

UC Davis

UC Davis Previously Published Works

Title

Single-nucleus transcriptomes reveal evolutionary and functional properties of cell types in the *Drosophila* accessory gland

Permalink

<https://escholarship.org/uc/item/5b81z2w7>

Journal

Genetics, 220(2)

ISSN

0016-6731

Authors

Majane, Alex C

Cridland, Julie M

Begun, David J

Publication Date

2022-02-04

DOI

10.1093/genetics/iyab213

Peer reviewed

Single-nucleus transcriptomes reveal evolutionary and functional properties of cell types in the *Drosophila* accessory gland

Alex C. Majane*, Julie M. Cridland, and David J. Begun

Department of Evolution and Ecology, University of California – Davis, Davis, CA 95616, USA

*Corresponding author: Email: acmajane@ucdavis.edu

Abstract

Many traits responsible for male reproduction evolve quickly, including gene expression phenotypes in germline and somatic male reproductive tissues. Rapid male evolution in polyandrous species is thought to be driven by competition among males for fertilizations and conflicts between male and female fitness interests that manifest in postcopulatory phenotypes. In *Drosophila*, seminal fluid proteins secreted by three major cell types of the male accessory gland and ejaculatory duct are required for female sperm storage and use, and influence female postcopulatory traits. Recent work has shown that these cell types have overlapping but distinct effects on female postcopulatory biology, yet relatively little is known about their evolutionary properties. Here, we use single-nucleus RNA-Seq of the accessory gland and ejaculatory duct from *Drosophila melanogaster* and two closely related species to comprehensively describe the cell diversity of these tissues and their transcriptome evolution for the first time. We find that seminal fluid transcripts are strongly partitioned across the major cell types, and expression of many other genes additionally defines each cell type. We also report previously undocumented diversity in main cells. Transcriptome divergence was found to be heterogeneous across cell types and lineages, revealing a complex evolutionary process. Furthermore, protein adaptation varied across cell types, with potential consequences for our understanding of selection on male postcopulatory traits.

Keywords: *Drosophila*; evolution; gene expression; selection; cell types; accessory gland; ejaculatory duct; reproduction; population genetics; single-cell RNA-seq

Introduction

Identifying and explaining variance in rates of evolution, which is commonly observed at all levels of biological organization, has been one of the great preoccupations of evolutionary biology. For example, some genes, proteins, and chromosomes evolve more quickly than others (White 1977; Kimura 1983), some traits evolve quickly in some lineages and slowly in others (Simpson 1944), and some traits evolve much more quickly in males than in females (Darwin 1871). This truism of evolutionary biology, that evolutionary rate variance is common and demands an explanation, extends to gene expression phenotypes, which tend to evolve relatively quickly in male reproductive tissues compared with most other tissues (reviewed in Ellegren and Parsch 2007). While the explanations proffered for faster expression evolution in male reproductive tissues often invoke rapidly changing selection pressures due to sexual selection or genomic conflicts, the biological processes driving rapid divergence of male reproductive tissues remain mostly unknown. Because the level of biological organization at which an evolutionary phenomenon is measured fundamentally shapes our understanding of evolutionary patterns, the level of analysis necessarily constrains the universe of testable hypotheses and the generation of new

hypotheses. In the context of *Drosophila* gene expression, the phenomenology of rapid male-biased expression divergence has often been observed at the whole animal level or the organ level (focusing primarily on gonads) (Meiklejohn et al. 2003; Ranz et al. 2003; Assis et al. 2012; Whittle and Extavour 2019). In reality, most organs are a complex mixture of many cell types, which suggests that while organ analysis is preferable to whole-animal analyses, layers of biological causation and evolutionary inferences are still missed. Indeed, since gene products are produced in individual cells, one could reasonably argue that the cell is the natural level of organization for understanding expression variation and generating hypotheses relating expression variation to downstream phenotypes.

Theoretical concepts underlying the evolution of cell-type diversity and the process of evolution in different cell types within a tissue are well-developed (reviewed in Musser and Wagner 2015; Arendt et al. 2016). Single-cell data in evolutionary contexts have generally been applied to distantly related taxa (Liang et al. 2018; Tosches et al. 2018; Hodge et al. 2019), typically focusing on cell-type diversity (La Manno et al. 2016; Seb e-Pedr os et al. 2018; Colquitt et al. 2021; Feregrino and Tschopp 2021; Wang et al. 2021). Evolutionary analysis of different cell types across species,

particularly on short-time scales, has received less attention (Liang et al. 2018). In this study, we use the polyandrous genus *Drosophila* as a model for evolution at the cellular level, with a focus on the tissues producing seminal fluid proteins (Sfps), which are transferred to females along with the sperm during mating. Many of these secreted proteins, which are produced in the accessory glands (AGs) and the ejaculatory duct, induce a set of physiological and behavioral changes in females collectively referred to as the postmating response (PMR; reviewed in Ravi Ram and Wolfner 2007). In *D. melanogaster*, the PMR includes increased rates of egg laying (Soller et al. 1999; Heifetz et al. 2000), decreased receptivity to remating (Liu and Kubli 2003), storage of sperm in specialized reproductive tract tissues (Neubaum and Wolfner 1999), elevated immune response (Peng et al. 2005), elevated feeding rates (Carvalho et al. 2006), increased activity rate, and decreased sleep (Isaac et al. 2010). Genetic variation in Sfps may also play a role in the outcome of sperm competition (Clark et al. 1995; Fiumera et al. 2005). Population genetic and comparative analyses of these proteins suggest they evolve unusually rapidly, often under the influence of directional selection (Tsauro et al. 1998; Aguadé 1999; Begun et al. 2000). These genes are frequently gained or lost during evolution (Mueller et al. 2005; Wagstaff and Begun 2005), even on short timescales (Begun and Lindfors 2005) and experimental evolution has shown that sexual conflict linked to PMR phenotypes may contribute to the rapid evolution of seminal fluid proteins (Hollis et al. 2019).

The *D. melanogaster* AG consists of two specialized, morphologically distinct, secretory epithelial cell types (Bairati 1968). Main cells (MC) are smaller, hexagonal, and squamous, while secondary cells (SC) are much larger, spherical, project into the lumen of the gland, and contain extensive vacuole-like compartments (Bairati 1968; Prince et al. 2019). MC, which constitute the vast majority of AG cells, are necessary and sufficient to initiate the PMR (Kalb et al. 1993; Sitnik et al. 2016; Hopkins et al. 2019). SC, which are located at the distal tip of the gland, appear to contribute in part to the long-term maintenance of the PMR, particularly with respect to remating phenotypes; females mated to males with deficient SC secretions exhibit greater receptivity to remating (Leiblich et al. 2012; Hopkins et al. 2019). It is difficult to dissect individual phenotypic contributions of each cell type, however, given their apparent interdependence in production of the seminal fluid (Hopkins et al. 2019). The ejaculatory duct consists of a single secretory epithelial cell type (Bairati 1968), contributing additional Sfps to the ejaculate (Rexhepaj et al. 2003; Takemori and Yamamoto 2009; Sepil et al. 2018). While the duct and its products contribute to the PMR (Xue and Noll 2000; Saudan et al. 2002; Rexhepaj et al. 2003), relatively little experimental work has been performed on this tissue.

While genetic and gene expression studies of the AG have revealed evidence of both shared and distinct properties of these three major cell types, and much has been learned from genetic mutants knocking out (Kalb et al. 1993; Xue and Noll 2000; Minami et al. 2012; Gligorov et al. 2013; Sitnik et al. 2016) or suppressing secretions of (Leiblich et al. 2012; Corrigan et al. 2014; Hopkins et al. 2019) specific cell types in the AG, no study has directly investigated patterns of cell-type expression bias from transcriptome data. Here, we carry out single-cell transcriptome analysis of the AG and ejaculatory duct in three closely related *Drosophila* species, *D. melanogaster*, *D. simulans*, and *D. yakuba*. We characterize MC, SC, and ejaculatory duct cells (EDC) to (1) reveal new biological attributes of the various cell types in the male somatic reproductive tract, (2) investigate rates of transcriptome divergence at the cellular level in multiple lineages, (3) determine

the degree to which expression evolution is concerted or independent across cell types, and (4) investigate the connection between cell-type-biased gene expression and adaptive protein divergence.

Methods

Fly stocks and single-nucleus RNA sequencing

Additional details of all methods in this study can be found in [Supplementary Material](#). We used the following sequenced stocks to generate AG and ejaculatory duct transcriptomes from 2- to 3-day-old virgin males for three *melanogaster* subgroup species: *D. melanogaster* RAL 517 (Mackay et al. 2012), *D. simulans* *w*⁵⁰¹, and *D. yakuba* Tai18E2 (hereafter referred to as *mel*, *sim*, and *yak*) (Begun et al. 2007). Nuclei were isolated into a suspension using a modified version of Luciano Martelotto's protocol (2019). FACS was used to purify single nuclei, and single-nucleus RNA-Seq libraries were created using the 10× Genomics Chromium platform and Illumina sequencing.

Bioinformatic assignment of species origin, RNA-Seq alignment, QC, and ortholog formatting

We parsed the 10× barcodes of raw reads and counted the number of unique molecular identifiers (UMIs) corresponding to each. We examined the distribution of UMI counts in descending rank order, using the “knee” inflection point method (Macosko et al. 2015) to identify putative nuclei and empty barcodes. We used a custom alignment-based bioinformatic pipeline (github.com/alexmajane/AG_single_nucleus) to assign species-of-origin to each nucleus. We aligned reads to the appropriate species genome (Flybase; *D. melanogaster* v6.33, *D. simulans* v2.02, *D. yakuba* v1.05) using STAR v2.7.5a (Dobin et al. 2013) with default parameters. We then filtered the set of nuclei according to alignment statistics to remove probable multiplets and nuclei with low sequencing depth. Next, we counted features from BAM files using HTSeq-count v0.12.3 (Anders et al. 2015) with default parameters. For comparative analyses we created a set of 1-to-1-to-1 orthologs (11,481 genes) using the *D. melanogaster* ortholog table from Flybase (http://ftp.flybase.net/releases/FB2020_02/precomputed_files/orthologs/dmel_orthologs_in_drosophila_species_fb_2020_02.tsv.gz [downloaded September 21, 2020]).

Marker gene identification and differential expression among species

Single-nucleus gene expression analyses were performed in R v3.6.1 using Seurat v3.2.2 (Satija et al. 2015; Butler et al. 2018; Stuart et al. 2019) using two parallel approaches. We did an integrated analysis (Stuart et al. 2019) of the data across species using our *mel/sim/yak* 1-to-1-to-1 orthologs. We also performed an independent analysis of *mel* using all annotated genes to gain a fuller picture of gene expression variation among cell types. We identified marker genes using Seurat's FindAllMarkers() method and assessed significance using a Wilcoxon rank-sum test. We required marker genes to be expressed in at least 25% of focal cluster cells and set a minimal average log₂(fold-change), hereafter referred to as logFC, requirement of 0.25. We filtered marker genes to those with Bonferroni-corrected P-values <0.05. To further investigate cell-type-specific expression bias of all Sfps, in addition to those strictly classified as marker genes, we did not impose minimum % cells expressing and average logFC thresholds. We additionally identified markers distinguishing MC subpopulations from one another using the FindMarkers() method. To further characterize these subpopulations, we estimated pseudotime using Slingshot (Street et al. 2018) and identified

dynamically differentially expressed (DE) genes with tradeSeq (Van den Berge et al. 2020).

We used limma v3.42.2 (Ritchie et al. 2015) to infer DE genes for each cell type. We performed pairwise contrasts among the three species and classified genes as DE with an FDR of 5% (Benjamini and Hochberg 1995). Further details of the limma analysis can be found in our R scripts (github.com/alexmajane/AG_single_nucleus). To compare the rate of qualitative expression divergence across cell types, we calculated ratios of DE genes at various logFC cutoffs across the three cell types for each of the three pairwise species contrasts, and tested for differences in these ratios using a G-test of goodness-of-fit (Sokal and Rohlf 2012). To test for differences in the magnitude of expression differences across cell types, we similarly compared distributions of absolute values of logFC using a Kruskal–Wallis test (Kruskal and Wallis 1952). Finally, we examined overall expression correlations between species within cell types by calculating average expression per gene and Pearson’s correlation coefficients.

