanogaster. *BMC Biol.* **20**, 279 (2022).

45. Hopkins, B. R. *et al.* Decoupled evolution of the *Sex Peptide* gene family and *Sex Peptide Receptor* in Drosophilidae. *Proc. Natl. Acad. Sci.* **121**, e2312380120 (2024).

46. McGeary, M. K. & Findlay, G. D. Molecular evolution of the sex peptide network in Drosophila. *J. Evol. Biol.* **33**, 629–641 (2020).

47. Tsuda, M., Peyre, J.-B., Asano, T. & Aigaki, T. Visualizing Molecular Functions and Cross-Species Activity of Sex-Peptide in Drosophila. *Genetics* **200**, 1161–1169 (2015).

48. Jenkins, V. K., Larkin, A. & Thurmond, J. Using FlyBase: A Database of Drosophila Genes and Genetics. in *Drosophila: Methods and Protocols* (ed. Dahmann, C.) 1–34 (Springer US, New York, NY, 2022). doi:10.1007/978-1-0716-2541-5_1.

49. Öztürk-Çolak, A. *et al.* FlyBase: updates to the Drosophila genes and genomes database. *Genetics* **227**, iyad211 (2024).

50. Pei, J., Kim, B.-H. & Grishin, N. V. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.* **36**, 2295–2300 (2008).

51. Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res.* **33**, W116-120 (2005).

52. Paysan-Lafosse, T. *et al.* InterPro in 2022. *Nucleic Acids Res.* **51**, D418–D427 (2023).

53. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinforma. Oxf. Engl.* **35**, 4453–4455 (2019).

54. Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information in 2025. *Nucleic Acids Res.* **53**, D20–D29 (2025).

55. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

56. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

57. EMBOSS Users Guide. http://emboss.open-bio.org/html/use/index.html.

58. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: Computing Large Minimum Evolution Trees with Profiles instead of a Distance Matrix. *Mol. Biol. Evol.* **26**, 1641–1650 (2009).

59. Katoh, K. & Standley, D. M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).

60. TREE2FASTA: a flexible Perl script for batch extraction of FASTA sequences from exploratory phylogenetic trees | BMC Research Notes | Full Text. https://bmcresnotes.biomedcentral.com/articles/10.1186/s13104-018-3268-y.

61. FigTree. https://tree.bio.ed.ac.uk/software/figtree/.

62. GeneRax: A Tool for Species-Tree-Aware Maximum Likelihood-Based Gene Family Tree Inference under Gene Duplication, Transfer, and Loss | Molecular Biology and Evolution | Oxford Academic. https://academic.oup.com/mbe/article/37/9/2763/5851843.

63. Rivera, A. M., Wilburn, D. B. & Swanson, W. J. Domain Expansion and Functional Diversification in Vertebrate Reproductive Proteins. *Mol. Biol. Evol.* **39**, msac105 (2022).

64. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing (2021).

65. Begun, D. J. & Lindfors, H. A. Rapid evolution of genomic Acp complement in the melanogaster subgroup of Drosophila. *Mol. Biol. Evol.* **22**, 2010–2021 (2005).

66. Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. Selection in the evolution of gene duplications. *Genome Biol.* **3**, research0008.1 (2002).

67. Lynch, M. & Conery, J. S. The Evolutionary Fate and Consequences of Duplicate Genes. *Science* **290**, 1151–1155 (2000).

68. Kelleher, E. S. & Markow, T. A. Duplication, Selection and Gene Conversion in a *Drosophila mojavensis* Female Reproductive Protein Family. *Genetics* **181**, 1451–1465 (2009).

69. Wagstaff, B. J. & Begun, D. J. Adaptive Evolution of Recently Duplicated Accessory Gland Protein Genes in Desert Drosophila. *Genetics* **177**, 1023–1030 (2007).

70. Findlay, G. D., Yi, X., MacCoss, M. J. & Swanson, W. J. Proteomics Reveals Novel Drosophila Seminal Fluid Proteins Transferred at Mating. *PLOS Biol.* **6**, e178 (2008).

71. Hurle, B., Swanson, W., NISC Comparative Sequencing Program & Green, E. D. Comparative sequence analyses reveal rapid and divergent evolutionary changes of the *WFDC* locus in the primate lineage. *Genome Res.* **17**, 276–286 (2007).

72. Ghenu, A.-H., Bolker, B. M., Melnick, D. J. & Evans, B. J. Multicopy gene family evolution on primate Y chromosomes. *BMC Genomics* **17**, 157 (2016).

73. Carlisle, J. A., Glenski, M. A. & Swanson, W. J. Recurrent Duplication and Diversification of Acrosomal Fertilization Proteins in Abalone. *Front. Cell Dev. Biol.* **10**, (2022).

74. Carlisle, J. A., Gurbuz, D. H. & Swanson, W. J. Recurrent Independent Pseudogenization Events of the Sperm Fertilization Gene ZP3r in Apes and Monkeys. *J. Mol. Evol.* **92**, 695–702 (2024).

75. Carlisle, J. A., Bickell, A. L. & Hart, M. W. Sexual selection and pseudogenization in primate fertilization. *Evolution* qpaf072 (2025).

76. Albalat, R. & Cañestro, C. Evolution by gene loss. *Nat. Rev. Genet.* **17**, 379–391 (2016).

77. Krylov, D. M., Wolf, Y. I., Rogozin, I. B. & Koonin, E. V. Gene Loss, Protein Sequence Divergence, Gene Dispensability, Expression Level, and Interactivity Are Correlated in Eukaryotic Evolution. *Genome Res.* **13**, 2229–2235 (2003).

78. Cavassim, M. I. A. *et al.* PRDM9 losses in vertebrates are coupled to those of paralogs ZCWPW1 and ZCWPW2. *Proc. Natl. Acad. Sci.* **119**, e2114401119 (2022).

