Title

Evolution of Gene Expression in the Drosophila Accessory Gland

Permalink

https://escholarship.org/uc/item/6dt2756x

Author

Majane, Alex

Publication Date

2022

Peer reviewed|Thesis/dissertation

Evolution of Gene Expression in the Drosophila Accessory Gland

By

ALEX MAJANE

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Integrative Genetics and Genomics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

David Begun, Chair

Joanna Chiu

Bruce Rannala

Committee in Charge

2022

TABLE OF CONTENTS

ABSTRACT

Gene expression evolution leads to much of the diversity observed among species. Expression divergence may be driven by regulatory evolution in *cis* loci and *trans* factors. Additionally, regulatory evolution may lead to hybrid incompatibilities in the form of misexpression linked to sterility. Many studies have shown that expression divergence varies widely among tissues, and male gonadal tissues tend to evolve particularly rapidly. There has been less work on somatic male reproductive tissues, and little work on the evolutionary properties of different cell types. Additionally, our understanding of how gene regulatory evolution and hybrid misexpression vary among tissues is especially limited. In this study I use the Drosophila male accessory gland and ejaculatory duct—organs that produce seminal fluid—as a for evolution of expression and regulation in cells and tissues. In Chapter I, I used single-nucleus RNA-Seq of the accessory gland and ejaculatory duct three Drosophila species to comprehensively describe the cell diversity of these tissues for the first time. I found that rates of transcriptome divergence were heterogenous among cell types and lineages, with ejaculatory duct cells evolving faster than the cells of the accessory gland. I also found that proteins characteristic of each cell type have variable rates of adaptive substitutions. In Chapter II, I used allele-specific expression in accessory glands and ejaculatory ducts of hybrids between *D. melanogaster* and *D. simulans* to estimate regulatory changes in *cis* and *trans* and quantify hybrid misexpression. I found an unexpected excess of *trans*-regulatory divergence in contrast to the prevailing expectation in the literature. I also found that the accessory gland has limited misexpression, potentially indicative of less hybrid dysgenesis in comparison to gonads. I integrated ATAC-Seq and RNA-Seq data to show that differences in chromatin accessibility correlate with divergence in both *cis* and *trans*. Taken together, these studies contribute to our knowledge of the unique evolution of the accessory gland and ejaculatory duct and underline the importance of tissue- and cell-type specific differences in expression and regulatory divergence.

CHAPTER 1: Single-nucleus transcriptomes reveal functional and evolutionary properties of the Drosophila accessory gland

Alex C Majane, Julie M Cridland, David J Begun, Single-nucleus transcriptomes reveal evolutionary and functional properties of cell types in the *Drosophila* accessory gland, *Genetics*, Volume 220, Issue 2, February 2022, iyab213, https://doi.org/10.1093/genetics/iyab213

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 post-copulatory 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 post-copulatory traits. Recent work has shown that these cell types have overlapping but distinct effects on female post-copulatory 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 define 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 post-copulatory traits.

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 to 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) (Ranz et al. 2003; Meiklejohn et al. 2003; Assis, Zhou, and Bachtrog 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 Arendt et al. 2016; Musser and Wagner 2015). Single-cell data in evolutionary contexts have generally been applied to distantly related taxa (Tosches et al. 2018; Hodge et al. 2019; Liang et al. 2018), typically focusing on cell type diversity (Sebé-Pedrós et al. 2018; La Manno et al. 2016; Colquitt et al. 2021; Feregrino and Tschopp 2021; J. 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 (AG) and the ejaculatory duct, induce a set of physiological and behavioral changes in females collectively referred to as the post-mating response (PMR; reviewed in Ravi Ram and Wolfner 2007). In *D. melanogaster*, the PMR includes increased rates of egg laying (Soller, Bownes, and Kubli 1999; Heifetz et al. 2000), decreased receptivity to re- mating (Liu and Kubli 2003), storage of sperm in specialized reproductive tract tissues (Neubaum and Wolfner 1999), elevated immune response (Peng, Zipperlen, and Kubli 2005), elevated feeding rates (Carvalho et al. 2006), increased activity rate, and decreased sleep (Isaac R. Elwyn et al. 2010). Genetic variation in Sfps may also play a role in the outcome of sperm competition (Clark et al. 1995; Fiumera, Dumont, and Clark 2005). Population genetic and comparative analyses of these proteins suggest they evolve unusually rapidly, often under the influence of directional selection (Begun et al. 2000; Tsaur, Ting, and Wu 1998; Aguadé 1999). These genes are frequently gained or lost during evolution (Wagstaff and Begun 2005, Muller *et al*. 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 are smaller, hexagonal, and squamous, while secondary cells are much larger, spherical, project into the lumen of the gland, and contain extensive vacuole-like compartments (Bairati 1968; Prince et al. 2018). Main cells, which constitute the vast majority of AG cells, are necessary and sufficient to initiate the PMR (Kalb, DiBenedetto, and Wolfner 1993; Sitnik et al. 2016; Hopkins et al. 2019). Secondary cells, 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 secondary cell 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 (Sepil et al. 2018; Rexhepaj et al. 2003; Takemori and Yamamoto 2009). While the duct and its products contribute to the PMR (Rexhepaj et al. 2003; Saudan et al. 2002; Xue and Noll 2000), 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, DiBenedetto, and Wolfner 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 accessory gland and ejaculatory duct in three closely related Drosophila species, *D. melanogaster*, *D. simulans*, and *D. yakuba*. We characterize main cells (MC), secondary cells (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 Information. We used the following sequenced stocks to generate AG and ejaculatory duct transcriptomes from 2-3 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 10X Genomics Chromium platform and Illumina sequencing.

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

We parsed the 10X 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, Pyl, and Huber 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 (2020 version 2).

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 orthologues. 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 $\log_2$(fold-change), hereafter referred to as logFC, requirement of 0.25. We filtered marker genes to those with Bonferroni-corrected p-values less than 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 percent 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 genes with tradeSeq (Van den Berge et al. 2020).

We used limma v3.42.2 (Ritchie et al. 2015) to infer differentially expressed (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 cut-offs 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 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 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 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, Puixeu Sala, and Vicoso 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 Supplemental 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 (L. 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 (Fig. 1*A-C*). We then used marker gene identification along with the relative sizes of clusters to assign cell type identity to clusters, identifying MC, SC, and EDC. MC

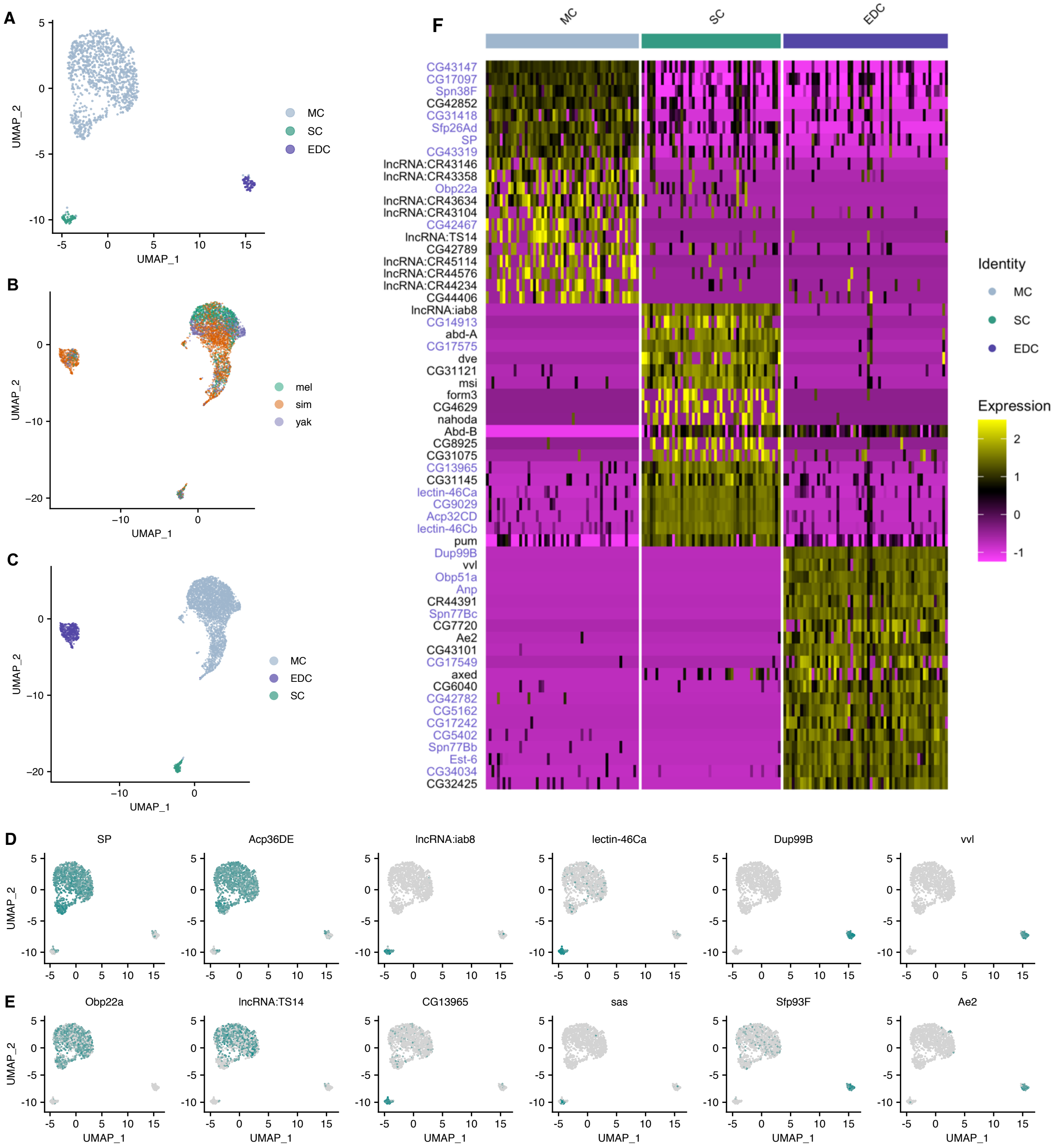

Figure 1. (*A*) UMAP showing clustering of *mel* single-nucleus transcriptomes into three major cell types: main cells (MC), secondary cells (SC), and ejaculatory duct cells (EDC). (*B*) Nuclei from three species cluster concordantly, (*C*) into the same three major cell types. Differences between (*A*) 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, (*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. Sfps 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 Fig. S1.

were identified as the cluster with the largest number of cells, and based on markers *Sex Peptide* (*SP*) (Styger 1992), *Acp36DE* (Wolfner et al. 1997), and *Acp95EF* (DiBenedetto, Harada, and Wolfner 1990; Kalb, DiBenedetto, and Wolfner 1993) (Fig. 1*D*). 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* orthologues are not annotated in *sim* or *yak*) (Fig. 1*D*). EDC were identified by expression of *vvl* (Junell et al. 2010) and *Dup99B* (Rexhepaj et al. 2003) (Fig. 1*D*). We additionally used *Abd-B* to characterize both SC and EDC (Maeda et al. 2018; Gligorov et al. 2013). 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 subpopulations (see Transcriptome heterogeneity among main cells below). For downstream analyses, we merged these sub-clusters 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 (Table S1, Dataset S1, Fig. 1*D,E*). Details of some of the most notable marker genes specific to each cell type can be found in our Supplemental Results.