To examine the relative level of concerted vs independent gene expression evolution across cell types, we subset the data to the set of DE genes exhibiting a logFC greater than one in at least one cell-type-specific pairwise species contrast. We then calculated pairwise Pearson’s correlation coefficients of logFC across cell types within each of the three pairwise species contrasts. We permuted logFC values across genes 10,000 times to obtain a distribution of Pearson’s correlation coefficients under the null expectation of entirely cell-type-independent change within our set of DE genes.

Population genetic inference of adaptive protein divergence of marker genes

To investigate potential differences in the prevalence of adaptive protein evolution across cell types, we used existing population data from *D. melanogaster* (Fraïsse et al. 2019) with *D. simulans* as the outgroup. We considered two summaries of the role of adaptation in protein divergence (McDonald and Kreitman 1991; Smith and Eyre-Walker 2002): the proportion of marker genes with $\alpha > 0$, and the distribution of α values amongst those genes with $\alpha > 0$. The proportions of positive α values were compared using Fisher’s exact test, with post hoc pairwise tests between cell types. The distributions of positive α values were visualized in ggplot2 v3.3.3 (Wickham 2016), and compared using a Kruskal–Wallis test with post hoc pairwise Wilcoxon tests.

To determine whether the prevalence of positive selection in AG-expressed genes correlates with differential gene expression, we intersected α values with DE genes. We selected the set of all genes expressed in the AG and filtered out genes expressed at a level lower than the lowest-expressed DE gene, to account for power to detect DE. We tested whether DE genes and non-DE genes had different likelihoods of showing positive selection by comparing the fraction of positive α values in each class of genes using a G-test. We tested whether the fraction of sites with evidence of positive selection differed among classes of genes by comparing distributions of positive α values using a Kruskal–Wallis test.

To catalog non-SFP genes narrowly expressed in the AG with evidence of recurrent protein adaptation, we used the index of tissue specificity, τ (Yanai et al. 2005), which we previously computed (Cridland et al. 2020) using FlyAtlas2 RNA-Seq data (Leader et al. 2018). We selected genes with the greatest expression in the AG and values of $\tau > 0.9$, indicative of highly AG-specific expression, $\alpha > 0.5$, and at least five fixed nucleotide substitutions, leading to a limited list of candidate non-SFPs with AG-specific

expression that may have undergone adaptive protein divergence between *mel* and *sim*.

De novo transcriptome assembly and identification of unannotated *D. melanogaster* transcripts

For *de novo* transcriptome assembly, we trimmed reads with TrimGalore! v0.6.5 (github.com/FelixKrueger/TrimGalore) and used Trinity v2.11.0 (Grabherr et al. 2011) to create the assembly. We augmented our assembly with two additional bulk RNA-Seq datasets (Leader et al. 2018; Immarigeon et al. 2021) (see Supplementary Methods). We quantified abundances of *de novo*-assembled transcripts in each cell-type population with Salmon v0.12.0 (Patro et al. 2017). We used a BLAST-based strategy (Camacho et al. 2009) to identify candidate unannotated transcripts in *D. melanogaster*. We then took the set of transcripts that had at least one BLAST hit to the *mel* reference sequence but no BLAST hits to *mel* gene annotations. We also used the Ensembl Metazoa BLAST search tool to verify that these candidate transcripts do not overlap with any annotated features (Howe et al. 2020). We filtered out very lowly expressed transcripts using counts from Salmon. We created a GTF file based on the BLAST coordinates of our candidate transcripts, and aligned our raw sequencing reads with STAR, performed feature counting with HTSeq, and removed ambient RNA using SoupX, as described earlier for transcriptome-wide analysis.

We used Ensembl Metazoa BLAST and the *mel* genome browser (Howe et al. 2020) to identify transcript coordinates, strand, and neighboring annotated genes. For cell-type-biased analysis of unannotated-transcript expression we added transcript counts to the broader *mel* dataset, post hoc. We used Seurat’s FindAllMarkers() method to identify cell-type expression bias, and significance was assessed using a Wilcoxon rank-sum test with Bonferroni multiple test correction. We assessed coding potential with CPAT v2.0.0 (Wang et al. 2013). To identify potential open-reading frames (ORFs), we used the getorf function in the EMBOSS software package (<http://emboss.sourceforge.net/apps/cvs/emboss/apps/getorf.html>). We attempted to characterize these potential ORFs further using Ensembl Metazoa Protein BLAST (Howe et al. 2020) to the database of all *mel* proteins, NCBI’s Conserved Domain Database search tool (Lu et al. 2020), and SignalP v5.0 (Almagro Armenteros et al. 2019) to identify putative signal sequences.

Results

Overview of single-nucleus RNA-Seq data

Following QC filtering to remove putative multiplets, we obtained a total of 4271 nuclei for single-cell analysis. The dataset comprised 1167 *mel*, 2116 *sim*, and 994 *yak* nuclei. While the overrepresentation of *sim* nuclei could be an artifact, given that tissue was pooled from nearly equal numbers of glands from each species prior to isolation of nuclei, it seems plausible that this difference results from divergence in cell number. Median counts per nucleus for *D. melanogaster*, *simulans*, and *yakuba* (hereafter referred to as *mel*, *sim*, and *yak*), were 1022, 1262.5, and 741.5, respectively, exhibiting the same species rank order as nuclei abundance, consistent with the idea of species differences in levels of seminal fluid production. We used k-nearest-neighbor-based clustering with UMAP visualization to identify three primary clusters of cells in both the *mel* and three-species dataset (Figure 1, A–C). We then used marker gene identification along with the relative sizes of clusters to assign cell-type identity to

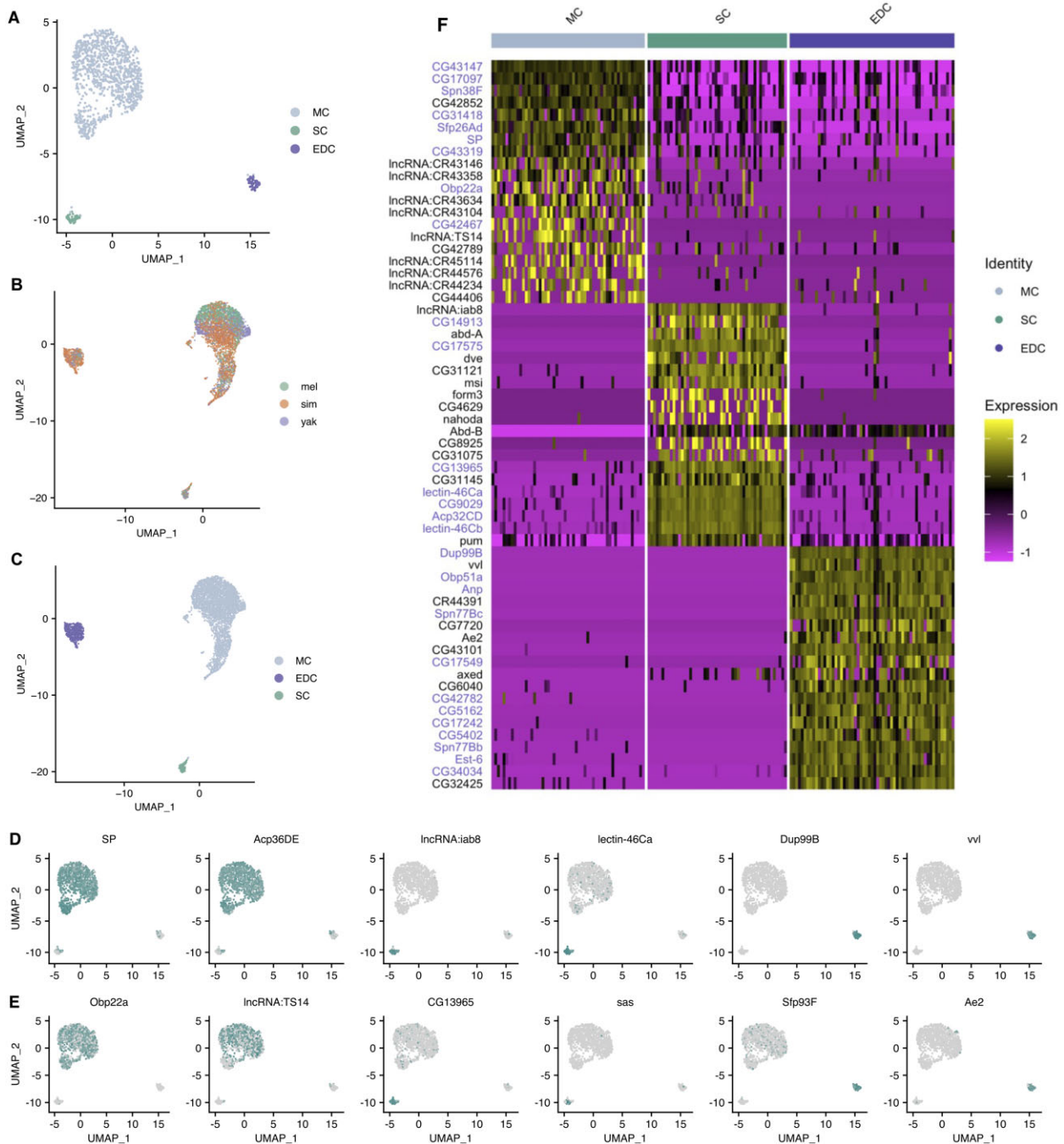


Figure 1 (A) UMAP showing clustering of *mel* single-nucleus transcriptomes into three major cell types: MC, SC, and EDC. (B) Nuclei from three species cluster concordantly; (C) into the same three major cell types. Differences between (A) and (B, C) are due to the nature of the UMAP algorithm (McInnes et al. 2018). Example marker genes in *mel*, with expression indicated in teal: (D) well-known markers and (E) novel markers. Cell-type clusters in (D) and (E) match those of (A). (F) Heatmap showing scaled expression of the top 20 markers of each cell type. Cell-type clusters in (D) and (E) are highlighted in blue text. Here we have down-sampled MC to 55 nuclei to aid visualization of SC and EDC, and so that scaled expression distributions are comparable among various marker genes. For the full population of MC, refer to [Supplementary Figure S1](#).

clusters, identifying MC, SC, and EDC. MC were identified as the cluster with the largest number of cells, and on the basis of markers Sex Peptide (SP) (Styger 1992), Acp36DE (Wolfner et al. 1997), and Acp95EF (DiBenedetto et al. 1990; Kalb et al. 1993) (Figure 1D). SC and EDC were classified as relatively smaller clusters. SC were identified by expression of lectin-46Ca (CG1652), lectin-46Cb (CG1656), *abd-A* (Maeda et al. 2018), and additionally by *iab-8* (Maeda et al. 2018) in the *mel*-only dataset (*iab-8* orthologs are not annotated in *sim* or *yak*) (Figure 1D). EDC were identified

by expression of *vvl* (Junell et al. 2010) and *Dup99B* (Rexhepaj et al. 2003) (Figure 1D). We additionally used *Abd-B* to characterize both SC and EDC (Gligorov et al. 2013; Maeda et al. 2018). In the *mel* dataset, we identified 1056 MC, 51 SC, and 60 EDC, with 6444, 2596, and 3445 expressed genes, respectively. In the three species dataset, we identified a total of 3629 MC, 139 SC, and 509 EDC, with 6978, 3573, and 5978 expressed orthologous genes, respectively. While our results revealed no evidence of subclusters within SC or EDC, we observed strong evidence of MC