79. LaFlamme, B. A., Avila, F. W., Michalski, K. & Wolfner, M. F. A Drosophila Protease Cascade Member, Seminal Metalloprotease-1, Is Activated Stepwise by Male Factors and Requires Female Factors for Full Activity. *Genetics* **196**, 1117–1129 (2014).

80. Allen, S. E., Chen, D. S., Misra, S. & Wolfner, M. F. Chapter 132 - Seminal metalloprotease-1. in *Handbook of Proteolytic Enzymes (Fourth Edition)* (eds Rawlings, N. D. & Auld, D. S.) 967–970 (Academic Press, 2025). doi:10.1016/B978-0-443-28849-4.00132-6.

81. van Hooff, J. J., Tromer, E., van Wijk, L. M., Snel, B. & Kops, G. J. Evolutionary dynamics of the kinetochore network in eukaryotes as revealed by comparative genomics. *EMBO Rep.* **18**, 1559–1571 (2017).

82. Moi, D., Kilchoer, L., Aguilar, P. S. & Dessimoz, C. Scalable phylogenetic profiling using MinHash uncovers likely eukaryotic sexual reproduction genes. *PLOS Comput. Biol.* **16**, e1007553 (2020).

83. Bloch, I. *et al.* Optimization of co-evolution analysis through phylogenetic profiling reveals pathway-specific signals. *Bioinformatics* **36**, 4116–4125 (2020).

84. Dobbelaere, J., Su, T. Y., Erdi, B., Schleiffer, A. & Dammermann, A. A phylogenetic profiling approach identifies novel ciliogenesis genes in Drosophila and C. elegans. *EMBO J.* **42**, e113616 (2023).

85. Ta, H. X., Koskinen, P. & Holm, L. A novel method for assigning functional linkages to proteins using enhanced phylogenetic trees. *Bioinformatics* **27**, 700–706 (2011).

86. Yapici, N., Kim, Y.-J., Ribeiro, C. & Dickson, B. J. A receptor that mediates the post-mating switch in Drosophila reproductive behaviour. *Nature* **451**, 33–37 (2008).

87. Kim, Y.-J. *et al.* MIPs are ancestral ligands for the sex peptide receptor. *Proc. Natl. Acad. Sci.* **107**, 6520–6525 (2010).

88. Papachristos, K., Sayadi, A. & Arnqvist, G. Comparative Genomic Analysis of the Pattern of Evolution of Male and Female Reproductive Proteins in Seed Beetles. *Genome Biol. Evol.* **16**, evae143 (2024).

89. Moi, D. & Dessimoz, C. Phylogenetic profiling in eukaryotes comes of age. *Proc. Natl. Acad. Sci. U. S. A.* **120**, e2305013120.

90. Ta, H. X., Koskinen, P. & Holm, L. A novel method for assigning functional linkages to proteins using enhanced phylogenetic trees. *Bioinformatics* **27**, 700–706 (2011).

91. Dobbelaere, J., Su, T. Y., Erdi, B., Schleiffer, A. & Dammermann, A. A phylogenetic profiling approach identifies novel ciliogenesis genes in Drosophila and C. elegans. *EMBO J.* **42**, e113616 (2023).

92. Clark, N. L., Alani, E. & Aquadro, C. F. Evolutionary Rate Covariation in Meiotic Proteins Results from Fluctuating Evolutionary Pressure in Yeasts and Mammals. *Genetics* **193**, 529–538 (2013).

93. Kowalczyk, A. *et al.* Evolutionary rate covariation identifies SLC30A9 (ZnT9) as a mitochondrial zinc transporter. *Biochem. J.* **478**, 3205–3220 (2021).

94. Tao, M. *et al.* Identification of a longevity gene through evolutionary rate covariation of insect mito-nuclear genomes. *Nat. Aging* **4**, 1076–1088 (2024).

95. Raza, Q. *et al.* Evolutionary rate covariation analysis of E-cadherin identifies Raskol as a regulator of cell adhesion and actin dynamics in Drosophila. *PLOS Genet.* **15**, e1007720 (2019).

96. Forsythe, E. S. *et al.* Phylogenomic prediction of interaction networks in the presence of gene duplication. 2024.08.06.606904 Preprint at https://doi.org/10.1101/2024.08.06.606904 (2024).

97. Allen, S. Seminal Insights: A Multidisciplinary Approach to Understanding Sex Peptide in Drosophila Reproductive Biology. (Cornell University, United States -- New York, 2024).

98. McCullough, E. L. *et al.* The life history of Drosophila sperm involves molecular continuity between male and female reproductive tracts. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2119899119 (2022).

99. dos Santos, G. *et al.* FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Res.* **43**, D690–D697 (2015).