Cell type transcriptomes in the *Drosophila melanogaster* accessory gland

Thresholding marker genes as expressed in at least 25% of cells in the focal cell type and minimum $\log_2$ of the fold change (logFC) = 0.25, we identified 540 *mel* marker genes (Fig. 1*F*, Dataset S1). Of these, 128 are annotated Sfps identified from proteomic studies of the male ejaculate (Findlay, MacCoss, and Swanson 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 (Swanson et al. 2001; Wolfner et al. 1997; Kalb, DiBenedetto, and Wolfner 1993). Marker Sfps for SC and EDC are summarized in 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 Dataset S10.

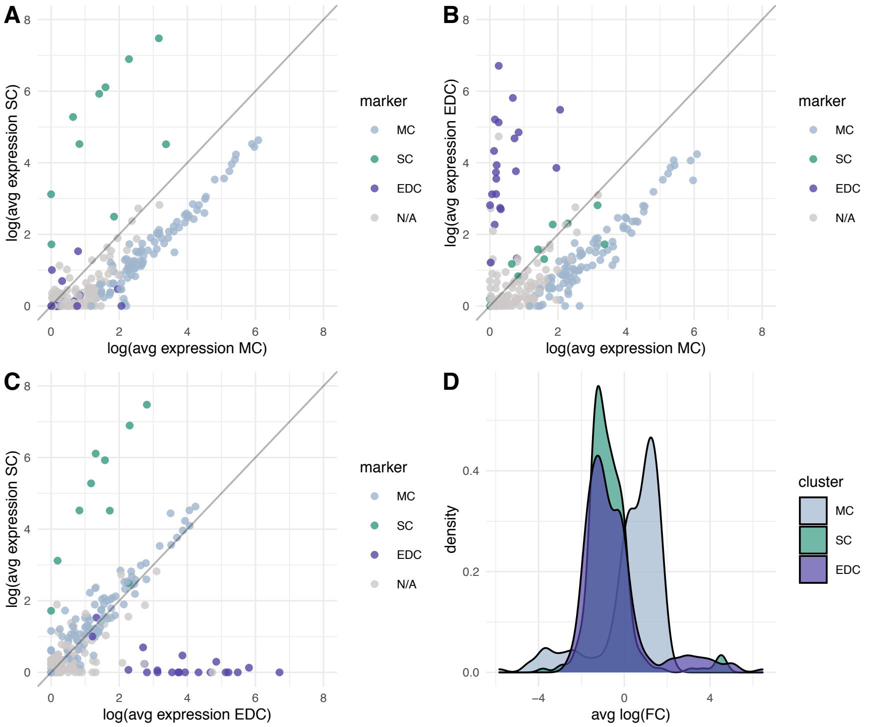

Figure 2. Expression of Sfps tends to be highly cell type-biased. (*A-C*) Expression levels of Sfps compared among cell types show a general pattern of MC enrichment and cell type bias. Colors indicate marker gene status for each SFP; N/A indicates that a gene does not show a strong cell type bias. (*D*) 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.

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 $\log_2$(average expression) for the focal cell type and the average logFC vs. all other cell types. Among the 224 Sfps detected in the data (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 (Fig. 2). Highly

MC-biased Sfps tend to also show expression in SC, though at a substantially lower level. Even among non-marker Sfps we observe a trend towards greater MC expression than SC expression (Fig. 2*A*, SC expression vs. MC expression gives a slope = 0.73, r^2 = 0.82). EDC vs. MC comparison for non-marker Sfps exhibits a similar pattern (Fig. 2*B*, 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 towards SC (Fig. 2*C*; slope = 0.67, r^2 = 0.598 Among the 97 non-marker 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 logFC of Sfps in MC. vs all other cells skews significantly greater than SC vs all others and EDC vs all others, respectively (Fig. 2*D*). The median logFC 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.

Using all annotated *mel* genes, marker genes for each *mel* cell type reveal both expected and novel markers (Dataset S1). In MC we identify many expected Sfps including *SP* (Fig 1*D*), *Acp36DE*, *Acp26Aa*, and *Acp95EF,* and relatively uncharacterized Sfps including *Obp22a* (Fig. *1E*). The top non-Sfp markers of MC are generally functionally uncharacterized: *CG42852*, *CG43254*, *CG42481*, *CG43392*, *lncRNA:CR43146*, *lncRNA:CR45013*, *CG34041, lncRNA:TS14* (Fig. 1*E*), and the genes *CG44388* and *lncRNA:CR44389*, which are neighbors. Despite its annotation as a lncRNA, *CR44389* possesses a 41 amino acid ORF strongly predicted to have a signal sequence, suggesting it could be a secreted protein. *Ugt50B3,* a UDP-glycosyltransferase, is another strong marker of MC.

Among the 10 Sfps identified as SC markers (Table S1), three were previously known to be SC-specific: *Acp32CD*, *lectin-46Ca* and *lectin-46Cb* (Maeda et al. 2018). Previous work with MC-null mutants identified *Acp32CD* as expressed in SC (Swanson et al. 2001), and here we additionally show that it exhibits very low expression in MC. The Sfps *CG17575, CG3349, CG9029*, *CG13695* (Fig. 1*E*), and *mfas* have also been previously identified as SC-expressed (Gligorov et al. 2013; Sitnik et al. 2016; Immarigeon et al. 2021). Here we show that these Sfps show very low expression in MC and EDC. We also identify the Sfp *Pgant9* as a novel SC marker. We additionally recovered expected non-Sfp markers: *lncRNA:iab8*, *abd-A (Maeda et al. 2018)*, *abd-B*, and *defective proventriculus* (*dve*) (Minami et al. 2012). We also identify non-Sfp SC markers *stranded-at-second* (*sas*) (Fig. 1*E*), *musashi* (*msi*), *form3*, *nahoda*, *CG31121*, *CG4629*, and *CG46430*. Additionally, we discovered that the unannotated transcript *DN2695* (see Identification of unannotated candidate genes in the AG, Table 1, Table S4, Fig. S6) is a strong SC marker.

We identified 24 Sfp EDC markers (Table S1). Of these, 1 had previously been identified as EDC-enriched: *Dup99B*, *Obp51a*, *Spn77Bc*, *Spn77Bb*, *Est-6*, *Gld*, *Anp*, *CG18258*, *CG5162*, *CG17242*, *CG5402, CG34034,* and *CG31704* (Takemori and Yamamoto 2009; Sepil et al. 2018; Samakovlis et al. 1991; Cavener 1985; Saudan et al. 2002). The remainder have not been previously identified as EDC-specific Sfps: *Treh*, *betaggt-I*, *Sfp93F* (Fig. 1*E*)*, 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* (*vvl*) (Junell et al. 2010) and *Abd-B (Gligorov et al. 2013)*. Novel EDC markers are *anion exchanger 2* (*Ae2*) (Fig. 1*E*), *axundead* (*axed*), *single-minded* (*sim*), *CG7720*, *CG43101*, *CG7342*, and *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* main cell subpopulations

During initial analysis we discovered an apparent subcluster of main cells characterized by unique SNN clusters at $k = 4$ and clear separation in UMAP space (Fig. 3*A*). Of a total 1057 MC, 942 are in subcluster one (MCsp1) and 115 are in subcluster two (MCsp2). 349 significant markers (Bonferroni-corrected $p < 0.05$) distinguish these subclusters (Dataset S3). In all three species, these subclusters are apparent and appear in roughly equal proportions (Fig. S2*A*, 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 (Fig. 3*B*). 26 show higher expression in MCsp2, while just eight show higher expression in MCsp1 (Dataset S3). Non-Sfps show the opposite pattern, with 102 genes showing increased expression in MCsp2, and 213 genes with higher expression in MCsp1 (Dataset S3). The most enriched non-Sfp genes for each subpopulation are shown in Fig. 3*D*.