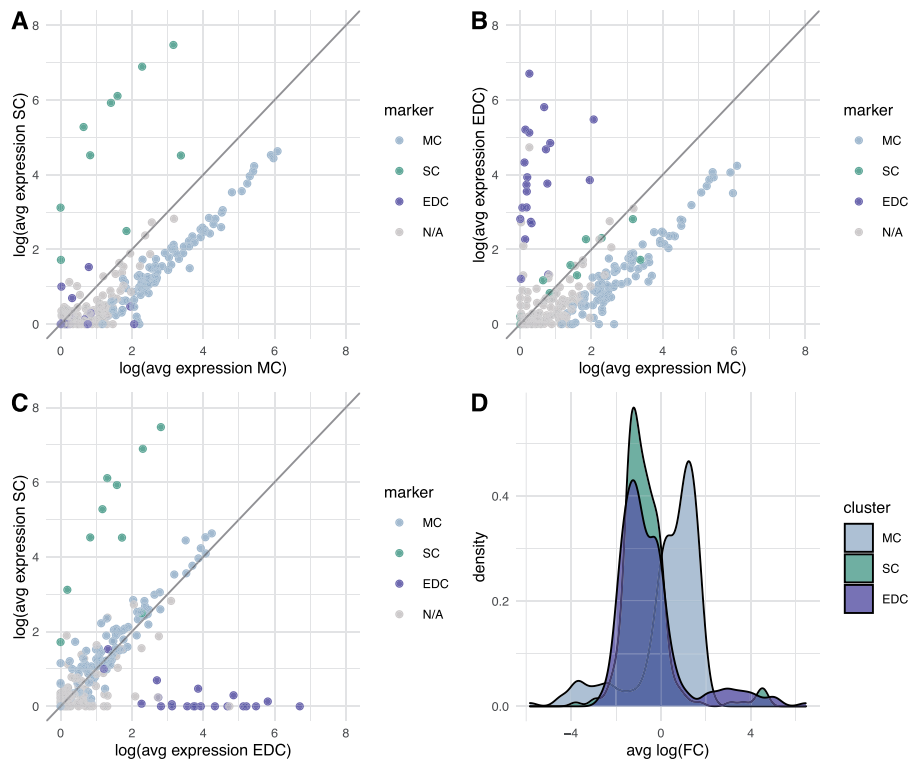


Figure 2 Expression of Sfps tends to be highly cell-type biased. (A–C) Expression levels of Sfps compared among cell types show a general pattern of MC enrichment and cell-type bias. Colors indicate marker gene status for each SFP; N/A indicates that a gene does not show a strong cell-type bias. (D) The average log(fold-change) of expression between each cell type and the other two shows that most SFPs are most highly expressed in MC, with few Sfps showing highest expression in SC or EDC.

subpopulations (see *Transcriptome heterogeneity among D. melanogaster MC subpopulations*). For downstream analyses, we merged these subclusters into a single MC cluster.

Using all annotated *mel* genes, marker genes for each *mel* cell type reveal both expected and novel markers, including Sfps and non-Sfps, and many lncRNAs (Supplementary Table S1 and Dataset S1; Figure 1, D and E). Details of some of the most notable marker genes specific to each cell type can be found in Supplementary Results.

Cell-type transcriptomes in the *Drosophila melanogaster* AG

Thresholding marker genes as expressed in at least 25% of cells in the focal cell type and minimum log₂ of the fold-change (log₂FC) = 0.25, we identified 540 *mel* marker genes (Figure 1F; Supplementary Dataset S1). Of these, 128 are annotated Sfps identified from proteomic studies of the male ejaculate (Findlay et al. 2009; Sepil et al. 2018). Of the 128 Sfp markers, 94 (73%) are MC markers, 10 (8%) are SC markers, and 24 (19%) are EDC markers, consistent with previous results that the majority of Sfps showing cell-type bias are expressed in MC (Kalb et al. 1993; Wolfner et al. 1997; Swanson et al. 2001). Marker Sfps for SC and EDC are summarized in Supplementary Table S1. Among the 214 total MC markers, 44% are Sfps. Among the 82 SC markers, only 12% are Sfps, and among the 262 EDC markers, 9% are Sfps. MC marker genes are significantly enriched for Sfps relative to both SC and EDC (pairwise G-tests, $P < 0.001$), while SC and EDC are not significantly different ($P = 0.43$). Thus, in contrast to MC, the distinct natures of SC and EDC transcriptomes are not driven primarily by Sfp expression. Tables of GO enrichment terms for cell-type markers can be found in Supplementary Dataset S10.

To investigate cell-type expression bias for all Sfps in addition to that of marker genes, we calculated for each of 264 *mel* Sfps the log₂(average expression) for the focal cell type and the average log₂FC vs all other cell types. Among the 224 Sfps detected in the data (Supplementary Dataset S2), 159 (71%) show greatest expression in MC, 25 (11%) show greatest expression in SC, and 40 (18%) show greatest expression in EDC. Expressed Sfps generally exhibit cell-type expression bias, with relatively few Sfps showing consistent expression among all three cell types (Figure 2). Highly MC-biased Sfps tend to also show expression in SC, though at a substantially lower level. Even among nonmarker Sfps we observe a trend toward greater MC expression than SC expression (Figure 2A, SC expression vs MC expression gives a slope = 0.73, $r^2 = 0.82$). EDC vs MC comparison for nonmarker Sfps exhibits a similar pattern (Figure 2B, slope = 0.79, $r^2 = 0.75$). Comparing SC vs EDC suggests a relatively more even spread of expression across these cell types, with some bias toward SC (Figure 2C; slope = 0.67, $r^2 = 0.598$). Among the 97 nonmarker Sfps, 66 show highest expression in MC, while 14 have highest expression in SC, and 17 have highest expression in EDC. Additionally, the distribution of average log₂FC of Sfps in MC vs all other cells skews significantly greater than SC vs all others and EDC vs all others, respectively (Figure 2D). The median log₂FC of MC vs all other cells is 0.75, while SC vs all others is -0.85, and EDC vs all others is -0.89.

We identified 24 Sfp EDC markers (Supplementary Table S1). Of these, one had previously been identified as EDC-enriched: *Dup99B*, *Obp51a*, *Spn77Bc*, *Spn77Bb*, *Est-6*, *Gld*, *Anp*, CG18258, CG5162, CG17242, CG5402, CG34034, and CG31704 (Cavener 1985; Samakovlis et al. 1991; Saudan et al. 2002; Takemori and Yamamoto 2009; Sepil et al. 2018). The remainder have not been previously identified as EDC-specific Sfps: *Treh*, *betaggt-I*, *Sfp93F* (Figure 1E), *trx*, *NT5E-2*, CG43101, CG33290, CG11590, CG17549, CG42782, and CG15394. CG42782 was previously identified as a likely mating plug protein gene, consistent

with origin in the ejaculatory duct or ejaculatory bulb (Avila et al. 2015). We also identified expected non-Sfps, *ventral veins lacking* (*vvf*) (Junell et al. 2010), and *Abd-B* (Gligorov et al. 2013). Novel EDC markers are *anion exchanger 2* (*Ae2*) (Figure 1E), *axundead* (*axed*), *single-minded* (*sim*), CG7720, CG43101, CG7342, CG13012, and CR44391. CR44391 is annotated as a pseudogene created by a tandem duplication of CG11400 (an EDC-biased gene), however, it has a homologous ORF with a strongly predicted signal sequence.

Transcriptome heterogeneity among *D. melanogaster* MC subpopulations

During initial analysis we discovered an apparent subcluster of MC characterized by unique SNN clusters at $k = 4$ and clear separation in UMAP space (Figure 3A). Of a total 1057 MC, 942 are in subcluster one (MCsp1) and 115 are in subcluster two (MCsp2). Three hundred and forty-nine significant markers (Bonferroni-corrected $P < 0.05$) distinguish these subclusters (Supplementary Dataset S3). In all three species, these subclusters are apparent and appear in roughly equal proportions (Supplementary Figure S2A and Dataset S4), strongly supporting the idea that they reflect a conserved, regulated phenomenon. Of the 349 markers distinguishing the MC subclusters, 34 are Sfps, all of which are MC markers and expressed in both subpopulations (Figure 3B). Twenty-six show higher expression in MCsp2, while just eight show higher expression in MCsp1 (Supplementary Dataset S3). Non-Sfps show the opposite pattern, with 102 genes showing increased expression in MCsp2, and 213 genes with higher expression in MCsp1 (Supplementary Dataset S3). The most enriched non-Sfp genes for each subpopulation are shown in Figure 3D. Genes significantly enriched in MCsp2 include 57 of the proteins comprising the large and small ribosomal subunits, along with *Eukaryotic Translation*

Elongation Factor 2 (*eEF2*), additional translation elongation factors *eEF5*, *eEF1 δ* , and *eEF1 α 1*, and translation initiation factors *eIF3a*, *eIF3b*, and *eIF3c*. Notably, MCsp1 has a lower level of RNA counts per nucleus than MCsp2, with 832 vs 1248 median counts (Figure 3C, Wilcoxon rank-sum test, $P < 0.001$). We find this same pattern of lower RNA counts in MCsp1 in *sim* and *yak* (Supplementary Figure S2B, Wilcoxon rank-sum tests, $P < 0.001$). Together with the quantitatively greater level of Sfp expression, these markers suggest a higher level of transcription accompanied by greater expression of translational machinery. Markers of MCsp1 include *Golgi microtubule-associated protein* (*Gmap*), *easily shocked* (*eas*), *taiman* (*tai*), and lncRNAs including *roX1*, *Hsro*, CR43104, CR43146, and CR45114 (Figure 3D). *roX1*, one of the strongest markers of MCsp1, plays a central role in dosage compensation (Mukherjee and Beermann 1965; Meller et al. 1997; Hallacli et al. 2012). We investigated patterns of broadly expressed genes using the methods of Mahadevaraju et al. (2021), but found no evidence of correlations between *roX1* abundance and X-to-autosome expression, or variation in X-to-autosome expression among subclusters or cell types. Thus, we find no evidence of differential dosage compensation between MC subpopulations.

We also used a pseudotime approach to model MCsp1 and MCsp2 as a continuous trajectory of differentiating cells. We found evidence of a continuous distribution of MC over pseudotime, strongly concordant with transcriptomic differences between MCsp1 and MCsp2, suggesting a range of expression within the entire population of MC (Supplementary Figure S3, A and B). These results are consistent with a dynamic process between MCsp1 and MCsp2, which could be explained by temporal or spatial factors. Visualizing dynamic differential gene expression with tradeSeq, we find a limited population of

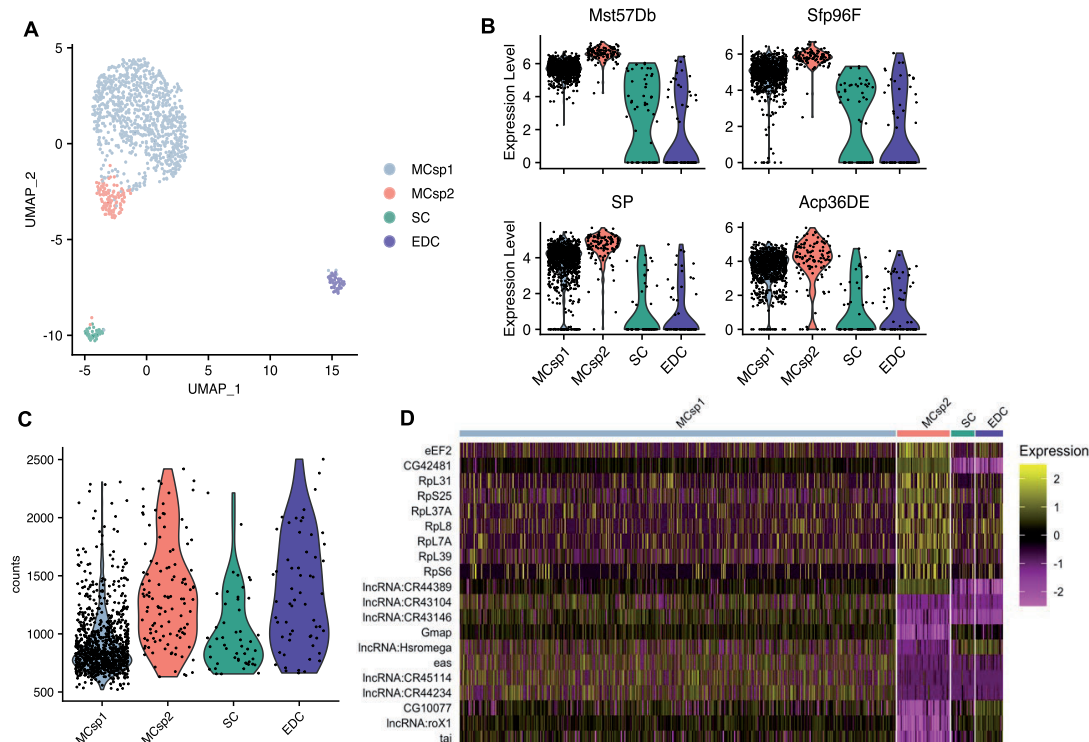


Figure 3 Transcriptome heterogeneity among subpopulations of MC in *mel*. (A) Subpopulations of MC are apparent in both UMAP space and SNN clustering with $k = 4$. (B) Examples of MC marker Sfps with greater expression in MCsp2. (C) MCsp1 has a significantly lower level of RNA counts per cell than MCsp2 or EDC (Kruskal–Wallis test and Wilcoxon rank-sum tests, $P < 0.001$), but not SC (Wilcoxon rank-sum test, $P > 0.05$). There is no significant difference between MCsp2 and EDC (Wilcoxon rank-sum test, $P > 0.05$). (D) Heatmap showing scaled expression of the top 10 non-Sfp markers for each subpopulation, suggesting enrichment of translational machinery in MCsp2.