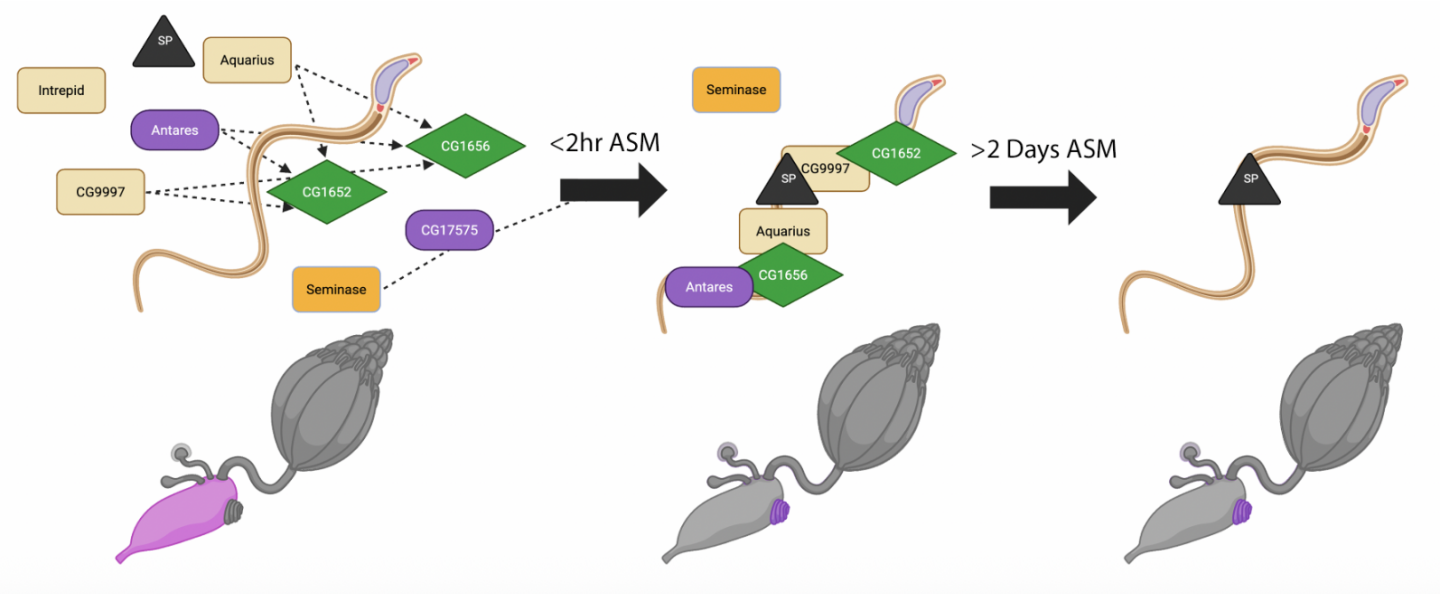


Figure 1. Seminal fluid proteins in the Sex Peptide Network are critical for Sex Peptide binding to sperm and for SP's long-term effects on female reproduction. Figure adapted from Allen 2024[97]. While all members of the Sex Peptide Network (SPN) are critical for storage of sperm with bound Sex Peptide (SP), there is evidence that these proteins mediate different steps of a cascade of events that allow for the long-term effects of SP[35]. These long-term effects are mediated through SP's long-term binding to sperm stored in the seminal receptacle, the primary sperm storage organ. Some of these seminal fluid proteins (Sfps) (CG1652 and CG1656) require other network members (Aqrs, Antr, CG9997) for proper transfer to the female. However, while essential for transfer of CG1652 and CG1656, Antr's transfer is independent of other known network members. After being transferred to the female, Sems and CG17575 do not bind sperm nor are stored in the seminal receptacle, but they are critical for SP and CG1656 to localize to the seminal receptacle where both proteins (along with CG1652, Antr, and CG9997) bind sperm. Sems can be detected in the seminal receptacle despite not binding sperm. While SP binds to sperm long-term, its network does not. Most of the SPN Sfps are no longer detectable on sperm 1 day after mating and none are detectable after 4 days although SP can still be found bound to sperm[35,98]. The functions of all SPN Sfps have not been fully characterized. For example, Intr's function within the network has not been described, and additional members of the network had yet to be identified. Two of the SPN Sfps, *CG1652* and *CG9997*, evolved under positive selection[46]. Of the 8 known members of the SPN, 4 are serine proteases, 2 are lectins, and 2 are proteins containing a CAP domain. Seminal fluid proteins across species, including in mammals, have been shown to contain these protein domains. Known interactions are indicated by dashed lines. The first panel depicts transfer of Sfps and sperm during copulation into the bursa, the second panel depicts SP and SPN proteins bound to sperm 2 hours after the start of mating, the third panel depicts only SP detectable as bound to sperm more than 2 days after mating. Orange squares are serine protease homologs of the SPN with the darker square being an active protease, purple ovals are CAP domain-containing proteins, green diamonds are lectins, and the black triangle is SP. Figure made with BioRender.