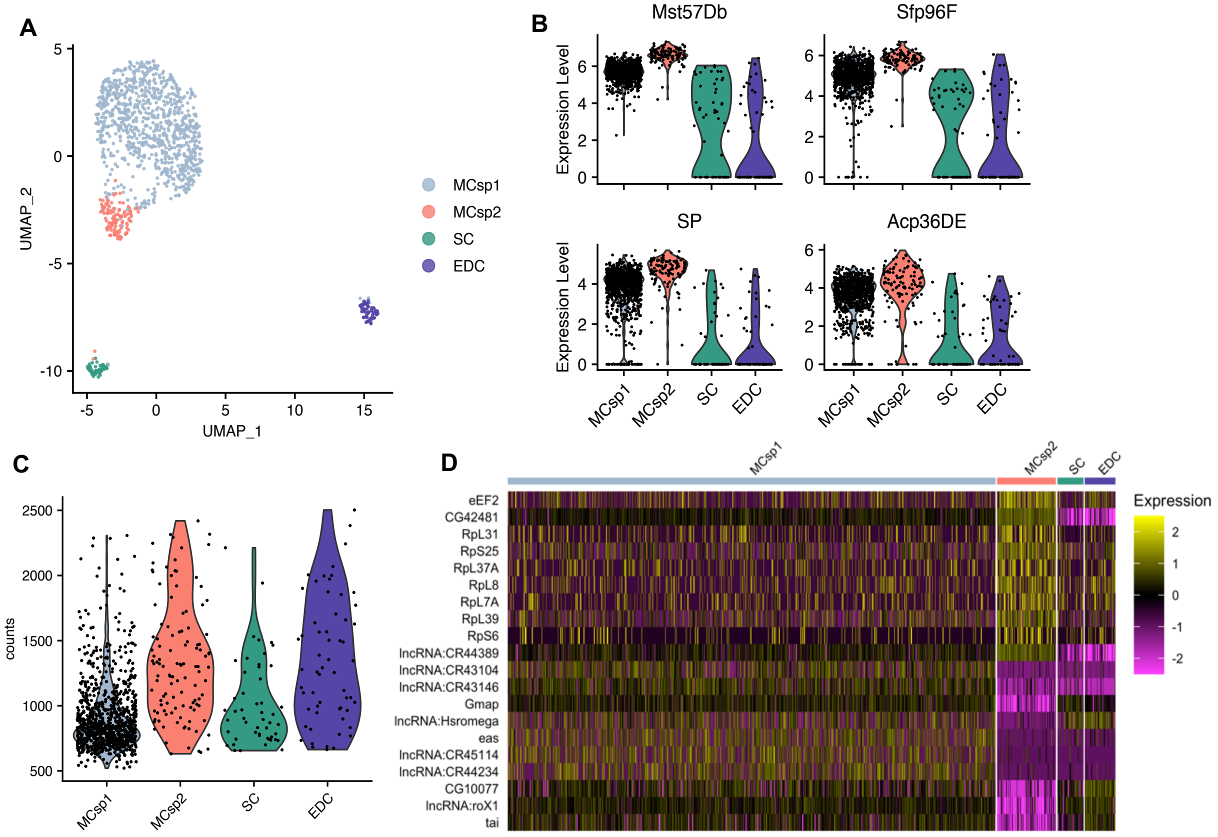

Figure 3. Transcriptome heterogeneity among subpopulations of MC in *mel*. (*A*) Subpopulations of MC are apparent in both UMAP space and SNN clustering with *k* = 4. (*B*) Examples of MC marker Sfps with greater expression in MCsp2. (*C*) MCsp1 has a significantly lower level of RNA counts per cell 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.

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*, *eEF1δ*, and *eEF1α1*, 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 (Fig. 3*C*, Wilcoxon rank sum test, $p < 0.001$). We find this same pattern of lower RNA counts in MCsp1 in *sim* and *yak* (Fig. S2*B*, 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*, *Hsrω*, *CR43104*, *CR43146*, and *CR45114* (Fig. 3*D*). *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 (Fig. S3*A*,*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 intermediate phase cells, but no obvious evidence of pseudotemporal variance in the onset of differential gene expression, pointing to a relatively simple process (Fig. S3*B*). Finally, we observe evidence of finer functional divisions within MC in an apparent third subpopulation (Fig. S2*A*; 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 (Fig. 1*B*,*C*). The top 12 DE genes for each cell type are summarized in Table S2, and expression of DE genes in all cell types can be found in Dataset S9. We found 132 genes that are DE (logFC > 1) in at least one pairwise species contrast among MC (Dataset S6), of which 40 (30%) are Sfps. Among SC we found 106 DE genes (Dataset S7), of which 21 (20%) are Sfps, while in EDC we found 221 (Dataset S8), of which just 32 (14%) are Sfps. The percentage of expressed genes that are DE for each species contrast and cell type (Fig. 4*A*) is significantly heterogeneous (G-test, p < 0.001, 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 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

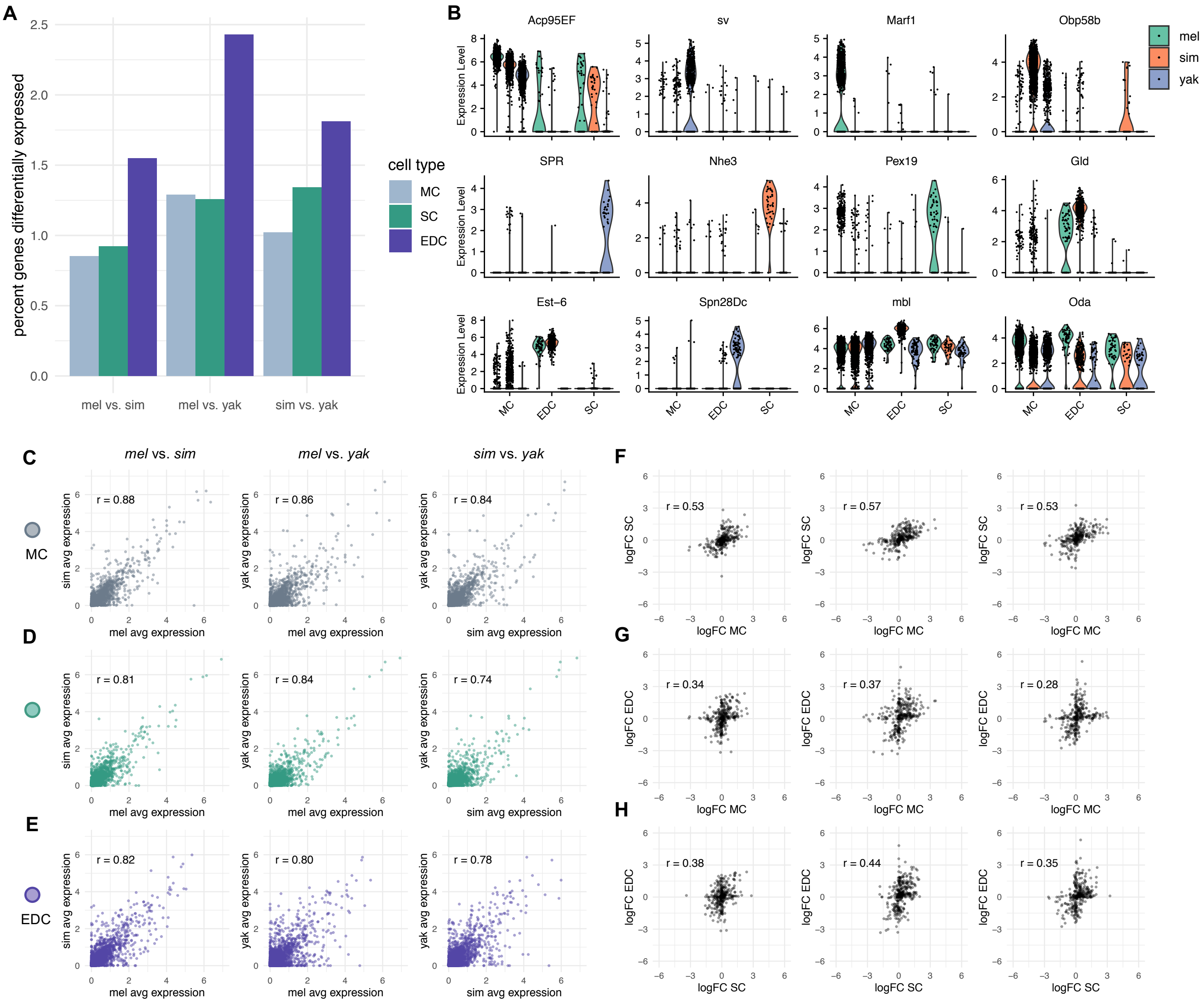

Figure 4. (*A*) Percentage of expressed genes DE by cell type and species contrast (*G*-test, $p < 0.001$). For significance values of pairwise tests see Table S2. (*B*) Examples of differential expression detected in this study. (*C-E*) Pearson correlations of transcriptome-wide expression show cell 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 correlations of logFC 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.

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 S4*A*-*D*). There is a trend towards elevated MC enrichment compared to 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 (Fig. S4*E*-*F*).

We used Pearson 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 (Fig. 4B; $rMC_{mel\text{-}sim} = 0.88$, $rMC_{mel\text{-}yak} = 0.86$, and $rMC_{sim\text{-}\mathrm{yak}} = 0.84$). Pearson correlations for SC and EDC are lower overall (Fig. 4*D*-*E*; $rSC_{mel\text{-}sim} = 0.81$, $rSC_{mel\text{-}yak} = 0.84$, $rSC_{sim\text{-}\mathrm{yak}} = 0.74$; $rEDC_{mel\text{-}sim} = 0.82$, $rEDC_{mel\text{-}yak} = 0.80$, $rEDC_{sim\text{-}\mathrm{yak}} = 0.78$). 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 Datasets S6-9, but below we wish to highlight a few interesting examples. The Sfp *Acp95EF* is strongly differentially expressed in MC, which has highest expression in *mel*, lower expression in *sim*, and lowest expression in *yak* (Fig 4*B*). 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* (Fig 4*B*), 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* (Fig 4*B*). 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 a 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 (Fig 4*B*). *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 (Y.-J. 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* (Fig 4*B*), 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* (Fig 4*B*). In general, we observed little DE among SC-biased Sfps. While 24 Spfs 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* (Dataset S7).

The EDC marker gene *Esterase 6* (*Est-6*) is highly expressed in *mel* and *sim*, and much more lowly expressed in *yak* (Fig 4*B*). *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 (Fig 4*B*). 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* (Fig 4*B*). 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 non-markers 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 differentially expressed in MC are MC markers, 75% of genes differentially expressed 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, *muscleblind* (*mbl*) 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 (Fig 4*B*). 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 (Fig 4*B*). We also identified nine cases where genes have shifted in their marker status among species (Fig. 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 (Fig 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 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 (Fig. 4*F-H*). MC and SC have the highest correlations (Fig. 4*F*); $r_{mel\text{-}sim}$ = 0.53, $r_{mel\text{-}yak}$ = 0.57, and $r_{sim\text{-yak}}$ = 0.53. SC and EDC are less correlated (Fig. 4*G*); $r_{mel\text{-}sim}$ = 0.38, $r_{mel\text{-}yak}$ = 0.44, and $r_{sim\text{-yak}}$ = 0.35. MC and EDC have the lowest correlations (Fig. 4*H*): $r_{mel\text{-}sim}$ = 0.34, $r_{mel\text{-}yak}$ = 0.37, and $r_{sim\text{-yak}}$ = 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 to 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 $x = y$ (Fig. 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%) (Fig. 5*A*; pairwise Fisher's exact tests, p = 0.002, p < 0.001, respectively), suggesting that compared to 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