intermediate phase cells, but no obvious evidence of pseudo-temporal variance in the onset of differential gene expression, pointing to a relatively simple process (Supplementary Figure S3B). Finally, we observe evidence of finer functional divisions within MC in an apparent third subpopulation (Supplementary Figure S2A and Dataset S5) that deserves further investigation. Unlike MCsp2, MCsp3 does not show significant differences from MCsp1 in Sfp expression. Some of the top genes characterizing MCsp3 include *Idgf4*, *Wnt6*, *pain*, *luna*, *CG18067*, and *CG9336*. However, given that this subpopulation is less well-supported than MCsp2, we do not wish to speculate about it here.

Cell-type-specific differential gene expression across species

We used our integrated three-species dataset to characterize differential gene expression (DE) across species. UMAP visualization reveals strongly concordant clustering of cell types across species

(Figure 1, B and C). The top 12 DE genes for each cell type are summarized in Supplementary Table S2, and expression of DE genes in all cell types can be found in Supplementary Dataset S9. We found 132 genes that are DE ($\log_{2}FC > 1$) in at least one pairwise species contrast among MC (Supplementary Dataset S6), of which 40 (30%) are Sfps. Among SC we found 106 DE genes (Supplementary Dataset S7), of which 21 (20%) are Sfps, while in EDC we found 221 (Supplementary Dataset S8), of which just 32 (14%) are Sfps. The percentage of expressed genes that are DE for each species contrast and cell type (Figure 4A) is significantly heterogeneous (G-test, $P < 0.001$, Supplementary Table S3). Notably, EDC show a consistently greater fraction of DE genes than MC and SC for each species comparison, except for *sim-yak* EDC vs SC. The fraction of DE genes does not differ between MC and SC for any species contrasts. The fraction of DE genes in different cell types tends not to vary significantly over species contrasts, except for EDC, where the *mel-yak* fraction is significantly greater than *mel-sim*, but not significantly different from *sim-yak*. To

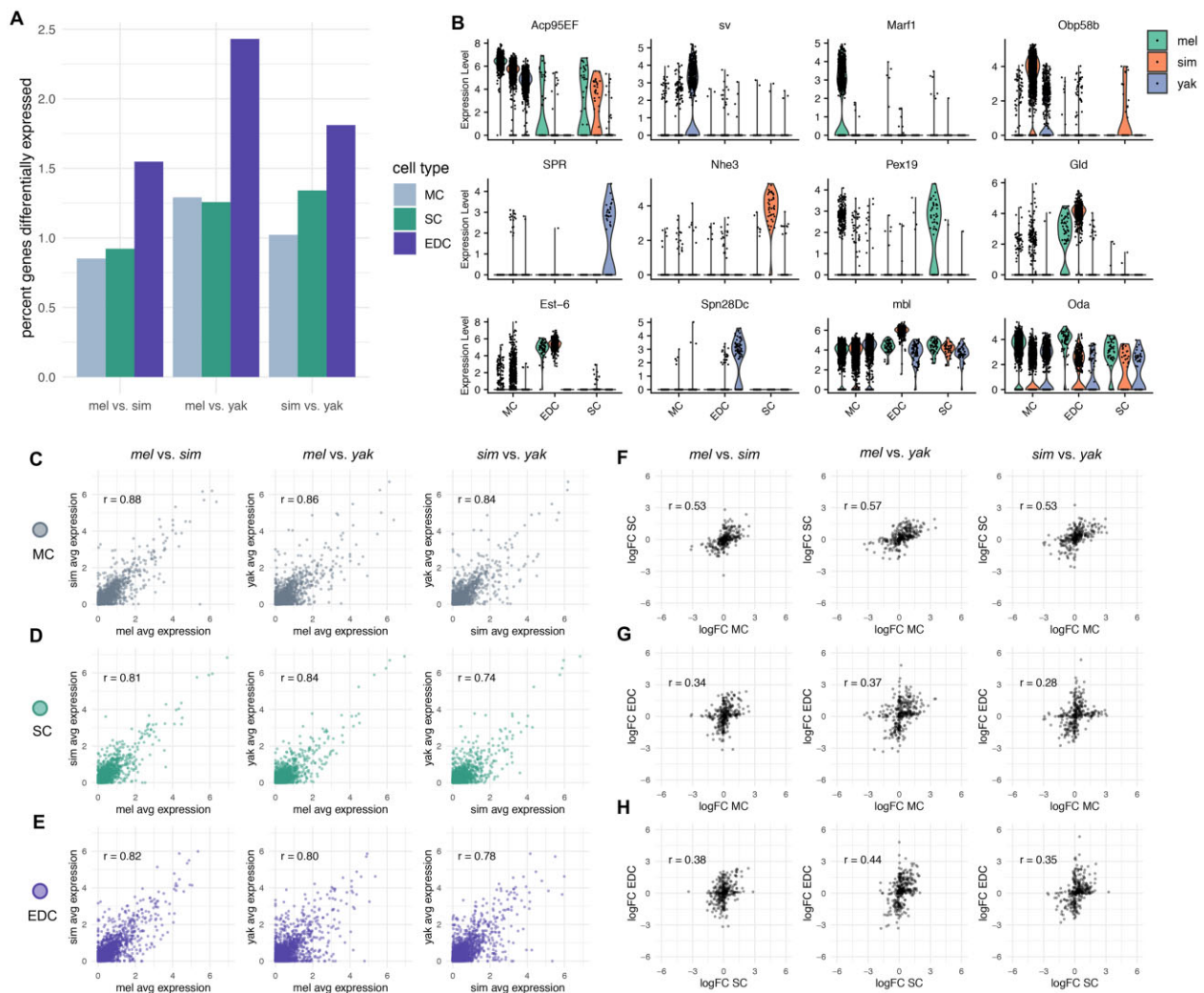


Figure 4 (A) Percentage of expressed genes DE by cell type and species contrast (G-test, $P < 0.001$). For significance values of pairwise tests (see Supplementary Table S2). (B) Examples of differential expression detected in this study. (C–E) Pearson's correlations of transcriptome-wide expression show cell-type- and species-specific patterns of divergence. The level of divergence among species is summarized by r . (C) MC, (D) SC, (E) EDC; columns indicate each of three species contrasts. Note the greater correlations among MC contrasts relative to SC and EDC, and lower correlations among *sim-yak* relative to other species contrasts. (F–H) Pearson's correlations of $\log_{2}FC$ of DE genes among contrasts reveal differences in the level of concerted vs independent DE among cell-type- and species-contrasts. The level of concerted DE among species is summarized by r . (F) MC vs SC, (G) MC vs EDC, (H) SC vs EDC. Columns indicate each of three species-contrasts. Note the overall greater level of concerted DE among MC and SC relative to the other cell-type contrasts.

determine the magnitude of DE among the genes that most distinguish each cell type we asked how many marker genes were DE in each cell type. In MC, 73 of 309 markers (24%) are DE, in SC, 25 of 121 markers (21%) are DE, and in EDC, 123 of 255 markers are DE (33%). EDC markers are significantly more likely to be DE than MC or SC (pairwise Fisher's exact tests, $P < 0.001$), while MC and SC are not significantly different ($P = 0.7$). Together, the data suggest an elevated level of DE for EDC relative to MC and SC, and an effect of lineage on DE in EDC; the *mel-yak* EDC contrast has significantly more DE genes than *sim-yak*, suggesting that DE genes accumulated faster in the *mel* EDC than the *sim* EDC. These conclusions are robust to different logFC cutoffs (Fig S4A-D). There is a trend toward elevated MC enrichment compared with SC at particularly high and low cutoffs, however, these differences are not statistically significant (Wilcoxon rank-sum tests, $P > 0.05$). We found no evidence of differences in the magnitude of DE across cell types and lineages; distributions of logFC among DE genes are not significantly different (Supplementary Figure S4, E and F).

We used Pearson's correlations of expression among all genes in species contrasts to investigate overall levels of transcriptome-wide divergence. A lower correlation coefficient (r) suggests a greater level of divergence. MC have the greatest overall correlations (Figure 4B; $r_{MC_{mel-sim}} = 0.88$, $r_{MC_{mel-yak}} = 0.86$, and $r_{MC_{sim-yak}} = 0.84$). Pearson's correlations for SC and EDC are lower overall (Figure 4, D and E; $r_{SC_{mel-sim}} = 0.81$, $r_{SC_{mel-yak}} = 0.84$, $r_{SC_{sim-yak}} = 0.74$; $r_{EDC_{mel-sim}} = 0.82$, $r_{EDC_{mel-yak}} = 0.80$, and $r_{EDC_{sim-yak}} = 0.78$). The data suggest an overall slower rate of expression evolution in MC than SC and EDC. Furthermore, the heterogeneous correlations for SC and EDC across species pairs suggest lineage by cell-type interactions on rates of transcriptome evolution.

DE genes are summarized in Supplementary Datasets S6–S9, but below we wish to highlight a few interesting examples. The Sfp *Acp95EF* is strongly DE in MC, which has highest expression in *mel*, lower expression in *sim*, and lowest expression in *yak* (Figure 4B). The transcription factor *shaven* (*sv*) is lowly expressed in *mel* and *sim*, but much more highly expressed in *yak* MC. Meiosis regulator and mRNA stability factor 1 (*Marf1*) has near-zero expression in *sim* and *yak*, but high expression and MC bias in *mel* (Figure 4B), supporting our previous work using bulk-tissue RNA-Seq characterizing this pattern of gain-of-expression specific to the *mel* AG (Cridland et al. 2020). Odorant-binding protein 58b (*Obp58b*) is highly expressed in *sim*, expressed moderately in *yak*, and rather lowly expressed in *mel* (Figure 4B). Findlay et al. (2009) detected peptides corresponding to *Obp58b* in a proteomic screen of *sim* seminal fluid but did not detect any corresponding peptides in *mel* or *yak* seminal fluid. Taken together, these results suggest *Obp58b* is an MC-expressed Sfp in *sim* but does not have a role as an Sfp in *mel*. The status of *Obp58b* in *yak* is less clear.

Sex Peptide Receptor (SPR), which is responsible for interactions with Sfp SP in the female reproductive tract (Yapici et al. 2008), is expressed in *yak* SC, but not in *sim* or *mel*, or in MC or EDC (Figure 4B). SPR is known to have additional ligands, and is expressed in the CNS of both males and females, but not in the *melanogaster* male reproductive tract (Kim et al. 2010; Poels et al. 2010), so potential functions of SPR in *yak* SC and whether it interacts with endogenous SP are interesting questions. Further examples of DE genes among SC include *Na⁺/H⁺ hydrogen exchanger 3* (*Nhe3*), with high expression in *sim* and near-zero expression in *mel* and *yak* (Figure 4B), consistent with *sim* gain-of-expression, and Peroxin 19 (*Pex19*), which exhibits what is likely gain-of-expression in *mel* SC and near-zero expression in *sim* and *yak*

(Figure 4B). In general, we observed little DE among SC-biased Sfps. While 24 Sfps exhibit SC DE, 22 of these are MC markers, with significantly lower expression in SC than MC. Two exceptions are *midline fascilin* (*mfas*) and CG3349 (Supplementary Dataset S7).