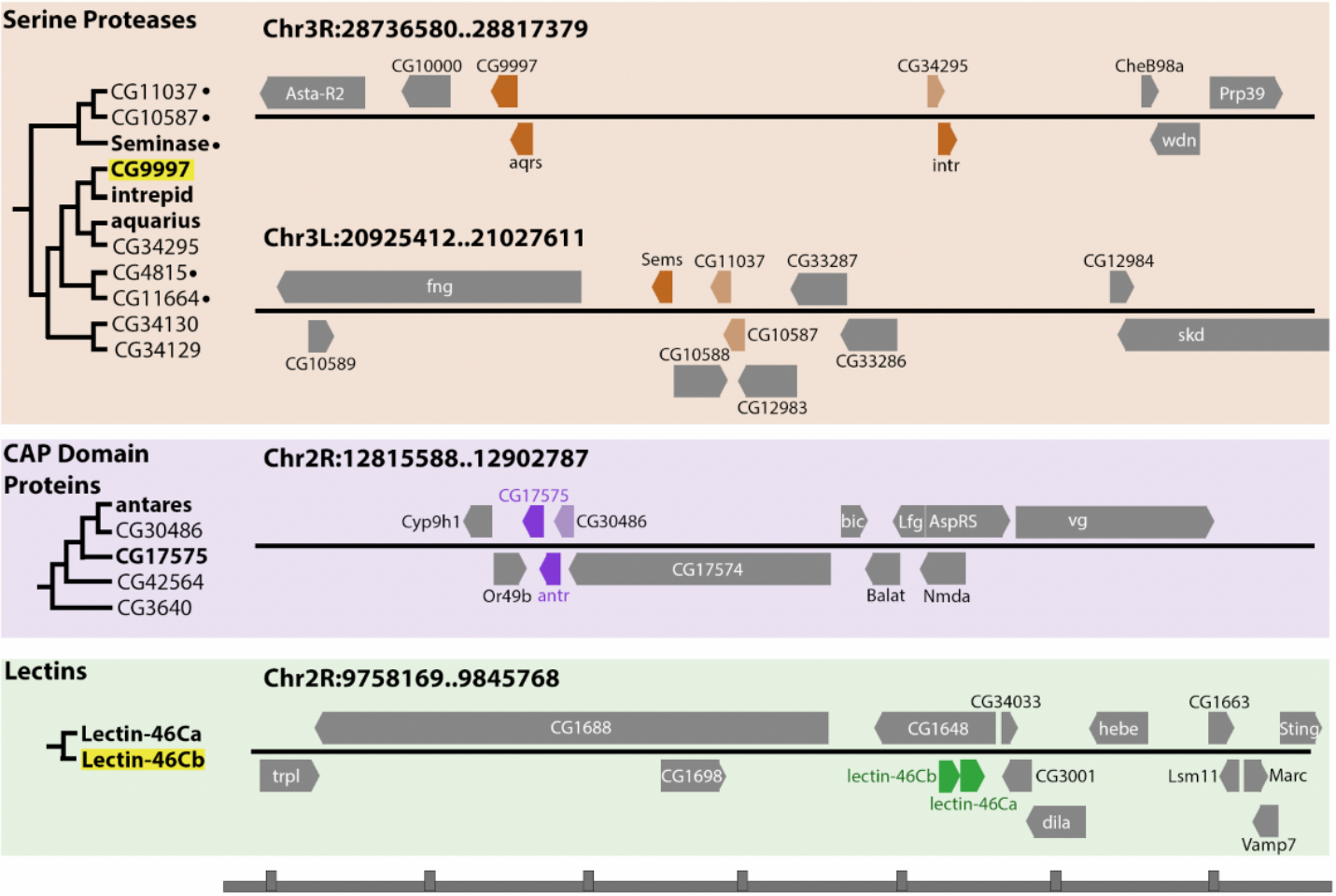


Figure 2: Sex Peptide Network seminal fluid proteins arose through gene-duplication. SPN proteins are closely paralogous to other Sfps including some that are known members of the SPN. Protein trees are derived from the phylogenetic clustering described in the Methods to establish orthology and paralogous relationships and supported by >50 bootstraps. All SPN proteins are bolded and they, along with their paralogs, are all known Sfps[15]. Circles to the right of serine proteases protein names indicate that they are predicted to be active proteases, others are annotated as inactive proteases. Highlighted genes have been shown to evolve under positive selection[46]. Genomic regions containing SPN proteins in *D. melanogaster* (r6.65) are depicted[99]. Tandem duplication appears to contribute to the origin of these Sfps. Darker colored boxes are SPN members, lighter colored boxes are paralogous Sfps whose SPN function is not known. ORFs and distances are approximate and are not shown perfectly to scale. Ticks on x-axis represent ~12.5 kb.

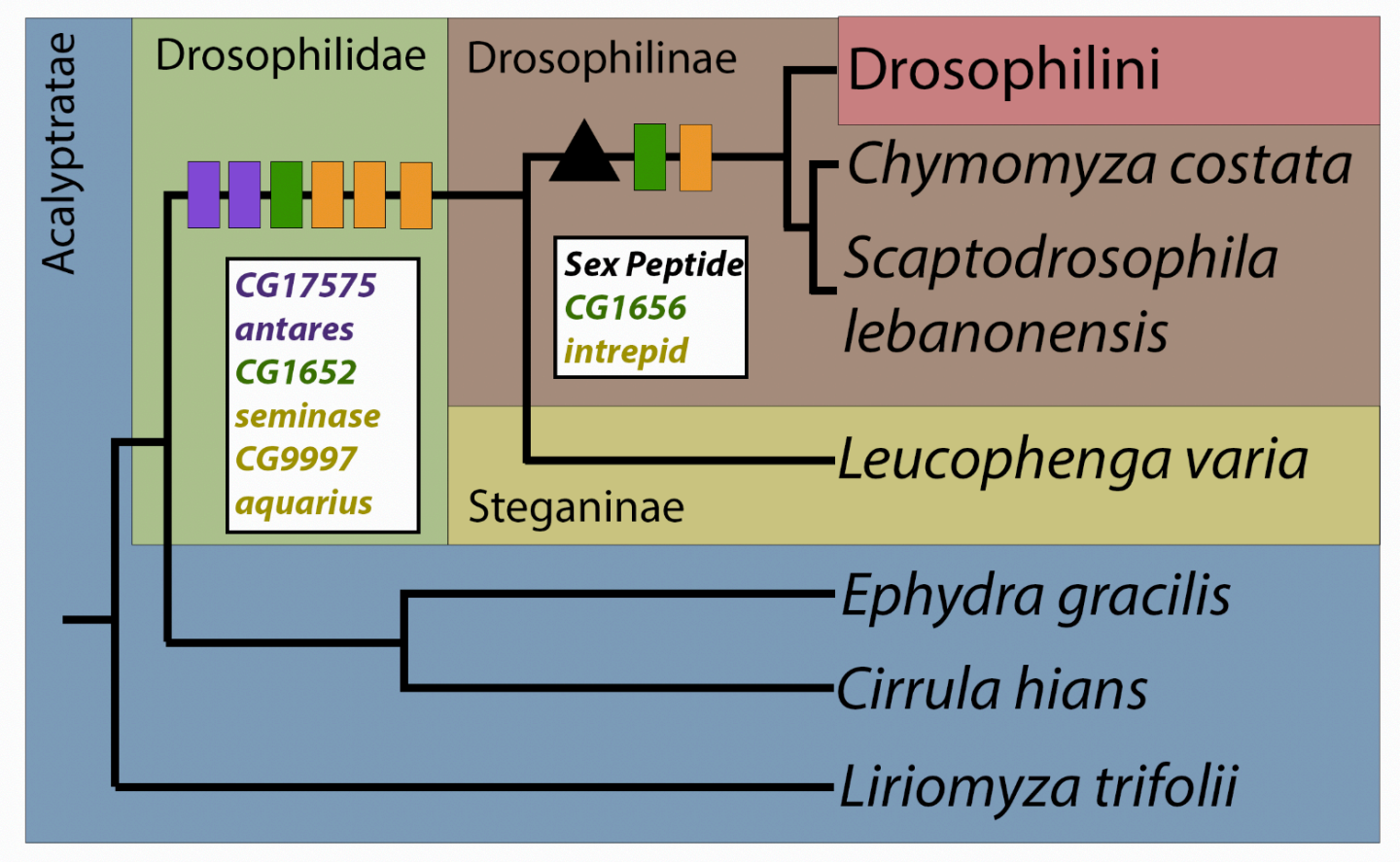


Figure 3. Many SPN Sfp genes arose before the Sex Peptide gene. Many of the SPN Sfp genes (*CG17575*, *antares*, *CG1652*, *seminase*, *aquarius* and *CG9997*) arose in *Drosophilidae*, and predated the origin of SP in *Drosophilinae*[12,45]. Along the lineage where SP arose, a duplication of *CG1652* and a duplication event of *CG9997* led to the origin of the two other known members of the SPN (*CG1656* and *intrepid* respectively).