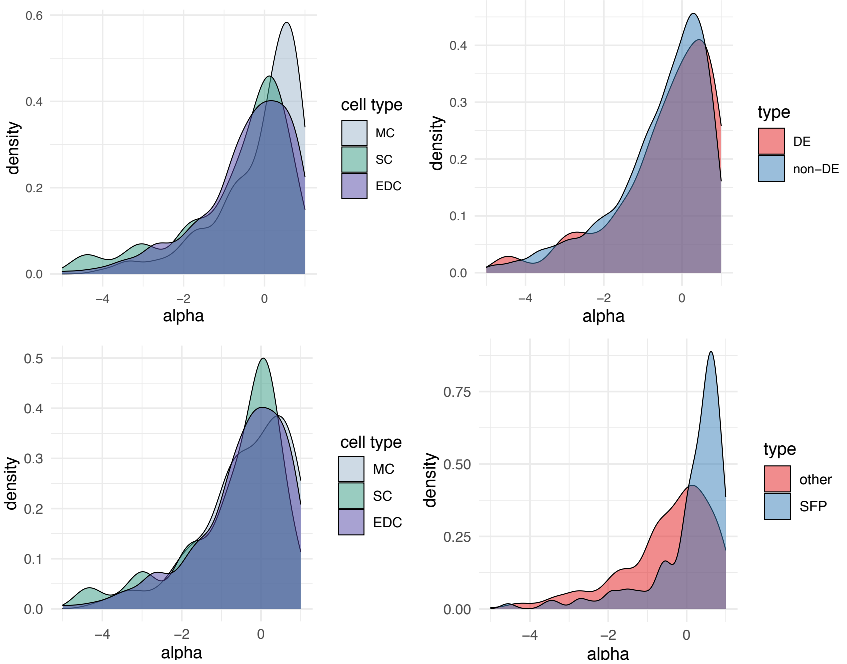

Figure 5. Distributions of α (*mel* population data vs *sim*) for marker genes. (*A*) α values by cell type show that MC markers are significantly greater than SC or EDC (Kruskal-Wallis test, *p* = 0.001). (*B*) Sfp markers have a dramatically greater median α than non-Sfps (Fisher test, *p* < 0.001). (*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).

(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 (Fig. 5*B*; 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 (Fig. 5*C*; 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.

To investigate whether genes that are differentially expressed 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 (Fig. 5*D*; 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 (Tsaur, Ting, and Wu 1998; Begun et al. 2000; Swanson et al. 2001; Kern, Jones, and Begun 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. Non-secreted 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 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, MacCoss, and Swanson 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 Supplemental Methods), we identified 11 unannotated, single-exon genes (Table 1, 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 (Fig. S6*A*), we used RNA-Seq data from FACS-sorted secondary cells (Immarigeon et al. 2021) to further improve our annotations. The median transcript length is 630 bp (range = 352 to 3,102 bp). None of these genes overlap annotated features in the *mel* genome.

Table 1. Unannotated candidate genes expressed in the *D. melanogaster* accessory gland. Length refers to the span of BLAST coordinates. logFC is the cell type with highest fraction of expression compared to the other two cell types. *p* is the result of a Wilcoxon Rank Sum test with Bonferroni correction. See Table S4 for additional details.

transcript	chromosome	length	expression bias	logFC	*p*
			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

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 to expressed annotated genes, but a relatively low level compared to marker genes (Fig. S6*B*). The two notable exceptions to this trend are *DN2695* in SC, and DN*818* in MC, which are expressed at a more intermediate level among markers. These two candidates additionally pass more stringent criteria (expressed in ≥ 25% of focal cells) to be considered marker genes (Dataset S1). *DN2695*, the 7th 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 79 bp from the EDC marker *sim*, but on the opposite strand. All 11 transcripts are predicted to be non-coding 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, Kokwaro, and Sequeira 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, Ravi Ram, and Wolfner 2012; Singh et al. 2018; Findlay et al. 2014; 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 Cl^{-}/HCO_3^{-} exchange in the midgut (Overend et al. 2016) and ovary (Benitez et al. 2019; Ulmschneider et al. 2016). *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 (Immarigeon et al. 2021; Cridland et al., *in press*). Our analysis revealed strong evidence of transcriptionally distinct main cell 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 non-coding 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 pre-print (Li et al. 2021), we investigated some of our top marker genes (Fig. 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 heterogenous 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 differentially expressed 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 (Minami et al. 2012; Xue and Noll 2000; Gligorov et al. 2013), compared to 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 (Fraïsse, Puixeu

Sala, and Vicoso 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 non-synonymous 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 a 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 differentially expressed 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.

ACKNOWLEDGEMENTS

This work was supported by the National Institutes of Health grant R35 GM134930 to DJB, a National Science Foundation Graduate Research Fellowship to ACM, and a Pilot and Feasibility Program award from the UC Davis Research Core Facilities Program. 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.

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.

LITERATURE CITED

Aguadé, Montserrat. 1999. "Positive Selection Drives the Evolution of the Acp29AB Accessory Gland Protein in Drosophila." *Genetics* 152 (2): 543–51.

Almagro Armenteros, José Juan, Konstantinos D. Tsirigos, Casper Kaae Sønderby, Thomas Nordahl Petersen, Ole Winther, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. 2019. "SignalP 5.0 Improves Signal Peptide Predictions Using Deep Neural Networks." *Nature Biotechnology* 37 (4): 420–23.

Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber. 2015. "HTSeq--a Python Framework to Work with High-Throughput Sequencing Data." *Bioinformatics* 31 (2): 166–69.

Arendt, Detlev, Jacob M. Musser, Clare V. H. Baker, Aviv Bergman, Connie Cepko, Douglas H. Erwin, Mihaela Pavlicev, et al. 2016. "The Origin and Evolution of Cell Types." *Nature Reviews. Genetics* 17 (12): 744–57.

Assis, Raquel, Qi Zhou, and Doris Bachtrog. 2012. “Sex-Biased Transcriptome Evolution in Drosophila.” *Genome Biology and Evolution* 4 (11): 1189–1200.

Avila, Frank W., Allie B. Cohen, Fatima S. Ameerudeen, David Duneau, Shruthi Suresh, Alexandra L. Mattei, and Mariana F. Wolfner. 2015. “Retention of Ejaculate by Drosophila Melanogaster Females Requires the Male-Derived Mating Plug Protein PEBme.” *Genetics* 200 (4): 1171–79.

Bairati, Aurelio. 1968. “Structure and Ultrastructure of the Male Reproductive System in Drosophila Melanogaster Meig. 2: The Genital Duct and Accessory Glands.” *Monitore Zoologico Italiano* 2 (3-4): 105–82.

Begun, David J., Alisha K. Holloway, Kristian Stevens, Ladeana W. Hillier, Yu-Ping Poh, Matthew W. Hahn, Phillip M. Nista, et al. 2007. “Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in Drosophila Simulans.” *PLoS Biology* 5 (11): e310.

Begun, David J., and Heather A. Lindfors. 2005. “Rapid Evolution of Genomic Acp Complement in the Melanogaster Subgroup of Drosophila.” *Molecular Biology and Evolution* 22 (10): 2010–21.

Begun, David J., Penn Whitley, Bridget L. Todd, Heidi M. Waldrip-Dail, and Andrew G. Clark. 2000. “Molecular Population Genetics of Male Accessory Gland Proteins in Drosophila.” *Genetics* 156 (4): 1879–88.

Benitez, Marimar, Sumitra Tatapudy, Yi Liu, Diane L. Barber, and Todd G. Nystul. 2019. “Drosophila Anion Exchanger 2 Is Required for Proper Ovary Development and Oogenesis.” *Developmental Biology* 452 (2): 127–33.

Benjamini, Yoav, and Yosef Hochberg. 1995. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 57 (1): 289–300.

Butler, Andrew, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. 2018. “Integrating Single-Cell Transcriptomic Data across Different Conditions, Technologies, and Species.” *Nature Biotechnology* 36 (5): 411–20.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. “BLAST+: Architecture and Applications.” *BMC Bioinformatics* 10 (December): 421.

Carvalho, Gil B., Pankaj Kapahi, David J. Anderson, and Seymour Benzer. 2006. “Allocrine Modulation of Feeding Behavior by the Sex Peptide of Drosophila.” *Current Biology: CB* 16 (7): 692–96.

Cavener, D. R. 1985. “Coevolution of the Glucose Dehydrogenase Gene and the Ejaculatory Duct in the Genus Drosophila.” *Molecular Biology and Evolution* 2 (2): 141–49.

Clark, A. G., M. Aguadé, T. Prout, L. G. Harshman, and C. H. Langley. 1995. “Variation in Sperm Displacement and Its Association with Accessory Gland Protein Loci in Drosophila Melanogaster.” *Genetics* 139 (1): 189–201.

Colquitt, Bradley M., Devin P. Merullo, Genevieve Konopka, Todd F. Roberts, and Michael S. Brainard. 2021. “Cellular Transcriptomics Reveals Evolutionary Identities of Songbird Vocal Circuits.” *Science* 371 (6530).

Corrigan, Laura, Siamak Redhai, Aaron Leiblich, Shih-Jung Fan, Sumeth M. W. Perera, Rachel Patel, Carina Gandy, et al. 2014. “BMP-Regulated Exosomes from Drosophila Male

Reproductive Glands Reprogram Female Behavior." *The Journal of Cell Biology* 206 (5): 671–88.

Cridland, Julie M., Alex C. Majane, Hayley K. Sheehy, and David J. Begun. 2020. "Polymorphism and Divergence of Novel Gene Expression Patterns in Drosophila Melanogaster." *Genetics* 216 (1): 79–93.

Cridland, Julie M., Alex C. Majane, Li Zhao, and David J. Begun, *in press*. "Population Biology of Accessory Gland-Expressed de Novo Genes in Drosophila Melanogaster." *Genetics*.

Dapper, Amy L., and Michael J. Wade. 2020. "Relaxed Selection and the Rapid Evolution of Reproductive Genes." *Trends in Genetics: TIG* 36 (9): 640–49.

Darwin, Charles. 1871. *The Descent of Man and Selection in Relation to Sex*. John Murray.

DiBenedetto, A. J., H. A. Harada, and M. F. Wolfner. 1990. "Structure, Cell-Specific Expression, and Mating-Induced Regulation of a Drosophila Melanogaster Male Accessory Gland Gene." *Developmental Biology* 139 (1): 134–48.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1): 15–21.

Ellegren, Hans, and John Parsch. 2007. "The Evolution of Sex-Biased Genes and Sex-Biased Gene Expression." *Nature Reviews. Genetics* 8 (9): 689–98.

Feregrino, Christian, and Patrick Tschopp. 2021. "Assessing Evolutionary and Developmental Transcriptome Dynamics in Homologous Cell Types." *bioRxiv*.