The EDC marker gene *Esterase 6* (*Est-6*) is highly expressed in *mel* and *sim*, and much more lowly expressed in *yak* (Figure 4B). *Est-6* transcript and *Est-6* protein expression in the ejaculatory duct is specific to *mel*, *sim*, and *D. sechellia*, and notably absent in the rest of the *melanogaster* subgroup, including *yak* (Richmond et al. 1990). *Serpin 28Dc* (*Spn28Dc*) has *yak*-specific EDC expression, with no expression in other cell types or species (Figure 4B). Serpins are a common component of seminal fluid (reviewed in Laflamme and Wolfner 2013), making *Spn28Dc* a good candidate for a *yak*-specific Sfp. *Glucose dehydrogenase* (*Gld*) has a high level of expression in *sim*, a lower level in *mel*, and near-zero expression in *yak* (Figure 4B). This same species-specific pattern was previously observed in enzymatic GLD assays (Cavener 1985), suggesting that variation in GLD abundance in the ejaculatory duct is ultimately controlled at the transcriptional level.

To determine the ratio of markers to nonmarkers among DE genes, we used singlet markers (characterizing just one cell type) called independently for each species to filter our list of DE genes, thereby allowing markers to be unique to one species or shared. We found 61% of DE genes were markers specific to a particular cell type. However, we find large differences in this ratio among cell types; 73% of genes DE in MC are MC markers, 75% of genes DE in EDC are EDC markers, while just 19% of DE genes in SC are SC markers. Thus, much DE is associated with cell-type biased expression for MC and EDC but not for SC. For example, *muscle-blind* (*mbi*) exhibits high EDC expression in *sim* relative to both *mel* and *yak*, while showing no DE in MC or SC, despite high expression in these cell types (Figure 4B). Alternatively, DE may be correlated in the same direction across multiple cell types. For example, *Ornithine decarboxylase antizyme* (*Oda*) is broadly expressed and shows the same pattern of increased *mel* expression in each cell type (Figure 4B). We also identified nine cases where genes have shifted in their marker status among species (Supplementary Figure S5). For example, *Sfp24C1*, the only Sfp in this gene set, is modestly expressed in *mel* MC, strongly EDC-biased in *sim*, and expressed in few *yak* cells. This rapid expression evolution is mirrored in its coding sequence, with high levels of adaptive amino acid substitutions between *mel* and *sim* ($\alpha = 0.75$, $dN/dS = 5$). *Glucuronyltransferase P* (*GlcAT-P*) shows a striking pattern of MC-biased expression in *mel*, with weaker MC expression in *sim* and *yak*, and very strong expression in *yak* EDC specifically (Supplementary Figure S5). *GlcAT-P* is expressed in the female spermatheca where it is thought to be involved in sperm maturation and/or preservation (Allen and Spradling 2008), but potential functions in the AG are unexplored.

To investigate the degree of concerted vs independent expression evolution across cell types we calculated pairwise Pearson's correlation coefficients (r) of logFC of DE genes for each cell type for each of the three species contrasts. A greater value of r suggests a greater overall level of concerted evolution, where expression evolution is more similar among different cell types. Conversely, a lower r would suggest relatively more independent expression evolution across cell types. We find r ranges between 0.28 and 0.57 for each comparison (Figure 4, F–H). MC and SC have the highest correlations (Figure 4F); $r_{mel-sim} = 0.53$, $r_{mel-yak} = 0.57$, and $r_{sim-yak} = 0.53$. SC and EDC are less correlated (Figure 4G); $r_{mel-sim} = 0.38$, $r_{mel-yak} = 0.44$, and $r_{sim-yak} = 0.35$. MC

and EDC have the lowest correlations (Figure 4H): $r_{mel-sim} = 0.34$, $r_{mel-yak} = 0.37$, and $r_{sim-yak} = 0.28$. To determine the expected distribution of r under a null model of cell-type-independent evolution, we permuted logFC 10,000 times and calculated values of r , as before. The 99th percentile of permuted r (0.123–0.133) was much lower than each observed r , supporting the hypothesis of correlated transcriptome divergence across cell types. Nevertheless, a gene is unlikely to pass our logFC ≥ 1 threshold for DE in multiple cell types; of 362 DE genes, 282 (78%) appear in a single-cell type, 51 (14%) appear in two, and just 25 (7%) appear in all three cell types. This pattern is reflected in plots of logFC across cell types, with relatively few points falling near the line $x = y$ (Figure 4, F–H). Thus, while the overall directionality of DE is similar among cell types, the largest interspecific expression differences tend to be limited to one cell-type.

Protein sequence evolution in *melanogaster*

To investigate the evidence for protein adaptation among marker genes of each cell type, we used the McDonald–Kreitman test estimator α (McDonald and Kreitman 1991; Smith and Eyre-Walker 2002). A positive value of α suggests a history of directional selection. Among positive values, α provides an estimate of the proportion of amino acid differences between *mel* and *sim* attributable to directional selection. We obtained estimates of α for 561 of 691 marker genes (called from joint analysis of *mel*, *sim*, and *yak*), of which 265 (47%) were positive. The proportion of MC markers with positive α (61%) is significantly greater than SC (41%) or EDC (40%) (Figure 5A; pairwise Fisher’s exact tests, $P = 0.002$, $P < 0.001$,

respectively), suggesting that compared with SC and EDC, MC markers are more likely to have a history of adaptive protein divergence. Median values among positive α for SC, MC, and EDC are 0.30, 0.57, and 0.52, respectively (Kruskal–Wallis test, $P = 0.01$), with SC being significantly smaller than MC and EDC (pairwise Wilcoxon rank-sum tests, $P = 0.008$). Overall, it appears MC-biased genes exhibit the greatest adaptive protein divergence and SC-biased genes the least. Given the enrichment for Sfp expression in MC we wanted to investigate whether this pattern of MC protein adaptation is driven by Sfp variation or is a general property of this cell type. Among marker genes, 91 of 126 Sfps (72%) have positive α values, while 174 of 432 non-Sfps (40%) have positive α values, a significant enrichment among Sfps (Figure 5B; Fisher’s exact test, $P < 0.001$). However, medians of positive α values are not significantly different for Sfps vs non-Sfps (Kruskal–Wallis test, $P = 0.11$). Among non-Sfp markers there is no significant difference in the proportion of positive vs negative α among cell types (Figure 5C; Fisher’s exact test, $P = 0.40$). However, non-Sfps show significant differences in distributions of positive α , with median α of 0.24 in SC, 0.54 in MC, and 0.50 in EDC (Kruskal–Wallis test, $P = 0.001$). Both MC and EDC are significantly greater than SC (pairwise Wilcoxon rank-sum tests, $P = 0.004$ and $P = 0.02$, respectively). Thus, while the unequal distribution of Sfps among marker genes in different cell types accounts for some of the observed cell-type heterogeneity in the proportion of markers showing excess protein divergence, the reduced effect of directional selection on protein divergence in SC-biased genes remains apparent as a general phenomenon.

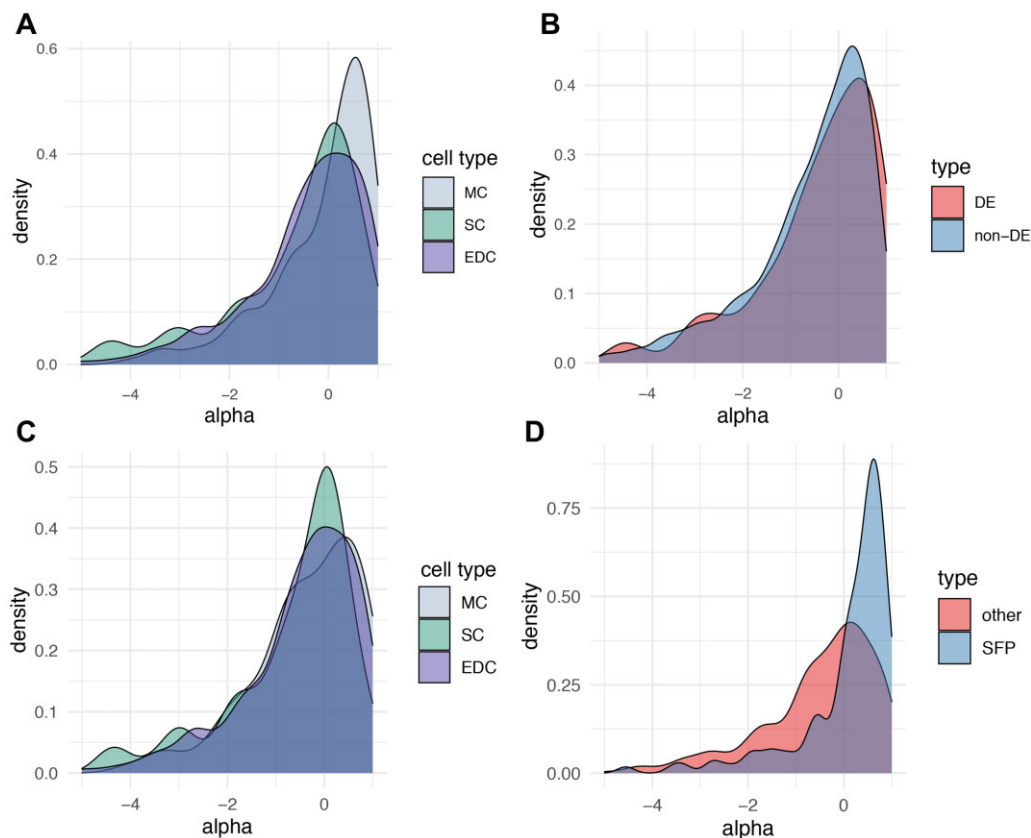


Figure 5 Distributions of α (*mel* population data vs *sim*) for marker genes. (A) α values by cell type show that MC markers are significantly greater than SC or EDC (Kruskal–Wallis test, $P = 0.001$). (B) Sfp markers have a dramatically greater median α than non-Sfps (Fisher’s exact test, $P < 0.001$). (C) Removing Sfps from the data shifts the distribution of MC α lower. MC and EDC are no longer significantly different, but SC is significantly less than MC and EDC (Kruskal–Wallis test, $P = 0.001$). (D) Genes that are DE between *mel* and *sim* have a modest but significantly greater α than non-DE markers (Kruskal–Wallis test, $P < 0.001$).

To investigate whether genes that are DE between *mel* and *sim* are also enriched for adaptive protein divergence for the *mel-sim* species pair, we compared α for genes that were DE vs non-DE. While the proportion of DE vs non-DE genes exhibiting $\alpha > 0$ (42.5% and 38.7%, respectively) were not significantly different (Figure 5D; G-test, $P=0.37$), the median positive α value for DE genes, 0.59, was significantly greater than median positive α for non-DE genes 0.46 (Kruskal-Wallis test, $P < 0.001$). Thus, expression divergence appears to be more strongly correlated with the proportion of protein divergence explained by selection than with the probability of a protein having elevated levels of fixed nonsynonymous substitutions.

Finally, we investigated some individual AG-expressed genes with unusually high values of α . While adaptive protein divergence in Sfps has been studied extensively (Tsaour et al. 1998; Begun et al. 2000; Swanson et al. 2001; Kern et al. 2004; Holloway and Begun 2004; Mueller et al. 2005; Begun and Lindfors 2005; Wagstaff and Begun 2005; Schully and Hellberg 2006; Wong et al. 2008; but see also Dapper and Wade 2020; Patlar et al. 2021), there has been no targeted study of adaptive protein evolution of non-Sfp genes exhibiting strongly AG-biased expression. Nonsecreted genes with evidence of rapid divergence might play important roles in the regulation of the seminal fluid at the level of transcription, post-translational modification, secretory pathway control, or other points in the production of the ejaculate. We report protein coding non-Sfps with extreme AG expression bias and high values of α in Supplementary Table S4. Most of these genes are uncharacterized, apart from Carbonic anhydrase 16 (CAH16). An alternative possibility is that these genes are unannotated Sfps, however, it seems unlikely that they would have escaped proteomic screening (Findlay et al. 2009; Sepil et al. 2018; Wigby et al. 2020) given their relatively high expression in the AG.