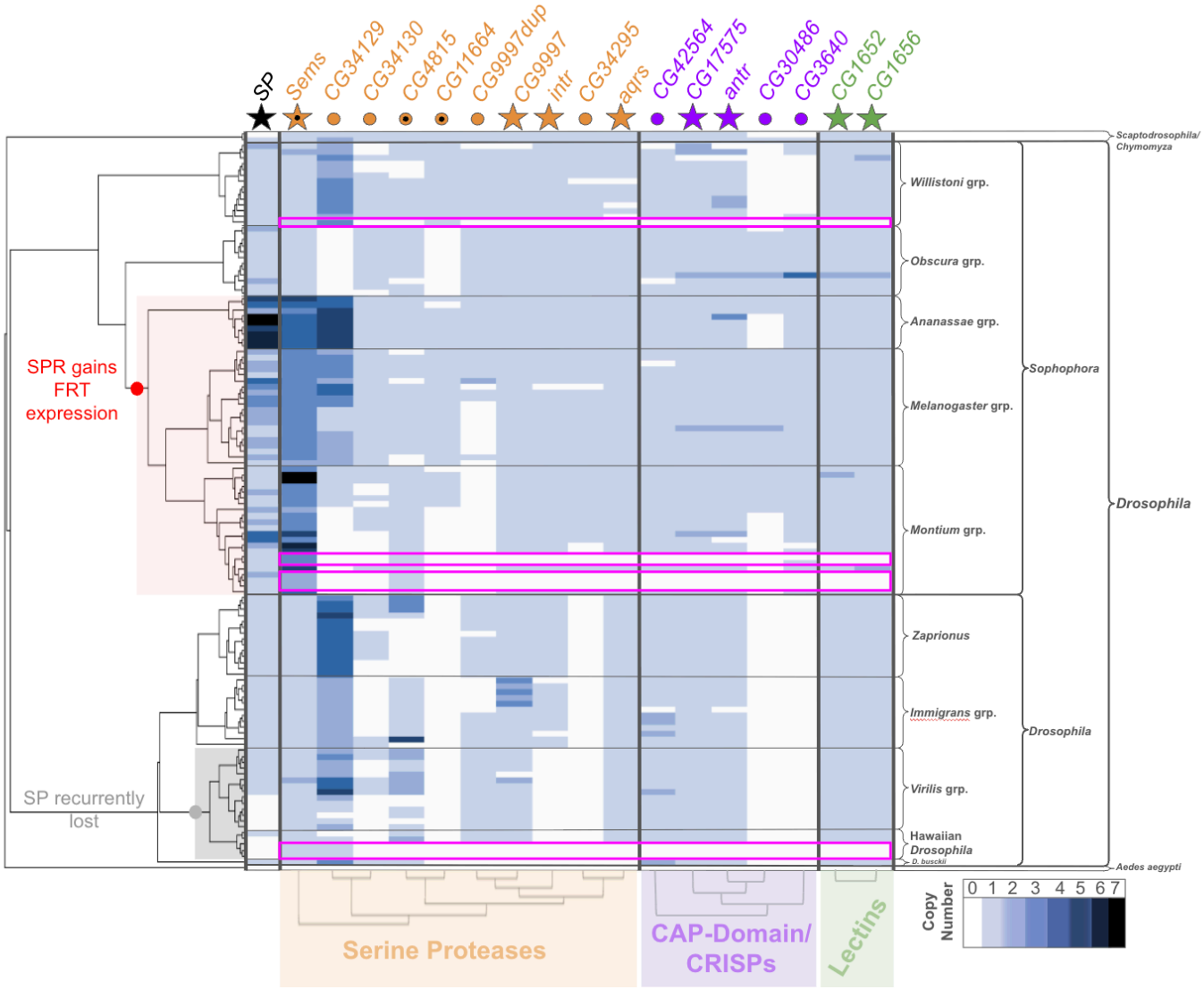


Figure 4. Copy number variation of SPN seminal fluid proteins in the genus *Drosophila*. Copy numbers of SPN genes and their close paralogs for 135 Drosophilids inferred via iterative blast and phylogenetic clustering (except *SP*, for which copy number data were obtained from Hopkins *et al.* (2024)). Genes are colored by protein type; Black = Sex Peptide; Orange = Serine Proteases; Purple = CAP-domain genes/CRISPs; Green = Lectins. Stars denote genes with known SPN function in *D. melanogaster*. For serine proteases, those stars/circles marked with a black dot denote active proteases, whereas unmarked shapes represent inactive protease homologs. The four main concerted losses of most network genes we identified are bordered by pink boxes. The point on the species tree where SPR was inferred to have gained expression in the female reproductive tract is marked in red[47]. Similarly, lineages where *SP* is recurrently lost are marked in gray. CNV data underlying the figure can be found in Supplementary File 1, Table 4.