Findlay, Geoffrey D., Michael J. MacCoss, and Willie J. Swanson. 2009. "Proteomic Discovery of Previously Unannotated, Rapidly Evolving Seminal Fluid Genes in Drosophila." *Genome Research* 19 (5): 886–96.

Findlay, Geoffrey D., Jessica L. Sitnik, Wenke Wang, Charles F. Aquadro, Nathan L. Clark, and Mariana F. Wolfner. 2014. "Evolutionary Rate Covariation Identifies New Members of a Protein Network Required for Drosophila Melanogaster Female Post-Mating Responses." *PLoS Genetics* 10 (1): e1004108.

Fiumera, Anthony C., Bethany L. Dumont, and Andrew G. Clark. 2005. "Sperm Competitive Ability in Drosophila Melanogaster Associated with Variation in Male Reproductive Proteins." *Genetics* 169 (1): 243–57.

Fraïsse, Christelle, Gemma Puixeu Sala, and Beatriz Vicoso. 2019. "Pleiotropy Modulates the Efficacy of Selection in Drosophila Melanogaster." *Molecular Biology and Evolution* 36 (3): 500–515.

Gligorov, Dragan, Jessica L. Sitnik, Robert K. Maeda, Mariana F. Wolfner, and François Karch. 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 Genetics* 9 (3): e1003395.

Grabherr, Manfred G., Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, et al. 2011. "Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome." *Nature Biotechnology* 29 (7): 644–52.

Hallacli, Erinc, Michael Lipp, Plamen Georgiev, Clare Spielman, Stephen Cusack, Asifa Akhtar, and Jan Kadlec. 2012. "Msl1-Mediated Dimerization of the Dosage Compensation Complex Is Essential for Male X-Chromosome Regulation in Drosophila." *Molecular Cell* 48 (4): 587–600.

Heifetz, Y., O. Lung, E. A. Frongillo Jr, and M. F. Wolfner. 2000. “The Drosophila Seminal Fluid Protein Acp26Aa Stimulates Release of Oocytes by the Ovary.” *Current Biology: CB* 10 (2): 99–102.

Hodge, Rebecca D., Trygve E. Bakken, Jeremy A. Miller, Kimberly A. Smith, Eliza R. Barkan, Lucas T. Graybuck, Jennie L. Close, et al. 2019. “Conserved Cell Types with Divergent Features in Human versus Mouse Cortex.” *Nature* 573 (7772): 61–68.

Hollis, Brian, Mareike Koppik, Kristina U. Wensing, Hanna Ruhmann, Eléonore Genzoni, Berra Erkosar, Tadeusz J. Kawecki, Claudia Fricke, and Laurent Keller. 2019. “Sexual Conflict Drives Male Manipulation of Female Postmating Responses in Drosophila Melanogaster.” *Proceedings of the National Academy of Sciences of the United States of America* 116 (17): 8437–44.

Holloway, Alisha K., and David J. Begun. 2004. “Molecular Evolution and Population Genetics of Duplicated Accessory Gland Protein Genes in Drosophila.” *Molecular Biology and Evolution* 21 (9): 1625–28.

Hopkins, Ben R., Irem Sepil, Sarah Bonham, Thomas Miller, Philip D. Charles, Roman Fischer, Benedikt M. Kessler, Clive Wilson, and Stuart Wigby. 2019. “BMP Signaling Inhibition in Drosophila Secondary Cells Remodels the Seminal Proteome and Self and Rival Ejaculate Functions.” *Proceedings of the National Academy of Sciences of the United States of America* 116 (49): 24719–28.

Howe, Kevin L., Bruno Contreras-Moreira, Nishadi De Silva, Gareth Maslen, Wasiu Akanni, James Allen, Jorge Alvarez-Jarreta, et al. 2020. “Ensembl Genomes 2020-Enabling Non-Vertebrate Genomic Research.” *Nucleic Acids Research* 48 (D1): D689–95.

Immarigeon, Clément, Yohan Frei, Sofie Y. N. Delbare, Dragan Gligorov, Pedro Machado Almeida, Jasmine Grey, Léa Fabbro, et al. 2021. “Identification of a Micropeptide and Multiple Secondary Cell Genes That Modulate Drosophila Male Reproductive Success.” *Proceedings of the National Academy of Sciences of the United States of America* 118 (15).

Isaac R. Elwyn, Li Chenxi, Leedale Amy E., and Shirras Alan D. 2010. “Drosophila Male Sex Peptide Inhibits Siesta Sleep and Promotes Locomotor Activity in the Post-Mated Female.” *Proceedings of the Royal Society B: Biological Sciences* 277 (1678): 65–70.

Junell, Anna, Hanna Uvell, Monica M. Davis, Esther Edlundh-Rose, Asa Antonsson, Leslie Pick, and Ylva Engström. 2010. “The POU Transcription Factor Drifter/Ventral Veinless Regulates Expression of Drosophila Immune Defense Genes.” *Molecular and Cellular Biology* 30 (14): 3672–84.

Kalb, J. M., A. J. DiBenedetto, and M. F. Wolfner. 1993. “Probing the Function of Drosophila Melanogaster Accessory Glands by Directed Cell Ablation.” *Proceedings of the National Academy of Sciences of the United States of America* 90 (17): 8093–97.

Kern, Andrew D., Corbin D. Jones, and David J. Begun. 2004. “Molecular Population Genetics of Male Accessory Gland Proteins in the Drosophila Simulans Complex.” *Genetics* 167 (2): 725–35.

Kimura, Motoo. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge: Cambridge University Press.

Kim, Young-Joon, Katarina Bartalska, Neil Audsley, Naoki Yamanaka, Nilay Yapici, Ju-Youn Lee, Yong-Chul Kim, et al. 2010. “MIPs Are Ancestral Ligands for the Sex Peptide

Receptor." *Proceedings of the National Academy of Sciences of the United States of America* 107 (14): 6520–25.

Kruskal, William H., and W. Allen Wallis. 1952. "Use of Ranks in One-Criterion Variance Analysis." *Journal of the American Statistical Association* 47 (260): 583–621.

LaFlamme, Brooke A., K. Ravi Ram, and Mariana F. Wolfner. 2012. "The Drosophila Melanogaster Seminal Fluid Protease 'Seminase' Regulates Proteolytic and Post-Mating Reproductive Processes." *PLoS Genetics* 8 (1): e1002435.

Laflamme, Brooke A., and Mariana F. Wolfner. 2013. "Identification and Function of Proteolysis Regulators in Seminal Fluid." *Molecular Reproduction and Development* 80 (2): 80–101.

La Manno, Gioele, Daniel Gyllborg, Simone Codeluppi, Kaneyasu Nishimura, Carmen Salto, Amit Zeisel, Lars E. Borm, et al. 2016. "Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells." *Cell* 167 (2): 566–80.e19.

Langley, Charles H., Kristian Stevens, Charis Cardeno, Yuh Chwen G. Lee, Daniel R. Schrider, John E. Pool, Sasha A. Langley, et al. 2012. "Genomic Variation in Natural Populations of Drosophila Melanogaster." *Genetics* 192 (2): 533–98.

Leader, David P., Sue A. Krause, Aniruddha Pandit, Shireen A. Davies, and Julian A. T. Dow. 2018. "FlyAtlas 2: A New Version of the Drosophila Melanogaster Expression Atlas with RNA-Seq, miRNA-Seq and Sex-Specific Data." *Nucleic Acids Research* 46 (D1): D809–15.

Leiblich, Aaron, Luke Marsden, Carina Gandy, Laura Corrigan, Rachel Jenkins, Freddie Hamdy, and Clive Wilson. 2012. "Bone Morphogenetic Protein- and Mating-Dependent Secretory Cell Growth and Migration in the Drosophila Accessory Gland." *Proceedings of the National Academy of Sciences of the United States of America* 109 (47): 19292–97.

Liang, Cong, Jacob M. Musser, Alison Cloutier, Richard O. Prum, and Günter P. Wagner. 2018. "Pervasive Correlated Evolution in Gene Expression Shapes Cell and Tissue Type Transcriptomes." *Genome Biology and Evolution* 10 (2): 538–52.

Li, Hongjie, Jasper Janssens, Maxime De Waegeneer, Sai Saroja Kolluru, Kristofer Davie, Vincent Gardeux, Wouter Saelens, et al. 2021. "Fly Cell Atlas: A Single-Cell Transcriptomic Atlas of the Adult Fruit Fly." *bioRxiv*.

Liu, Huanfa, and Eric Kubli. 2003. "Sex-Peptide Is the Molecular Basis of the Sperm Effect in Drosophila Melanogaster." *Proceedings of the National Academy of Sciences of the United States of America* 100 (17): 9929–33.

Lu, Shennan, Jiyao Wang, Farideh Chitsaz, Myra K. Derbyshire, Renata C. Geer, Noreen R. Gonzales, Marc Gwadz, et al. 2020. "CDD/SPARCLE: The Conserved Domain Database in 2020." *Nucleic Acids Research* 48 (D1): D265–68.

Mackay, Trudy F. C., Stephen Richards, Eric A. Stone, Antonio Barbadilla, Julien F. Ayroles, Dianhui Zhu, Sònia Casillas, et al. 2012. "The Drosophila Melanogaster Genetic Reference Panel." *Nature* 482 (7384): 173–78.

Macosko, Evan Z., Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, et al. 2015. "Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets." *Cell* 161 (5): 1202–14.

Maeda, Robert K., Jessica L. Sitnik, Yohan Frei, Elodie Prince, Dragan Gligorov, Mariana F. Wolfner, and François Karch. 2018. "The lncRNA Male-Specific Abdominal Plays a

Critical Role in Drosophila Accessory Gland Development and Male Fertility." *PLoS Genetics* 14 (7): e1007519.

Mahadevaraju, Sharvani, Justin M. Fear, Miriam Akeju, Brian J. Galletta, Mara M. L. Pinheiro, Camila C. Avelino, Diogo C. Cabral-de-Mello, et al. 2021. "Dynamic Sex Chromosome Expression in Drosophila Male Germ Cells." *Nature Communications* 12 (1): 1–16.

Martelotto, Luciano. 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.

McDonald, J. H., and M. Kreitman. 1991. "Adaptive Protein Evolution at the Adh Locus in Drosophila." *Nature* 351 (6328): 652–54.

McGeary, Meaghan K., and Geoffrey D. Findlay. 2020. "Molecular Evolution of the Sex Peptide Network in Drosophila." *Journal of Evolutionary Biology* 33 (5): 629–41.