Identification of unannotated genes expressed in the AG

Following stringent filtering (see Supplementary Methods), we identified 11 unannotated, single-exon genes (Table 1; Supplementary Table S4; github.com/alexmajane/AG_single_nucleus). Transcript assemblies of FlyAtlas2 data were used to improve our annotation for seven of these candidates. Since DN100097 and DN2695 are SC-limited in expression (Supplementary Figure S6A), we used RNA-Seq data from FACS-sorted SC (Immarigeon et al. 2021) to further improve our annotations. The median transcript length is 630bp (range = 352–3102bp). None of these genes overlap annotated features in the *mel* genome. Among these genes, four show strong MC bias, two are SC-biased, and two are EDC-biased. In general, these candidates are expressed at a relatively high level compared with expressed annotated genes, but a relatively low level compared with marker genes (Supplementary Figure S6B). The two notable exceptions to this trend are DN2695 in SC, and DN818 in MC, which are expressed at a more intermediate level among markers. These two candidates additionally pass more stringent criteria (expressed in $\geq 25\%$ of focal cells) to be considered marker genes (Supplementary Dataset S1). DN2695, the seventh most significant SC marker, is expressed in 47% of SC yet shows no evidence of MC or EDC expression. Interestingly, the two candidate SC-biased genes, DN2695 and DN10097, lie 5.4 kb apart within a 20.1-kb intergenic region on chromosome 2L. Both EDC-biased candidates, DN16089 and DN10930, are exclusively detected in EDC, although they do not meet our criteria for marker genes. DN16089 is expressed in 18% of EDC, DN10930 is expressed in 15% of EDC; neither exhibit SC or MC expression. DN16089 is located just 79bp from the EDC marker *sim*, but on the opposite strand. All 11

Table 1 Unannotated candidate genes expressed in the *D. melanogaster* accessory gland

Transcript	Chromosome	Length	Expression bias	logFC	P
DN4707	3R	352	Broad	0.544	0.309
DN8354	2R	530	Broad	0.255	1
DN35169	3R	630	Broad/MC	0.595	0.087
DN10930	3R	863	EDC	0.750	<0.001
DN16089	3R	572	EDC	0.718	<0.001
DN11110	X	352	MC	0.923	0.001
DN2736	2L	739	MC	0.856	0.006
DN5813	2R	1278	MC	0.981	<0.001
DN818	3R	3102	MC	1.170	<0.001
DN10097	2L	353	SC	0.826	<0.001
DN2695	2L	2176	SC	2.130	<0.001

Length refers to the span of BLAST coordinates. logFC is the cell type with highest fraction of expression compared with the other two cell types. P is the result of a Wilcoxon rank-sum test with Bonferroni correction (see Supplementary Table S4 for additional details).

transcripts are predicted to be noncoding by CPAT. Although getorf identified many putative ORFs (github.com/alexmajane/AG_single_nucleus), BLAST comparisons of predicted proteins to the *D. melanogaster* protein database and the NCBI database of conserved domains returned no significant matches. SignalP revealed no evidence of signal sequences.

Discussion

Our single-nucleus transcriptome analysis of the primary *Drosophila* seminal fluid producing organs has validated conjectures in the literature and revealed several new findings. As expected, MC are the primary source of Sfp diversity and exhibit transcriptomes biased toward Sfp production. While several individual Sfps are produced in all three major cell types investigated here, it is notable that the majority of Sfps exhibit strong cell-biased expression, raising the question of why this occurs. Given that these three cell types are spatially separated along the reproductive tract, with the SC distal, the EDC proximal, and the MC intermediate, perhaps there are Sfp “order effects” in assembling the seminal fluid prior to transfer to the female. Order effects have been observed in assembly of the spermatophore in *Pieris rapae* butterflies (Meslin et al. 2017) and seminal fluid in tsetse flies (Odhiambo et al. 1983). Such order effects could influence the details of how Sfps bind sperm or interact directly with the female reproductive tract. In spite of the important role for MC in Sfp production, many genes showing MC bias are not annotated as Sfps; their roles in AG function remain to be investigated. SC and EDC transcriptomes are much less biased toward Sfp expression. Indeed, most SC and EDC markers are not Sfps, and most of the genes exhibiting strongly biased expression in these cell types have no known functions in male reproduction. Thus, much of the biology of the AG and ejaculatory duct is still mysterious. Especially notable is the relatively small number of Sfps produced in SC, as first reported by Immarigeon et al. (2021).

Our data confirm that expression of the “Sex Peptide network”—Sfps that interact with SP in the female reproductive tract and enhance the PMR (Ravi Ram and Wolfner 2007, 2009; LaFlamme et al. 2012; Findlay et al. 2014; Singh et al. 2018; McGeary and Findlay 2020)—is divided across cell types. lectin-46Ca, lectin-46Cb, and CG17575 are SC markers, while SP, *aqrs*, *antr*, *intr*, CG9997, and *Sems* are MC markers, and *Esp* appears EDC biased. *frma* and *hdly*, remaining members of the known Sex Peptide network, are not strongly expressed in our dataset. Discovery of the EDC marker Anion exchanger 2 (Ae2), provides a

clue about possible functions of the ejaculatory duct apart from Sfp production. In *D. melanogaster*, Ae2 regulates intracellular pH through $\text{Cl}^-/\text{HCO}_3^-$ exchange in the midgut (Overend et al. 2016) and ovary (Ulmschneider et al. 2016; Benitez et al. 2019). Ae2 is a highly conserved membrane protein, responsible for pH regulation in the mouse epididymal epithelium, seminiferous tubules, and developing spermatocytes, and is essential for spermatogenesis (Medina et al. 2003). Thus, EDC-biased expression of Ae2 suggests that the ejaculatory duct may regulate ejaculate pH.

Many of our strongest marker genes are lncRNAs, including markers of our newly defined MC subpopulations. Aside from *iab-8* and *msa* (Maeda et al. 2018), the roles of lncRNAs in AG biology and male reproduction more broadly are uncharacterized, though the possibility that some of these RNAs code for small proteins cannot be ruled out (Cridland et al. 2021; Immariageon et al. 2021). Our analysis revealed strong evidence of transcriptionally distinct MC subclusters. The most obvious distinction between them is that one exhibits evidence of higher transcriptional and translational activity. Many of the markers for these MC subclusters are annotated as lncRNAs, further supporting the possible importance of noncoding RNAs in AG biology. Given that we observe no correlation between *roX1* expression and dosage compensation, *roX1* might have other, uncharacterized functions in the AG. Whether MC subpopulations represent cell subtypes, transitory states, or developmental states, and whether communication among these subclusters occurs, are important questions. To compare our MC subcluster inference with a similar inference made in the Fly Cell Atlas preprint (Li et al. 2021), we investigated some of our top marker genes (Supplementary Figure S7) and found concordant patterns of expression in their data, consistent with the same subpopulations identified in the two experiments.

We found evidence for 11 unannotated *D. melanogaster* genes expressed in seminal fluid producing tissues, most of which are strongly cell-type biased. Given the low coding potential of these transcripts, and that predicted ORFs exhibit no homology to known proteins and show no evidence of signal sequences required for secretion, their possible functions are mysterious, yet likely relevant to the biology of these three cell types. The two SC-biased genes DN2695 and DN10097, located proximal to one another in a large intergenic region, are particularly interesting candidates for future research into their role in SC biology.

The transcriptomes of the three major cell types investigated here show many similarities between species, as expected given their recent common ancestor. Moreover, interspecific transcriptome divergence among cell types is not occurring independently, supporting the notion that these cell types have correlated functions. Nevertheless, each cell-type exhibits a distinct transcriptome and has distinct evolutionary properties. MC and SC, the two cell types of the AG proper, have less transcriptional divergence from each other than either has from EDC, consistent with more functional and developmental overlap between MC and SC. Overall, interspecific transcriptome divergence is substantially slower for MC than for SC or EDC. However, divergence rates are heterogeneous among lineages. For example, SC transcriptome divergence is substantially greater in the *sim* vs *yak* comparison than the *mel* vs *yak* comparison, consistent with the hypothesis of accelerated transcriptome evolution along the *sim* lineage for this cell type.

A slightly different picture emerges if one focuses on the most strongly DE genes between species rather than on overall transcriptome divergence. While the directionality of DE is similar

among cell types, the largest expression changes tend to be exhibited in a single-cell type, suggesting that the mechanisms driving divergence operate heterogeneously across cell types. EDC generally show the greatest interspecific divergence, though again, the data are consistent with the hypothesis of lineage differences in evolutionary rates. Whether the greater proportion of DE genes among EDC results from directional selection or relaxed stabilizing selection (Dapper and Wade 2020) is an open question. Many DE genes are Sfps, as expected since Sfps are a major component of these transcriptomes, but notably, most DE genes are not Sfps, raising important questions about the functional axes along which species differences are evolving in these cell types. Indeed, many of the most strongly differentiated genes, which include genes expressed at a high level in some species and apparently unexpressed in others, have unknown functions in these cells in any of the three species. Consistent with transcriptome-wide results, correlations of logFC for DE genes among cell types suggest concerted change, as expected given the closely shared developmental origins of these cell types (Musser and Wagner 2015; Liang et al. 2018) and short-time scales examined in this study. Indeed, correlations of logFC are greatest between MC and SC, which differentiate later in development (Xue and Noll 2000; Minami et al. 2012; Gligorov et al. 2013), compared with EDC cells. Given the limited inquiry into the phenomenon of DE across related cell types in *Drosophila*, however, it is difficult to establish a baseline expectation of concerted change. Finally, we identified a small set of genes that have shifted their marker gene status to different cell types among species. These appear to be relatively rare evolutionary events, at least on the time scales examined here, but the regulatory basis and functional significance of these shifts remain to be determined.

Our investigation of the interaction of protein divergence with cell-biased expression and interspecific expression divergence revealed a few salient patterns. As expected, given genome wide results (Begun et al. 2007; Langley et al. 2012), directional selection appears to play an important role in driving protein evolution for cell-biased genes. Indeed, α values for marker genes, though high, are not obviously different from genome-wide estimates (Fraisse et al. 2019), raising interesting questions about whether protein divergence of the AG is unusual in any way. Nevertheless, the relative importance of adaptive divergence appears to vary across cell types. MC-biased genes are more likely than SC- or EDC-biased genes to show evidence of directional selection. Much of this enrichment results from the strongly Sfp-biased expression of MC, and cell-biased genes that are not Sfps are equally likely to show evidence of protein adaptation for all three cell types. However, conditioning on positive α , the relative importance of directional selection is much lower for SC-biased genes than for MC- or EDC-biased genes. Overall, it seems that while adaptive protein evolution is likely common for all cell types, it is most pronounced for MC and least for SC. A speculative hypothesis for this observation is that more beneficial nonsynonymous mutations are associated with phenotypes related to establishment of the female PMR, which is primarily a MC function, than with long-term maintenance of receptivity to remating, which is in part an SC function (Sitnik et al. 2016). However, it is difficult to make strong statements about the agents of selection driving protein divergence in marker genes without more information on their biological functions in the AG or other tissues and cell types. Finally, we found DE genes are not more likely than other genes to show evidence of protein adaptation, however, there is a small, significant elevation of positive α for DE genes vs non-DE genes. Thus, while there appear to be some

correlations between expression divergence and protein adaptation, the relationship is neither particularly strong nor simple.