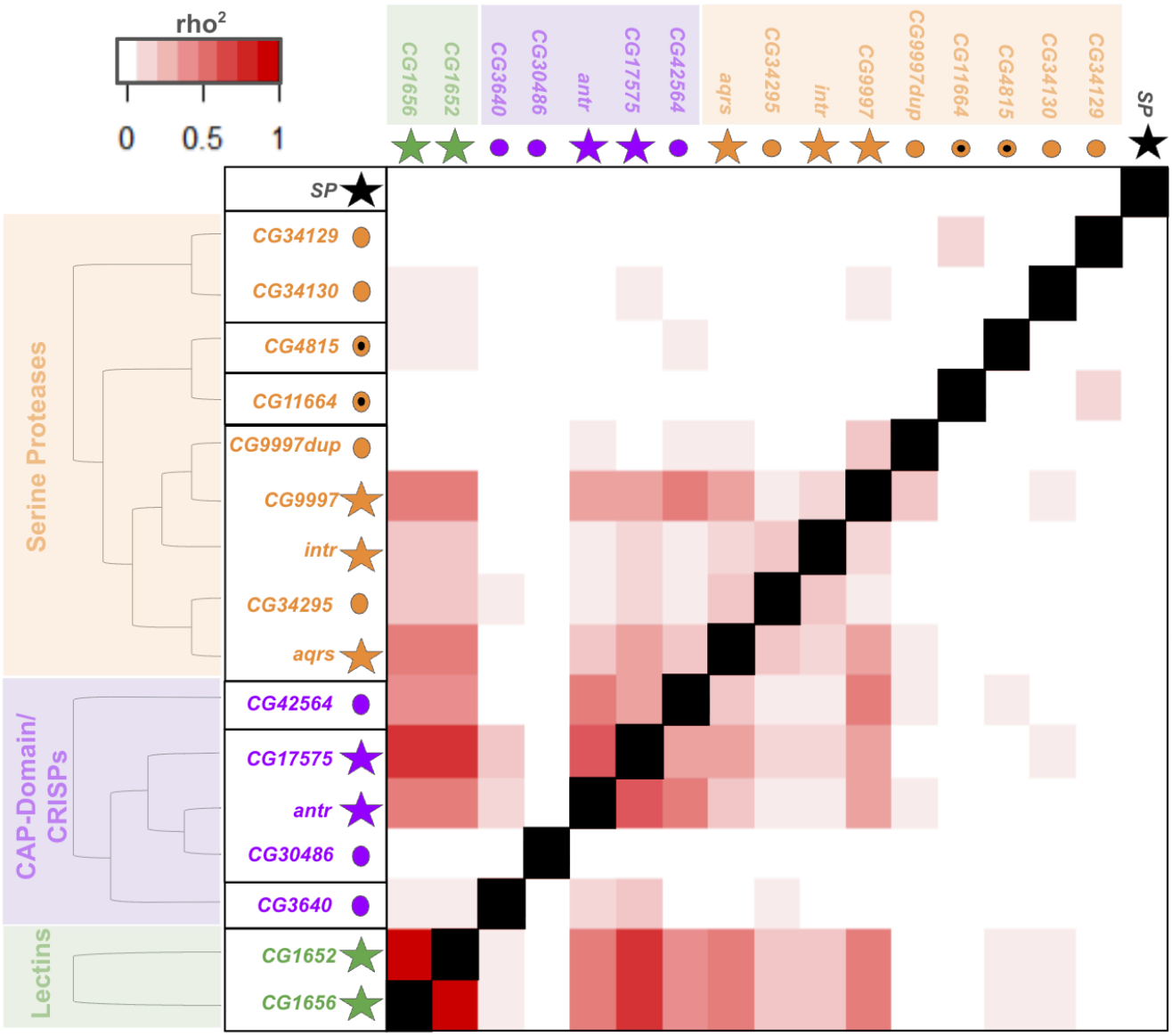
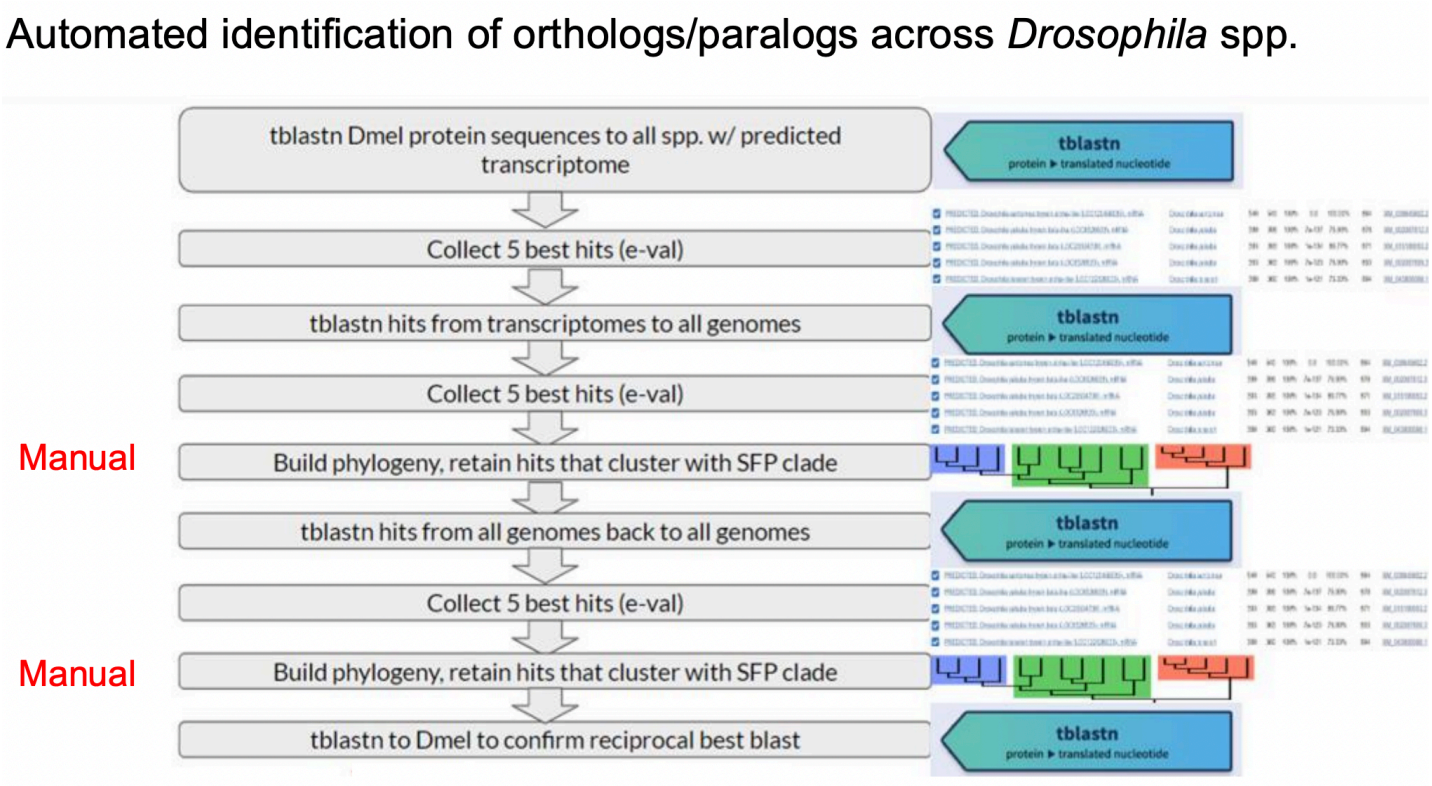
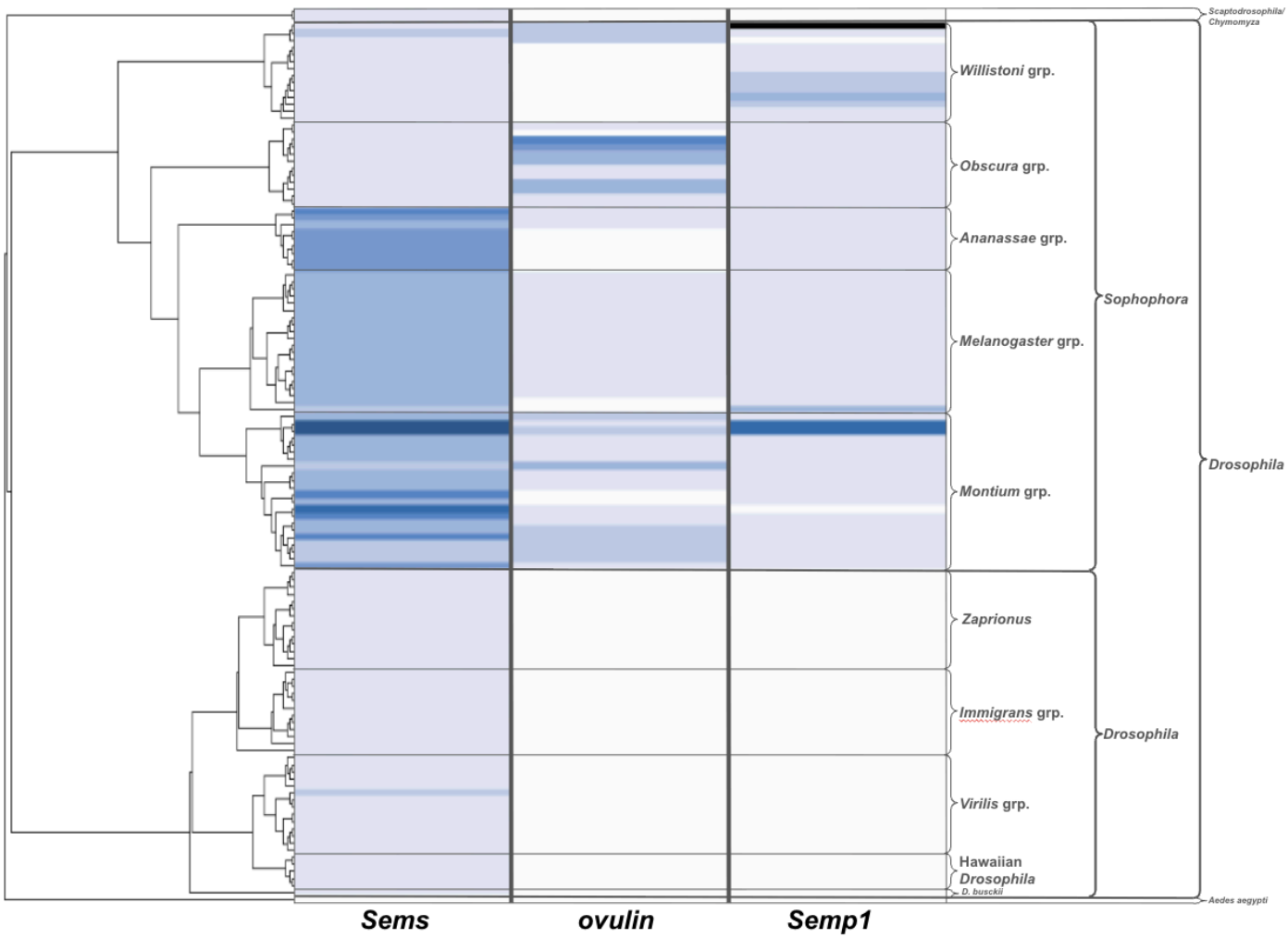