Medina, Juan F., Sergio Recalde, Jesús Prieto, Jon Lecanda, Elena Saez, Colin D. Funk, Paola Vecino, et al. 2003. "Anion Exchanger 2 Is Essential for Spermiogenesis in Mice." *Proceedings of the National Academy of Sciences of the United States of America* 100 (26): 15847–52.

Meiklejohn, Colin D., John Parsch, José M. Ranz, and Daniel L. Hartl. 2003. "Rapid Evolution of Male-Biased Gene Expression in Drosophila." *Proceedings of the National Academy of Sciences of the United States of America* 100 (17): 9894–99.

Meller, V. H., K. H. Wu, G. Roman, M. I. Kuroda, and R. L. Davis. 1997. "roX1 RNA Paints the X Chromosome of Male Drosophila and Is Regulated by the Dosage Compensation System." *Cell* 88 (4): 445–57.

Meslin, Camille, Tamara S. Cherwin, Melissa S. Plakke, Jason Hill, Brandon S. Small, Breanna J. Goetz, Christopher W. Wheat, Nathan I. Morehouse, and Nathan L. Clark. 2017. "Structural Complexity and Molecular Heterogeneity of a Butterfly Ejaculate Reflect a Complex History of Selection." *Proceedings of the National Academy of Sciences of the United States of America* 114 (27): E5406–13.

Minami, Ryunosuke, Miyuki Wakabayashi, Seiko Sugimori, Kiichiro Taniguchi, Akihiko Kokuryo, Takao Imano, Takashi Adachi-Yamada, Naoko Watanabe, and Hideki Nakagoshi. 2012. "The Homeodomain Protein Defective Proventriculus Is Essential for Male Accessory Gland Development to Enhance Fecundity in Drosophila." *PloS One* 7 (3): e32302.

Mueller, J. L., K. Ravi Ram, L. A. McGraw, M. C. Bloch Qazi, E. D. Siggia, A. G. Clark, C. F. Aquadro, and M. F. Wolfner. 2005. "Cross-Species Comparison of Drosophila Male Accessory Gland Protein Genes." *Genetics* 171 (1): 131–43.

Mukherjee, A. S., and W. Beermann. 1965. "Synthesis of Ribonucleic Acid by the X-Chromosomes of Drosophila Melanogaster and the Problem of Dosage Compensation." *Nature* 207 (998): 785–86.

Musser, Jacob M., and Günter P. Wagner. 2015. "Character Trees from Transcriptome Data: Origin and Individuation of Morphological Characters and the so-Called 'Species Signal.'" *Journal of Experimental Zoology. Part B, Molecular and Developmental Evolution* 324 (7): 588–604.

Neubaum, D. M., and M. F. Wolfner. 1999. "Mated Drosophila Melanogaster Females Require a Seminal Fluid Protein, Acp36DE, to Store Sperm Efficiently." *Genetics* 153 (2): 845–57.

Odhiambo, Thomas R., Elizabeth D. Kokwaro, and Lina M. Sequeira. 1983. “Histochemical and Ultrastructural Studies of the Male Accessory Reproductive Glands and Spermatophore of the Tsetse, Glossina Morsitans Morsitans Westwood.” *International Journal of Tropical Insect Science* 4 (3): 227–36.

Overend, Gayle, Yuan Luo, Louise Henderson, Angela E. Douglas, Shireen A. Davies, and Julian A. T. Dow. 2016. “Molecular Mechanism and Functional Significance of Acid Generation in the Drosophila Midgut.” *Scientific Reports* 6 (June): 27242.

Patlar, Bahar, Vivek Jayaswal, José M. Ranz, and Alberto Civetta. 2021. “Nonadaptive Molecular Evolution of Seminal Fluid Proteins in Drosophila.” *Evolution; International Journal of Organic Evolution* 75 (8): 2102–13.

Patro, Rob, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. 2017. “Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression.” *Nature Methods* 14 (4): 417–19.

Peng, Jing, Peder Zipperlen, and Eric Kubli. 2005. “Drosophila Sex-Peptide Stimulates Female Innate Immune System after Mating via the Toll and Imd Pathways.” *Current Biology: CB* 15 (18): 1690–94.

Poels, Jeroen, Tom Van Loy, Hans Peter Vandersmissen, Boris Van Hiel, Sofie Van Soest, Ronald J. Nachman, and Jozef Vanden Broeck. 2010. “Myoinhibiting Peptides Are the Ancestral Ligands of the Promiscuous Drosophila Sex Peptide Receptor.” *Cellular and Molecular Life Sciences: CMLS* 67 (20): 3511–22.

Prince, Elodie, Benjamin Kroeger, Dragan Gligorov, Clive Wilson, Suzanne Eaton, François Karch, Marko Brankatschk, and Robert K. Maeda. 2018. “Rab-Mediated Trafficking in the Secondary Cells of Drosophila Male Accessory Glands and Its Role in Fecundity.” *Traffic* , November.

Ranz, José M., Cristian I. Castillo-Davis, Colin D. Meiklejohn, and Daniel L. Hartl. 2003. “Sex-Dependent Gene Expression and Evolution of the Drosophila Transcriptome.” *Science* 300 (5626): 1742–45.

Ravi Ram, K., and Mariana F. Wolfner. 2007. “Seminal Influences: Drosophila Acps and the Molecular Interplay between Males and Females during Reproduction.” *Integrative and Comparative Biology* 47 (3): 427–45.

———. 2009. “A Network of Interactions among Seminal Proteins Underlies the Long-Term Postmating Response in Drosophila.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (36): 15384–89.

Rexhepaj, Albana, Huanfa Liu, Jing Peng, Yves Choffat, and Eric Kubli. 2003. “The Sex-Peptide DUP99B Is Expressed in the Male Ejaculatory Duct and in the Cardia of Both Sexes.” *European Journal of Biochemistry / FEBS* 270 (21): 4306–14.

Richmond, Rollin C., Karen M. Nielsen, James P. Brady, and Elizabeth M. Snella. 1990. “Physiology, Biochemistry and Molecular Biology of the Est-6 Locus in Drosophila Melanogaster.” In *Ecological and Evolutionary Genetics of Drosophila*, edited by J. S. F. Barker, William T. Starmer, and Ross J. MacIntyre, 273–92. Boston, MA: Springer US.

Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. “Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies.” *Nucleic Acids Research* 43 (7): e47.

Samakovlis, C., P. Kylsten, D. A. Kimbrell, A. Engström, and D. Hultmark. 1991. “The Andropin Gene and Its Product, a Male-Specific Antibacterial Peptide in Drosophila Melanogaster.” *The EMBO Journal* 10 (1): 163–69.

Satija, Rahul, Jeffrey A. Farrell, David Gennert, Alexander F. Schier, and Aviv Regev. 2015. “Spatial Reconstruction of Single-Cell Gene Expression Data.” *Nature Biotechnology* 33 (5): 495–502.

Saudan, Philippe, Klaus Hauck, Matthias Soller, Yves Choffat, Michael Ottiger, Michael Spörri, Zhaobing Ding, et al. 2002. “Ductus Ejaculatorius Peptide 99B (DUP99B), a Novel Drosophila Melanogaster Sex-Peptide Pheromone.” *European Journal of Biochemistry / FEBS* 269 (3): 989–97.

Schully, Sheri Dixon, and Michael E. Hellberg. 2006. “Positive Selection on Nucleotide Substitutions and Indels in Accessory Gland Proteins of the Drosophila Pseudoobscura Subgroup.” *Journal of Molecular Evolution* 62 (6): 793–802.

Sebé-Pedrós, Arnau, Baptiste Saudemont, Elad Chomsky, Flora Plessier, Marie-Pierre Mailhé, Justine Renno, Yann Loe-Mie, et al. 2018. “Cnidarian Cell Type Diversity and Regulation Revealed by Whole-Organism Single-Cell RNA-Seq.” *Cell* 173 (6): 1520–34.e20.

Sepil, Irem, Ben R. Hopkins, Rebecca Dean, Marie-Laëtitia Thézénas, Philip D. Charles, Rebecca Konietzny, Roman Fischer, Benedikt Kessler, and Stuart Wigby. 2018. “Quantitative Proteomics Identification of Seminal Fluid Proteins in Male Drosophila Melanogaster.” *Molecular & Cellular Proteomics: MCP*, October.

Simpson, George Gaylord. 1944. *Tempo and Mode in Evolution*. New York Chichester, West Sussex: Columbia University Press.

Singh, Akanksha, Norene A. Buehner, He Lin, Kaitlyn J. Baranowski, Geoffrey D. Findlay, and Mariana F. Wolfner. 2018. “Long-Term Interaction between Drosophila Sperm and Sex Peptide Is Mediated by Other Seminal Proteins That Bind Only Transiently to Sperm.” *Insect Biochemistry and Molecular Biology* 102 (November): 43–51.

Sitnik, Jessica L., Dragan Gligorov, Robert K. Maeda, François Karch, and Mariana F. Wolfner. 2016. “The Female Post-Mating Response Requires Genes Expressed in the Secondary Cells of the Male Accessory Gland in Drosophila Melanogaster.” *Genetics* 202 (3): 1029–41.

Smith, Nick G. C., and Adam Eyre-Walker. 2002. “Adaptive Protein Evolution in Drosophila.” *Nature* 415 (6875): 1022–24.

Sokal, Robert R., and F. James Rohlf. 2012. *Biometry: The Principles and Practice of Statistics in Biological Research, 4th Ed*. New York, NY: Freeman and Co.

Soller, Matthias, Mary Bownes, and Eric Kubli. 1999. “Control of Oocyte Maturation in Sexually Mature Drosophila Females.” *Developmental Biology* 208 (2): 337–51.

Street, Kelly, Davide Risso, Russell B. Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. 2018. “Slingshot: Cell Lineage and Pseudotime Inference for Single-Cell Transcriptomics.” *BMC Genomics* 19 (1): 477.

Stuart, Tim, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck 3rd, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. 2019. “Comprehensive Integration of Single-Cell Data.” *Cell* 177 (7): 1888–1902.e21.

Styger, D. 1992. “Molekulare Analyse Des Sexpeptidgens Aus Drosophila Melanogaster.” *University of Zurich, Zurich, Switzerland*.

Swanson, W. J., A. G. Clark, H. M. Waldrip-Dail, M. F. Wolfner, and C. F. Aquadro. 2001. “Evolutionary EST Analysis Identifies Rapidly Evolving Male Reproductive Proteins in Drosophila.” *Proceedings of the National Academy of Sciences of the United States of America* 98 (13): 7375–79.