While our analyses of single-nucleus transcriptomes in an evolutionary genetics framework has led to many functional and evolutionary findings and hypotheses, perhaps what is most apparent is how little we still understand the biology and evolution of these cells. Many open questions remain about the regulation and function of the seminal fluid producing cells, the biological consequences of species divergence in these cells, and the evolutionary mechanisms shaping this divergence. Continued investigation of closely related species for single-cell phenotypes and population genetic variation will facilitate the fruitful investigation of both functional and evolutionary mechanisms, and help to draw additional connections between these two research domains.

Data availability

Count data for single nuclei in each of the three species, fasta and GTF files for unannotated genes, R scripts, our orthology table, and the list of Sfps used in this study are available at github.com/alexmajane/AG_single_nucleus. Sequence data are available at the NCBI SRA under accession number PRJNA741528.

Supplementary material is available at GENETICS online.

Acknowledgments

We thank the UC Davis Flow Cytometry Shared Resource for performing FACS and the UC Davis Genome Center for single-nucleus library preparation and sequencing. We thank Ben Hopkins and Rachel Thayer for feedback and thoughtful discussions, the anonymous reviewers, and the Chiu, Lott, and Kopp labs for sharing materials and equipment.

Funding

This work was supported by the National Institutes of Health, grant number R35 GM134930 to D.J.B., a National Science Foundation Graduate Research Fellowship to A.C.M., and a Pilot and Feasibility Program award from the UC Davis Research Core Facilities Program.

Conflicts of interest

The authors declare that there is no conflict of interest.

Literature cited

- Aguadé M. 1999. Positive selection drives the evolution of the Acp29AB accessory gland protein in *Drosophila*. *Genetics*. 152: 543–551.
- Allen AK, Spradling AC. 2008. The Sf1-related nuclear hormone receptor Hr39 regulates *Drosophila* female reproductive tract development and function. *Development*. 135:311–321. doi: 10.1242/dev.015156.
- Almagro Armenteros JJ, Tsirigos, KD Sønderby, CK Petersen, TN Winther, O, et al. 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat Biotechnol*. 37:420–423.
- Anders S, Pyl PT, Huber W. 2015. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics*. 31: 166–169.
- Arendt D, Musser JM, Baker CVH, Bergman A, Cepko C, et al. 2016. The origin and evolution of cell types. *Nat Rev Genet*. 17:744–757.
- Assis R, Zhou Q, Bachtrog D. 2012. Sex-biased transcriptome evolution in *Drosophila*. *Genome Biol Evol*. 4:1189–1200.
- Avila FW, Cohen AB, Ameerudeen FS, Duneau D, Suresh S, et al. 2015. Retention of ejaculate by *Drosophila melanogaster* females requires the male-derived mating plug protein PEBme. *Genetics*. 200:1171–1179.
- Bairati A. 1968. Structure and ultrastructure of the male reproductive system in *Drosophila melanogaster* Meig. 2: the genital duct and accessory glands. *Monitore Zoologico Italiano*. 2:105–182.
- Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh Y-P, et al. 2007. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol*. 5:e310.
- Begun DJ, Lindfors HA. 2005. Rapid evolution of genomic Acp complement in the *melanogaster* subgroup of *Drosophila*. *Mol Biol Evol*. 22:2010–2021.
- Begun DJ, Whitley P, Todd BL, Waldrip-Dail HM, Clark AG. 2000. Molecular population genetics of male accessory gland proteins in *Drosophila*. *Genetics*. 156:1879–1888.
- Benitez M, Tatapudy S, Liu Y, Barber DL, Nystul TG. 2019. *Drosophila* anion exchanger 2 is required for proper ovary development and oogenesis. *Dev Biol*. 452:127–133.
- Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc Ser B Stat Methodol*. 57: 289–300.
- Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. 2018. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol*. 36:411–420.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, et al. 2009. BLAST+: architecture and applications. *BMC Bioinformatics*. 10:421.
- Carvalho GB, Kapahi P, Anderson DJ, Benzer S. 2006. Allocrine modulation of feeding behavior by the sex peptide of *Drosophila*. *Curr Biol*. 16:692–696.
- Cavener DR. 1985. Coevolution of the glucose dehydrogenase gene and the ejaculatory duct in the genus *Drosophila*. *Mol Biol Evol*. 2: 141–149.
- Clark AG, Aguadé M, Prout T, Harshman LG, Langley CH. 1995. Variation in sperm displacement and its association with accessory gland protein loci in *Drosophila melanogaster*. *Genetics*. 139: 189–201.
- Colquitt BM, Merullo DP, Konopka G, Roberts TF, Brainard MS. 2021. Cellular transcriptomics reveals evolutionary identities of songbird vocal circuits. *Science*. 371: eabd9704.
- Corrigan L, Redhai S, Leiblich A, Fan S-J, Perera SMW, et al. 2014. BMP-regulated exosomes from *Drosophila* male reproductive glands reprogram female behavior. *J Cell Biol*. 206:671–688.
- Cridland JM, Majane AC, Sheehy HK, Begun DJ. 2020. Polymorphism and divergence of novel gene expression patterns in *Drosophila melanogaster*. *Genetics*. 216:79–93.
- Cridland JM, Majane AC, Zhao L, Begun DJ. 2021. Population biology of accessory gland-expressed de novo genes in *Drosophila melanogaster*. *Genetics*.
- Dapper AL, Wade MJ. 2020. Relaxed selection and the rapid evolution of reproductive genes. *Trends Genet*. 36:640–649.
- Darwin C. 1871. *The Descent of Man and Selection in Relation to Sex*. London: John Murray.
- DiBenedetto AJ, Harada HA, Wolfner MF. 1990. Structure, cell-specific expression, and mating-induced regulation of a *Drosophila melanogaster* male accessory gland gene. *Dev Biol*. 139:134–148.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, et al. 2013. STAR: ultrafast universal RNA-Seq aligner. *Bioinformatics*. 29: 15–21.

- Ellegren H, Parsch J. 2007. The evolution of sex-biased genes and sex-biased gene expression. *Nat Rev Genet.* 8:689–698.
- Feregrino C, Tschopp P. 2021. Assessing evolutionary and developmental transcriptome dynamics in homologous cell types. *Dev Dyn.* 2021;1–18.
- Findlay GD, MacCoss MJ, Swanson WJ. 2009. Proteomic discovery of previously unannotated, rapidly evolving seminal fluid genes in *Drosophila*. *Genome Res.* 19:886–896.
- Findlay GD, Sitnik JL, Wang W, Aquadro CF, Clark NL, et al. 2014. Evolutionary rate covariation identifies new members of a protein network required for *Drosophila melanogaster* female post-mating responses. *PLoS Genet.* 10:e1004108.
- Fiumera AC, Dumont BL, Clark AG. 2005. Sperm competitive ability in *Drosophila melanogaster* associated with variation in male reproductive proteins. *Genetics.* 169:243–257.
- Fraïsse C, Puixeu Sala G, Vicoso B. 2019. Pleiotropy modulates the efficacy of selection in *Drosophila melanogaster*. *Mol Biol Evol.* 36:500–515.
- Gligorov D, Sitnik JL, Maeda RK, Wolfner MF, Karch F. 2013. A novel function for the Hox gene Abd-B in the male accessory gland regulates the long-term female post-mating response in *Drosophila*. *PLoS Genet.* 9:e1003395.
- Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol.* 29:644–652.
- Hallacli E, Lipp M, Georgiev P, Spielman C, Cusack S, et al. 2012. Msl1-mediated dimerization of the dosage compensation complex is essential for male X-chromosome regulation in *Drosophila*. *Mol Cell.* 48:587–600.
- Heifetz Y, Lung O, Frongillo EA, Jr, Wolfner MF. 2000. The *Drosophila* seminal fluid protein Acp26Aa stimulates release of oocytes by the ovary. *Curr Biol.* 10:99–102.
- Hodge RD, Bakken TE, Miller JA, Smith KA, Barkan ER, et al. 2019. Conserved cell types with divergent features in human versus mouse cortex. *Nature.* 573:61–68.
- Hollis B, Koppik M, Wensing KU, Ruhmann H, Genzoni E, et al. 2019. Sexual conflict drives male manipulation of female postmating responses in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A.* 116:8437–8444.
- Holloway AK, Begun DJ. 2004. Molecular evolution and population genetics of duplicated accessory gland protein genes in *Drosophila*. *Mol Biol Evol.* 21:1625–1628.
- Hopkins BR, Sepil I, Bonham S, Miller T, Charles PD, et al. 2019. BMP signaling inhibition in *Drosophila* secondary cells remodels the seminal proteome and self and rival ejaculate functions. *Proc Natl Acad Sci U S A.* 116:24719–24728.
- Howe KL, Contreras-Moreira B, De Silva N, Maslen G, Akanni W, et al. 2020. Ensembl Genomes 2020-enabling non-vertebrate genomic research. *Nucleic Acids Res.* 48:D689–D695.
- Immarigeon C, Frei Y, Delbare SYN, Gligorov D, Machado Almeida P, et al. 2021. Identification of a micropeptide and multiple secondary cell genes that modulate *Drosophila* male reproductive success. *Proc Natl Acad Sci U S A.* 118e2001897118.
- Isaac RE, Li C, Leedale AE, Shirras AD. 2010. *Drosophila* male sex peptide inhibits siesta sleep and promotes locomotor activity in the post-mated female. *Proc Biol Sci.* 277:65–70.
- Junell A, Uvell H, Davis MM, Edlundh-Rose E, Antonsson A, et al. 2010. The POU transcription factor drifter/ventral veinless regulates expression of *Drosophila* immune defense genes. *Mol Cell Biol.* 30:3672–3684.
- Kalb JM, DiBenedetto AJ, Wolfner MF. 1993. Probing the function of *Drosophila melanogaster* accessory glands by directed cell ablation. *Proc Natl Acad Sci U S A.* 90:8093–8097.
- Kern AD, Jones CD, Begun DJ. 2004. Molecular population genetics of male accessory gland proteins in the *Drosophila simulans* complex. *Genetics.* 167:725–735.
- Kimura M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.
- Kim Y-J, Bartalska K, Audsley N, Yamanaka N, Yapici N, et al. 2010. MIPs are ancestral ligands for the sex peptide receptor. *Proc Natl Acad Sci U S A.* 107:6520–6525.
- Kruskal WH, Wallis WA. 1952. Use of ranks in one-criterion variance analysis. *J Am Stat Assoc.* 47:583–621.
- LaFlamme BA, Ravi Ram K, Wolfner MF. 2012. The *Drosophila melanogaster* seminal fluid protease ‘Seminase’ regulates proteolytic and post-mating reproductive processes. *PLoS Genet.* 8:e1002435.
- Laflamme BA, Wolfner MF. 2013. Identification and function of proteolytic regulators in seminal fluid. *Mol Reprod Dev.* 80:80–101.
- La Manno G, Gyllborg D, Codeluppi S, Nishimura K, Salto C, et al. 2016. Molecular diversity of midbrain development in mouse, human, and stem cells. *Cell.* 167:566–580.e19.
- Langley CH, Stevens K, Cardeno C, Lee YCG, Schrider DR, et al. 2012. Genomic variation in natural populations of *Drosophila melanogaster*. *Genetics.* 192:533–598.
- Leader DP, Krause SA, Pandit A, Davies SA, Dow JAT. 2018. FlyAtlas 2: a new version of the *Drosophila melanogaster* expression atlas with RNA-Seq, miRNA-Seq and sex-specific data. *Nucleic Acids Res.* 46:D809–D815.
- Leiblich A, Marsden L, Gandy C, Corrigan L, Jenkins R, et al. 2012. Bone morphogenetic protein- and mating-dependent secretory cell growth and migration in the *Drosophila* accessory gland. *Proc Natl Acad Sci U S A.* 109:19292–19297.
- Liang C, Musser JM, Cloutier A, Prum RO, Wagner GP. 2018. Pervasive correlated evolution in gene expression shapes cell and tissue type transcriptomes. *Genome Biol Evol.* 10:538–552.
- Li H, Janssens J, De Waegeneer M, Saroja Kolluru S, Davie K, et al. 2021. Fly cell atlas: a single-cell transcriptomic atlas of the adult fruit fly. *bioRxiv* 2021.07.04.451050.
- Liu H, Kubli E. 2003. Sex-peptide is the molecular basis of the sperm effect in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A.* 100:9929–9933.
- Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, et al. 2020. CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* 48: D265–D268.
- Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, et al. 2012. The *Drosophila melanogaster* genetic reference panel. *Nature.* 482:173–178.
- Macosko EZ, Basu A, Satija R, Nemes J, Shekhar K, et al. 2015. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell.* 161:1202–1214.
- Maeda RK, Sitnik JL, Frei Y, Prince E, Gligorov D, et al. 2018. The lncRNA male-specific abdominal plays a critical role in *Drosophila* accessory gland development and male fertility. *PLoS Genet.* 14: e1007519.
- Mahadevaraju S, Fear JM, Akeju M, Galletta BJ, Pinheiro MML, et al. 2021. Dynamic sex chromosome expression in *Drosophila* male germ cells. *Nat Commun.* 12:1–16.
- Martelotto L. 2019. ‘Frankenstein’ protocol for nuclei isolation from fresh and frozen tissue for snRNAseq. May 27, 2019. <https://www.protocols.io/view/frankenstein-protocol-for-nuclei-isolation-from-f-3eqgjdw?version=warning=no>. (Accessed: 2021 November 29).
- McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature.* 351:652–654.
- McGeary MK, Findlay GD. 2020. Molecular evolution of the sex peptide network in *Drosophila*. *J Evol Biol.* 33:629–641.
- McInnes L, John H, Nathaniel S, Großberger L. 2018. UMAP: Uniform Manifold Approximation and Projection. *J Open Source Softw.* 3:861.