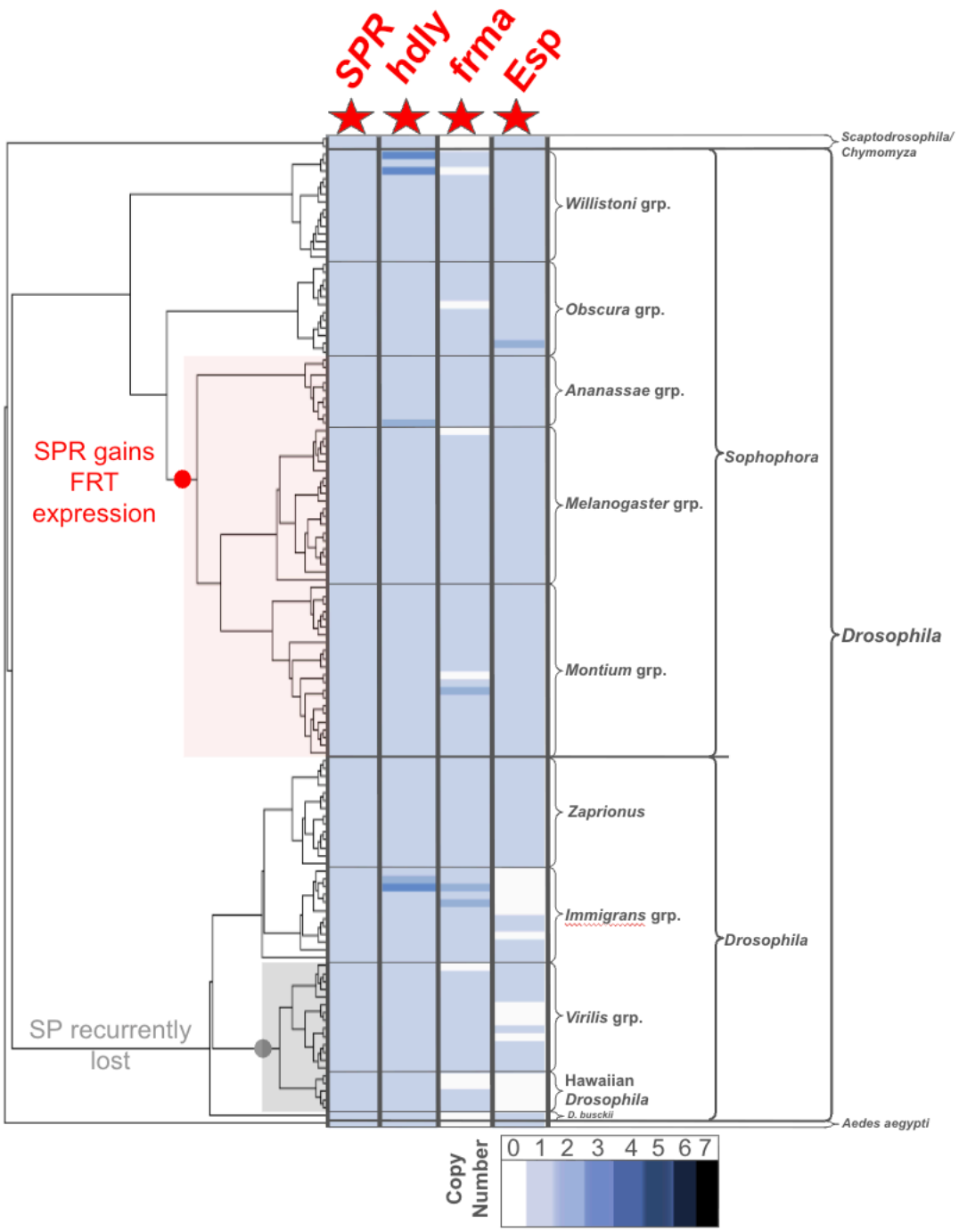
Figure 5. Correlated loss events between members of the SPN. Pairwise Spearman's rank correlation coefficients (rho; squared for visualization purposes) between SPN genes and their close paralogs for concerted total orthogroup losses by branch (associated values can be found in Supplementary File 1, Table 7). The location of loss events on the species tree was inferred by gene tree-species tree reconciliation. For the correlations, each branch (n = 258) was assigned a value (0 or 1) denoting whether or not a loss occurred on that branch, and the resulting vectors for each orthogroup were correlated. Genes are colored by protein type; Black = Sex Peptide; Orange = Serine Proteases; Purple = CAP-domain genes/CRISPs; Green = Lectins. Stars denote genes with known SPN function in *D. melanogaster* prior to this study. For serine proteases, those stars/circles marked with a black dot denote active proteases, whereas unmarked shapes represent inactive protease homologs.