Takemori, Nobuaki, and Masa-Toshi Yamamoto. 2009. “Proteome Mapping of the Drosophila Melanogaster Male Reproductive System.” *Proteomics* 9 (9): 2484–93.

Tosches, Maria Antonietta, Tracy M. Yamawaki, Robert K. Naumann, Ariel A. Jacobi, Georgi Tushev, and Gilles Laurent. 2018. “Evolution of Pallium, Hippocampus, and Cortical Cell Types Revealed by Single-Cell Transcriptomics in Reptiles.” *Science* 360 (6391): 881–88.

Tsaur, S. C., C. T. Ting, and C. I. Wu. 1998. “Positive Selection Driving the Evolution of a Gene of Male Reproduction, Acp26Aa, of Drosophila: II. Divergence versus Polymorphism.” *Molecular Biology and Evolution* 15 (8): 1040–46.

Ulmschneider, Bryne, Bree K. Grillo-Hill, Marimar Benitez, Dinara R. Azimova, Diane L. Barber, and Todd G. Nystul. 2016. “Increased Intracellular pH Is Necessary for Adult Epithelial and Embryonic Stem Cell Differentiation.” *The Journal of Cell Biology* 215 (3): 345–55.

Van den Berge, Koen, Hector Roux de Bézieux, Kelly Street, Wouter Saelens, Robrecht Cannoodt, Yvan Saeys, Sandrine Dudoit, and Lieven Clement. 2020. “Trajectory-Based Differential Expression Analysis for Single-Cell Sequencing Data.” *Nature Communications* 11 (1): 1201.

Wagstaff, Bradley J., and David J. Begun. 2005. “Molecular Population Genetics of Accessory Gland Protein Genes and Testis-Expressed Genes in Drosophila Mojavensis and D. Arizonae.” *Genetics* 171 (3): 1083–1101.

Wang, Jingjing, Huiyu Sun, Mengmeng Jiang, Jiaqi Li, Peijing Zhang, Haide Chen, Yuqing Mei, et al. 2021. “Tracing Cell-Type Evolution by Cross-Species Comparison of Cell Atlases.” *Cell Reports* 34 (9): 108803.

Wang, Liguo, Hyun Jung Park, Surendra Dasari, Shengqin Wang, Jean-Pierre Kocher, and Wei Li. 2013. “CPAT: Coding-Potential Assessment Tool Using an Alignment-Free Logistic Regression Model.” *Nucleic Acids Research* 41 (6): e74.

White, M. J. D. 1977. *Animal Cytology and Evolution*. CUP Archive.

Whittle, Carrie A., and Cassandra G. Extavour. 2019. “Selection Shapes Turnover and Magnitude of Sex-Biased Expression in Drosophila Gonads.” *BMC Evolutionary Biology* 19 (1): 60.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer.

Wigby, Stuart, Nora C. Brown, Sarah E. Allen, Snigdha Misra, Jessica L. Sitnik, Irem Sepil, Andrew G. Clark, and Mariana F. Wolfner. 2020. “The Drosophila Seminal Proteome and Its Role in Postcopulatory Sexual Selection.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 375 (1813): 20200072.

Wolfner, M. F., H. A. Harada, M. J. Bertram, T. J. Stelick, K. W. Kraus, J. M. Kalb, Y. O. Lung, D. M. Neubaum, M. Park, and U. Tram. 1997. “New Genes for Male Accessory Gland Proteins in Drosophila Melanogaster.” *Insect Biochemistry and Molecular Biology* 27 (10): 825–34.

Wong, Alex, Michael C. Turchin, Mariana F. Wolfner, and Charles F. Aquadro. 2008. "Evidence for Positive Selection on Drosophila Melanogaster Seminal Fluid Protease Homologs." *Molecular Biology and Evolution* 25 (3): 497–506.

Xue, Lei, and Markus Noll. 2000. "Drosophila Female Sexual Behavior Induced by Sterile Males Showing Copulation Complementation." *Proceedings of the National Academy of Sciences of the United States of America* 97 (7): 3272–75.

Yanai, Itai, Hila Benjamin, Michael Shmoish, Vered Chalifa-Caspi, Maxim Shklar, Ron Ophir, Arren Bar-Even, et al. 2005. "Genome-Wide Midrange Transcription Profiles Reveal Expression Level Relationships in Human Tissue Specification." *Bioinformatics* 21 (5): 650–59.

Yapici, Nilay, Young-Joon Kim, Carlos Ribeiro, and Barry J. Dickson. 2008. "A Receptor That Mediates the Post-Mating Switch in Drosophila Reproductive Behaviour." *Nature* 451 (7174): 33–37.

Chapter II: Regulatory basis of gene expression evolution in the Drosophila accessory gland

Alex Majane, Julie Cridland, and David Begun

ABSTRACT

Gene expression is an important phenotype that evolves and leads to divergence among species. The genetic basis of expression may include *cis*-regulatory loci and *trans*-acting factors. Additionally, incompatibilities between regulatory factors among species may give rise to hybrid sterility. Studies of allele-specific expression with interspecific hybrids have given major insights into gene-regulatory evolution and hybrid incompatibilities. However, properties of regulatory differences vary widely among species, sex, and tissues. Tissue-specific allele-specific expression studies may improve our understanding of how regulatory evolution and accrual of hybrid incompatibilities proceeds in different biological contexts. In this study we use a hybrid between sister species, *Drosophila melanogaster* and *D. simulans,* to characterize gene regulatory evolution and hybrid misexpression in a somatic male sex organ, the accessory gland. The accessory gland produces seminal fluid proteins (Sfps), a class of proteins involved in sexual conflict with extremely rapid rates of protein sequence evolution. We find that *trans* differences are relatively more abundant than *cis* in this organ, in contrast to most of the interspecific hybrid literature. However, large-effect size *trans* differences are rare. Sfps and accessory gland- biased genes have significantly elevated levels of expression divergence and tend to be regulated through both *cis* and *trans* divergence. We find limited misexpression in this organ compared to other Drosophila studies. As with previous studies, male-biased genes are overrepresented among misexpressed genes and are much more likely to be underexpressed. Finally, we integrate ATAC-Seq data to show chromatin accessibility is correlated with expression differences among species and hybrid allele-specific expression. In summary, this work identifies unique regulatory evolution and hybrid misexpression properties in the accessory gland, contributing to our understanding of this organ's evolution and suggesting a general importance of tissue-specific allele-specific expression studies.

INTRODUCTION

Gene expression is a key phenotype on which stabilizing, directional, or diversifying selection may act. The gene regulatory variants that are targeted by selection and must ultimately explain within and between species expression variation can be broadly classified into *cis*-acting components, such as promoters and enhancers, and *trans*-acting components, such as transcription factors (Rabinow and Dickinson 1981; Dickinson, Rowan, and Brennan 1984; Patricia J. Wittkopp, Haerum, and Clark 2004; Gibson et al. 2004; P. J. Wittkopp 2005; Ronald et al. 2005; Ronald and Akey 2007). Because the genetic control of gene expression can involve several sites spanning both cis- and trans-acting factors, selection could plausibly have many potential substrates on which to act. Thus, understanding the relative importance of these factors in regulatory evolution is critical for achieving a comprehensive view of expression evolution.

A commonly used method for detecting and estimating magnitudes of *cis* and *trans* effects is measuring allele-specific expression (ASE) in hybrids and their parents. ASE can be used to classify genes into regulatory types based on the presence and directionality of *cis* and *trans* components (McManus et al. 2010). It has been broadly applied to both intraspecific and interspecific hybrids to study the genetics of expression variation within species and divergence between species. The comparison of within- and between species regulatory genetics informs our understanding of evolutionary mechanisms because the concordance or discordance of phenomena on these two timescales can narrow the range of evolutionary explanations for the variation. A major conclusion from accumulated ASE research is that intraspecific gene expression evolution in animals is mediated predominantly through *trans* effects, while interspecific evolution proceeds predominantly via *cis* effects (reviewed in Signor and Nuzhdin 2018; Hill, Vande Zande, and Wittkopp 2021). It is thought that *trans* regulation has a broader mutational target but generally deleterious pleiotropic effects, underlying this general observation (P. J. Wittkopp 2005; Gruber et al. 2012; Lemos et al. 2008). However, this pattern is not always observed. For example, Sánchez-Ramírez et al. (2021) found *cis*-regulatory divergence was more frequent than *trans* in male *Caenorhabditis* but not in females, and some studies of flies have found relatively more *trans* effects between species (McManus et al. 2010; Coolon et al. 2014). Deeper investigation into regulatory evolution across sexes, tissues, cell types, or environmental conditions might reveal more nuanced patterns than expected based on whole-organism, single-condition studies.

There is no reason to expect that the genetics of regulatory variation will be identical across tissues in multicellular organisms, as the cell and developmental biology, as well as the influence of mutation and selection on expression phenotypes may vary across cell types, tissues, and organs. Indeed, empirical evidence supports the view that tissues and cell types exhibit varying rates of expression divergence (Gu and Su 2007; Brawand et al. 2011; Romero, Ruvinsky, and Gilad 2012; Kryuchkova-Mostacci and Robinson-Rechavi 2015; Liang et al. 2018; J. Chen et al. 2019; Pal, Oliver, and Przytycka 2021). Intraspecific studies of mouse (Babak et al. 2015; Andergassen et al. 2017; St Pierre et al. 2022), humans (Babak et al. 2015; Leung et al. 2015; Castel et al. 2020), flycatchers (Wang, Uebbing, and Ellegren 2017), and Drosophila (Combs et al. 2018) have revealed tissue-specific variance in *cis*-effects; genes may exhibit ASE in some tissues but not others, and the total number and magnitude of *cis*-effects also varies across tissues. These studies did not identify *trans*-effects, however, limiting the insight we have into how regulatory evolution varies among tissues.