- Medina JF, Recalde S, Prieto J, Lecanda J, Saez E, et al. 2003. Anion exchanger 2 is essential for spermiogenesis in mice. *Proc Natl Acad Sci U S A*. 100:15847–15852.
- Meiklejohn CD, Parsch J, Ranz JM, Hartl DL. 2003. Rapid evolution of male-biased gene expression in *Drosophila*. *Proc Natl Acad Sci U S A*. 100:9894–9899.
- Meller VH, Wu KH, Roman G, Kuroda MI, Davis RL. 1997. roX1 RNA paints the X chromosome of male *Drosophila* and is regulated by the dosage compensation system. *Cell*. 88:445–457.
- Meslin C, Cherwin TS, Plakke MS, Hill J, Small BS, et al. 2017. Structural complexity and molecular heterogeneity of a butterfly ejaculate reflect a complex history of selection. *Proc Natl Acad Sci U S A*. 114:E5406–E5413.
- Minami R, Wakabayashi M, Sugimori S, Taniguchi K, Kokuryo A, et al. 2012. The homeodomain protein defective proventriculus is essential for male accessory gland development to enhance fecundity in *Drosophila*. *PLoS One*. 7:e32302.
- Mueller JL, Ravi Ram K, McGraw LA, Bloch Qazi MC, Siggia ED, et al. 2005. Cross-species comparison of *Drosophila* male accessory gland protein genes. *Genetics*. 171:131–143.
- Mukherjee AS, Beermann W. 1965. Synthesis of ribonucleic acid by the X-chromosomes of *Drosophila melanogaster* and the problem of dosage compensation. *Nature*. 207:785–786.
- Musser JM, Wagner GP. 2015. Character trees from transcriptome data: origin and individuation of morphological characters and the so-called 'species signal'. *J Exp Zool B Mol Dev Evol*. 324:588–604.
- Neubaum DM, Wolfner MF. 1999. Mated *Drosophila melanogaster* females require a seminal fluid protein, Acp36DE, to store sperm efficiently. *Genetics*. 153:845–857.
- Odhiambo TR, Kokwaro ED, Sequeira LM. 1983. Histochemical and ultrastructural studies of the male accessory reproductive glands and spermatophore of the Tsetse, *Glossina morsitans* Westwood. *Int J Trop Insect Sci*. 4:227–236.
- Overend G, Luo Y, Henderson L, Douglas AE, Davies SA, et al. 2016. Molecular mechanism and functional significance of acid generation in the *Drosophila* midgut. *Sci Rep*. 6:27242.
- Patlar B, Jayaswal V, Ranz JM, Civetta A. 2021. Nonadaptive molecular evolution of seminal fluid proteins in *Drosophila*. *Evolution*. 75:2102–2113.
- Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*. 14:417–419.
- Peng J, Zipperlen P, Kubli E. 2005. *Drosophila* sex-peptide stimulates female innate immune system after mating via the Toll and Imd pathways. *Curr Biol*. 15:1690–1694.
- Poels J, Van Loy T, Vandersmissen HP, Van Hiel B, Van Soest S, et al. 2010. Myoinhibiting peptides are the ancestral ligands of the promiscuous *Drosophila* sex peptide receptor. *Cell Mol Life Sci*. 67:3511–3522.
- Prince E, Kroeger B, Gligorov D, Wilson C, Eaton S, et al. 2019. Rab-mediated trafficking in the secondary cells of *Drosophila* male accessory glands and its role in fecundity. *Traffic* 20:137–151.
- Ranz JM, Castillo-Davis CI, Meiklejohn CD, Hartl DL. 2003. Sex-dependent gene expression and evolution of the *Drosophila* transcriptome. *Science*. 300:1742–1745.
- Ravi Ram K, Wolfner MF. 2007. Seminal influences: *Drosophila* Acp's and the molecular interplay between males and females during reproduction. *Integr Comp Biol*. 47:427–445.
- Ravi Ram K, Wolfner MF. 2009. A network of interactions among seminal proteins underlies the long-term postmating response in *Drosophila*. *Proc Natl Acad Sci U S A*. 106:15384–15389.
- Rexhepaj A, Liu H, Peng J, Choffat Y, Kubli E. 2003. The sex-peptide DUP99B is expressed in the male ejaculatory duct and in the cardia of both sexes. *Eur J Biochem*. 270:4306–4314.
- Richmond RC, Nielsen KM, Brady JP, Snella EM. 1990. Physiology, biochemistry and molecular biology of the Est-6 locus in *Drosophila melanogaster*. In: JSF Barker, WT Starmer, RJ MacIntyre, editors. *Ecological and Evolutionary Genetics of Drosophila*. Boston, MA: Springer US. p. 273–292.
- Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, et al. 2015. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 43:e47.
- Samakovlis C, Kylsten P, Kimbrell DA, Engström A, Hultmark D. 1991. The Andropin gene and its product, a male-specific antibacterial peptide in *Drosophila melanogaster*. *EMBO J*. 10:163–169.
- Satija R, Farrell JA, Gennert D, Schier AF, Regev A. 2015. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol*. 33:495–502.
- Saudan P, Hauck K, Soller M, Choffat Y, Ottiger M, et al. 2002. Ductus ejaculatorius peptide 99B (DUP99B), a novel *Drosophila melanogaster* sex-peptide pheromone. *Eur J Biochem*. 269:989–997.
- Schully SD, Hellberg ME. 2006. Positive selection on nucleotide substitutions and indels in accessory gland proteins of the *Drosophila pseudoobscura* subgroup. *J Mol Evol*. 62:793–802.
- Sebé-Pedrós A, Saudemont B, Chomsky E, Plessier F, Mailhé M-P, et al. 2018. Cnidarian cell type diversity and regulation revealed by whole-organism single-cell RNA-seq. *Cell*. 173:1520–1534.e20.
- Sepil I, Hopkins BR, Dean R, Thézénas M-L, Charles PD, et al. 2018. Quantitative proteomics identification of seminal fluid proteins in male *Drosophila melanogaster*. *Mol Cell Proteomics*. 18:S46–S58.
- Simpson GG. 1944. *Tempo and Mode in Evolution*. Chichester, NY: West Sussex: Columbia University Press.
- Singh A, Buehner NA, Lin H, Baranowski KJ, Findlay GD, et al. 2018. 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.
- Sitnik JL, Gligorov D, Maeda RK, Karch F, Wolfner MF. 2016. The female post-mating response requires genes expressed in the secondary cells of the male accessory gland in *Drosophila melanogaster*. *Genetics*. 202:1029–1041.
- Smith NGC, Eyre-Walker A. 2002. Adaptive protein evolution in *Drosophila*. *Nature*. 415:1022–1024.
- Sokal RR, Rohlf JF. 2012. *Biometry: The Principles and Practice of Statistics in Biological Research*. 4th ed. New York, NY: Freeman and Co.
- Soller M, Bownes M, Kubli E. 1999. Control of oocyte maturation in sexually mature *Drosophila* females. *Dev Biol*. 208:337–351.
- Street K, Risso D, Fletcher RB, Das D, Ngai J, et al. 2018. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*. 19:477.
- Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, et al. 2019. Comprehensive integration of single-cell data. *Cell*. 177:1888–1902.e21.
- Styger D. 1992. *Molekulare Analyse Des Sexpeptidgens Aus Drosophila melanogaster*. Zurich, Switzerland: University of Zurich.
- Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF. 2001. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. *Proc Natl Acad Sci U S A*. 98:7375–7379.
- Takemori N, Yamamoto M-T. 2009. Proteome mapping of the *Drosophila melanogaster* male reproductive system. *Proteomics*. 9:2484–2493.
- Tosches MA, Yamawaki TM, Naumann RK, Jacobi AA, Tushev G, et al. 2018. Evolution of pallium, hippocampus, and cortical cell

- types revealed by single-cell transcriptomics in reptiles. *Science*. 360: 881–888.
- Tsaur SC, Ting CT, Wu CI. 1998. Positive selection driving the evolution of a gene of male reproduction, *Acp26Aa*, of *Drosophila*: II. Divergence versus polymorphism. *Mol Biol Evol*. 15: 1040–1046.
- Ulmschneider B, Grillo-Hill BK, Benitez M, Azimova DR, Barber DL, et al. 2016. Increased intracellular pH is necessary for adult epithelial and embryonic stem cell differentiation. *J Cell Biol*. 215: 345–355.
- Van den Berge K, Roux de Bézieux H, Street K, Saelens W, Cannoodt R, et al. 2020. Trajectory-based differential expression analysis for single-cell sequencing data. *Nat Commun*. 11: 1201.
- Wagstaff BJ, Begun DJ. 2005. Molecular population genetics of accessory gland protein genes and testis-expressed genes in *Drosophila mojavensis* and *D. arizonae*. *Genetics*. 171: 1083–1101.
- Wang J, Sun H, Jiang M, Li J, Zhang P, et al. 2021. Tracing cell-type evolution by cross-species comparison of cell atlases. *Cell Rep*. 34: 108803.
- Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, et al. 2013. CPAT: coding-potential assessment tool using an alignment-free logistic regression model. *Nucleic Acids Res*. 41: e74.
- White MJD. 1977. *Animal Cytology and Evolution*. CUP Archive.
- Whittle CA, Extavour CG. 2019. Selection shapes turnover and magnitude of sex-biased expression in *Drosophila gonads*. *BMC Evol Biol*. 19: 60.
- Wickham H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Berlin: Springer.
- Wigby S, Brown NC, Allen SE, Misra S, Sitnik JL, et al. 2020. The *Drosophila* seminal proteome and its role in postcopulatory sexual selection. *Philos Trans R Soc Lond B Biol Sci*. 375: 20200072.
- Wolfner MF, Harada HA, Bertram MJ, Stelick TJ, Kraus KW, et al. 1997. New genes for male accessory gland proteins in *Drosophila melanogaster*. *Insect Biochem Mol Biol*. 27: 825–834.
- Wong A, Turchin MC, Wolfner MF, Aquadro CF. 2008. Evidence for positive selection on *Drosophila melanogaster* seminal fluid protease homologs. *Mol Biol Evol*. 25: 497–506.
- Xue L, Noll M. 2000. *Drosophila* female sexual behavior induced by sterile males showing copulation complementation. *Proc Natl Acad Sci U S A*. 97: 3272–3275.
- Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, et al. 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. *Bioinformatics*. 21: 650–659.
- Yapici N, Kim Y-J, Ribeiro C, Dickson BJ. 2008. A receptor that mediates the post-mating switch in *Drosophila* reproductive behaviour. *Nature*. 451:33–37.

Communicating editor: P. Wittkopp