Supplemental Figure 1: Graphical overview of computational pipeline used for sequence collection and determination of orthology/paralogy relationships via iterative blast and phylogenetic clustering. See Materials and Methods for a more detailed description of the approaches used. Associated code can be found in Supplementary File 2 and additional instruction on implementation in Supplementary File 3. Briefly, *D. melanogaster* protein sequences representing all orthogroups in a protein family of interest were used as the query of a tblastn search to the predicted transcriptomes of 35 *Drosophila* species. The five best hits were then collected for each query protein. The corresponding amino acid sequences were subsequently obtained from NCBI and themselves used as the query for a new tblastn search, this time to the genomes of 135 Drosophilids. The five best hits for each query were translated, aligned, and used to build a protein tree containing the hits recovered from all species. The hits that convincingly clustered with the gene family of interest on this tree used as the query of a final tblastn search of the same 135 genomes. The subsequent filtering and phylogenetic clustering were repeated, and final orthogroup designations manually annotated on the resulting tree. To additional validate our annotations, we confirmed that each protein sequence, when used as the query of a tblastn search of the *D. melanogaster* predicted transcriptome, hits the expected *D. melanogaster* transcript.



Supplementary Figure 2. Copy number variation in the genus *Drosophila* of other Sfps functionally related to *Sems*. Copy numbers of *Sems*, *ovulin*, and *Semp1* genes and their close paralogs for 135 Drosophilids inferred via iterative blast and phylogenetic clustering. CNV of *Semp1* is not correlated with *seminase* or *SP*.



Supplementary Figure 3. Copy number variation of genes that mediate downstream effects of SP in females and SP's receptor in the genus *Drosophila*. CNV of SPR from Hopkins *et al* 2024[45]. *Frma*, *Hdly*, and *Esp* CNV determined via same pipeline used on SPN genes. No significant CNV was observed for any of these genes across the genus, although *Hdly* and *Esp* were occasionally lost in some lineages. Therefore, as was the case for SPR, there is no CNV correlation between these genes with each other, *SP*, or the SPN.