There is, however, little literature investigating the genetics of interspecific expression divergence at the level of organ or tissue. Indeed, much of the influential Drosophila literature on this topic analyzes parental and hybrid expression in whole animals (Patricia J. Wittkopp, Haerum, and Clark 2004; Landry et al. 2005; Patricia J. Wittkopp, Haerum, and Clark 2008; McManus et al. 2010; Wei, Clark, and Barbash 2014; Coolon et al. 2014) or heads (Graze et al. 2009); while this literature has provided valuable generalities about regulatory evolution, the possibility remains of henceforth undiscovered genetically and evolutionary important heterogeneity across organs. Additional studies used ASE to investigate interspecific divergence in Drosophila testes (Haerty and Singh 2006; Lu et al. 2010; Llopart 2012; Brill et al. 2016; Banho et al. 2021). While this work has shed light on the regulatory basis of hybrid incompatibilities, given the propensity for hybrid dysgenesis in the testis, conclusions from studies of this organ may not apply to gene regulatory evolution more broadly. Few studies of interspecific ASE in animals have investigated single somatic tissues and characterized both *cis* and *trans* regulatory divergence (Goncalves et al. 2012; Davidson and Balakrishnan 2016).

Our goal here is to contribute to the literature on the genetics of interspecific regulatory divergence using the accessory gland (AG) of Drosophila as a model. Seminal fluid proteins (Sfps) are secreted by the accessory glands, ejaculatory duct, and ejaculatory bulb and transferred to females along with sperm during mating and are essential for fertilization, similarly to the seminal fluid of the mammalian prostate (reviewed in Poiani 2006; Wilson et al. 2017). The genus Drosophila has a polyandrous mating system featuring competition between males for matings and sperm competition (Boorman and Parker 1976; Imhof et al. 1998; Clark, Begun, and Prout 1999). Sfps of Drosophila and many other insects induce a range of physiological and behavioral changes in females comprising the post-mating response (PMR; reviewed in Ravi Ram and Wolfner 2007; Avila et al. 2011; Sirot et al. 2014; Wigby et al. 2020), including increased egg laying, facilitation of sperm storage, immune system responses, increased feeding rates, increased activity level and decreased sleep, and decreased receptivity to remating. PMR phenotypes evolve in response to sperm competition and sexual conflict (Hollis et al. 2019). Sfps play a key role in mediating sperm competition; genetic variation in Sfp loci is linked to competitiveness (Clark et al. 1995; Chapman et al. 2000; Fiumera, Dumont, and Clark 2005), and males respond to perceived level of competition through differential allocation of Sfps to the ejaculate (Sirot, Wolfner, and Wigby 2011; Hopkins et al. 2019).

As expected, given their possible roles in sexual conflict (Swanson and Vacquier 2002; Haerty et al. 2007), Sfp protein sequences evolve at an especially rapid rate, often under the influence of recurrent directional selection (Tsaur, Ting, and Wu 1998; Aguadé 1999; Begun et al. 2000; Holloway and Begun 2004; Begun et al. 2006-3; Schully and Hellberg 2006; Wong et al. 2008; Majane, Cridland, and Begun 2022), though relaxed selective constraint may also contribute to their rapid divergence (Dapper and Wade 2020; Patlar et al. 2021). Sfp genes have rapid rates of turnover (Wagstaff and Begun 2005; Mueller et al. 2005), exhibiting gene gain and loss even among closely related species (Begun and Lindfors 2005).

While there is a long history of work on sequence evolution and turnover in Sfps (reviewed in Hurtado et al. 2022), less is known about gene expression evolution among Sfps or in the accessory gland more broadly. RNAi knockdowns demonstrate that PMR phenotypes are sensitive to expression level of many Sfps (Ravi Ram and Wolfner 2007; Patlar and Civetta 2022), suggesting that Sfp expression is a relevant phenotype on which selection could act via male reproductive success. Consistent with the observation that male-biased genes tend to have higher levels of interspecific expression divergence (Meiklejohn et al. 2003; Parisi et al. 2004; Ellegren and Parsch 2007; Brawand et al. 2011; Graveley et al. 2011; Assis, Zhou, and Bachtrog 2012; Whittle and Extavour 2019; Pal, Oliver, and Przytycka 2021), we recently reported rapid expression divergence as well as the evolution of novel genes and expression phenotype in the accessory gland (Cridland et al. 2020). However, Cridland et al. did not focus on the general properties of accessory gland transcriptome divergence, did not compare Sfp expression to expression divergence of other gene classes expressed in the accessory gland, and did not address the genetics of accessory gland expression divergence between species. To our knowledge there has been no work on regulatory evolution in the accessory gland.

While hybrid males derived from crosses between *D. melanogaster* and its sibling species, *D. simulans* are generally completely sterile or inviable (Sturtevant 1920), with severely atrophied or absent testes, a previous report noted that hybrids between male *D. melanogaster* and female *D. simulans* have morphologically normal accessory glands that produce seminal fluid that can induce the PMR in females (Stumm-Zollinger and Chen 1988). In this study, we use ASE analyses derived from measures of gene expression in accessory glands of *D. melanogaster, D. simulans*, and their hybrids to estimate *cis* and *trans* expression effects in the accessory gland and ejaculatory duct. We investigate regulatory evolution of the gland and uncover unique evolutionary properties of Sfps and other accessory-gland biased genes. We also quantify inheritance of expression phenotypes and characterize misexpressed genes that may be related to hybrid incompatibilities. We tie regulatory evolution to divergence in upstream noncoding regions and protein sequence evolution. Finally, we integrate ATAC-Seq data to link changes in chromatin state with expression divergence.

METHODS

RNA-Seq

We performed RNA-Seq on each of three samples: *D. melanogaster* (Raleigh 517, Mackay et al. 2012), *D. simulans* (w501), and a *D. melanogaster* X *D. simulans* interspecific hybrid, with three replicates per sample. We raised all Drosophila stocks on cornmeal-molasses-agar medium at 25C and 60% relative humidity, on a 12:12 light/dark cycle. When we crossed the parents to produce the interspecific hybrid, we used a ratio of five female *D. simluans* to 25 male *D. melanogaster*, as females are unlikely to mate with heterospecific males. We collected virgin male adult flies and grouped them together with five males per vial. We aged flies for two days before dissection. On the day of the experiment, we anesthetized flies with CO2, dissected their accessory glands and anterior ejaculatory duct in cold 1X PBS, and collected the tissue in TRIzol (Thermo-Fisher 15596026) on ice. We confirmed that the hybrid organs appeared morphologically normal with seminal fluid production. After we collected tissue from 20 males we flash-froze the TRIzol tubes containing tissue in liquid nitrogen and stored the material at -80C. We extracted RNA using the standard TRIzol protocol followed by DNAse digestion (Invitrogen AM1907) and clean-up with AMPure beads (Beckman-Coulter A63881). Novogene performed RNA-Seq library preparation and Illumina sequencing (paired-end 150bp).

Assigning species-of-origin

To determine the species-of-origin for each allele in the hybrid, we used an alignment-based approach relying on differences in the number of mismatches between the reads and each reference. We aligned each sample to each of two references, *D. melanogaster* (custom reference based on Flybase release 6.04 with SNPs included for the strain used here, Raleigh 517), and *D. simulans* (Princeton University, release 3.0), using HiSat2 and requiring a MAPQ score ≥ 30. We sorted reads using custom perl, bash and R scripts (github.com/alexmajane/hybridASE). We sorted reads into groups that mapped to one reference uniquely or mapped to both references. For read pairs where at least one mate aligned uniquely, we assigned both reads to that species. For the remaining reads, we analyzed the number of nucleotide mismatches algorithmically to assign species-of-origin. Reads that aligned to one species with at least six fewer mismatches were assigned to the species with fewer mismatches. We also subjected our *D. melanogaster* and *D. simulans* samples to the same workflow, to A) account for artifactual effects of the procedure on expression analysis, B) establish a ground-truth false-positive rate for species-assignment, and C) identify problematic gene regions with high rates of erroneous species-assignment. To address the latter point, we calculated the fold change of counts (see below) per gene with and without inclusion of misassigned parental reads. If the absolute value of $\log_2$(fold change) was greater than 0.025 in either species, we removed that gene from our downstream analyses, a total of 382 genes.

Quantification of gene expression

After we assigned hybrid reads to species-of-origin and filtered out unassignable reads in hybrid and parental samples, we quantified gene expression using Salmon (Patro et al. 2017). We used Salmon's alignment-free approach because it can account for differences in transcript length and GC content across samples. Since length and GC content vary between species due to evolution or annotation differences, we think it is important to account for this in comparative RNA-Seq studies. Following quantification, we used tximport (Soneson, Love, and Robinson 2015) to estimate counts per gene. We limited our analysis to 1-to-1 orthologs between *D. melanogaster* and *D. simulans*, based on the FlyBase annotation (02/2020 release) with additional Sfp orthologs (Majane, Cridland, and Begun 2022).

Differential expression

We used estimated counts from tximport as the basis of all downstream analysis. We performed independent analyses of autosomal and X-linked genes because of their different inheritance in the hybrid. We used DESeq2 (Love, Huber, and Anders 2014) to normalize count data with the median-of-ratios method (Anders and Huber 2010), identify DE genes using Wald tests, and estimate moderated log-fold changes with the *apeglm* model (Zhu, Ibrahim, and Love 2019).

Regulatory and inheritance classifications

We refer to counts from *D. melanogaster* as P_{mel}, *D. simulans* as P_{sim}, and allele-specific hybrid counts as $F1_{mel}$ and $F1_{sim}$. We calculated total F1 expression ($F1_{total}$) as $F1_{mel}$ + $F1_{sim}$. We classified genes into regulatory and inheritance groups using the algorithm outlined in McManus et al. (2010). For this purpose, we define DE as a significant Wald test (Bonferonni adjusted $p < 0.05$) and make comparisons between A) parental expression: P_{mel} and P_{sim} (**P**), B) ASE within the hybrid: $F1_{mel}$ and $F1_{sim}$ (**H**), and C) between parent expression and expression of parental-specific alleles in the hybrid: P_{mel} and $F1_{mel}$ or P_{sim} and $F1_{sim}$ (**T**). We define regulatory classes as follows:

A. conserved: no significant **P**, **H**, or **T**.
B. *cis*: significant **P** and **H**, no significant **T**.
C. *trans*: significant **P** and **T**, no significant **H**.
D. *cis* + *trans*: significant **P**, **H**, and **T**, same directionality between the parental contrast and hybrid ASE. *cis* and *trans* effects favor expression of the same allele.
E. *cis* by *trans*: significant **P**, **H**, and **T**, opposite directionality between the parental contrast and hybrid ASE. *cis* and *trans* effects favor expression of different alleles.

F. compensatory: significant **H** and **T**, but no significant **P**. *cis* and *trans* effects complement one another such that there has been no evolved expression difference among species.

G. ambiguous: all other patterns of expression without classical interpretations, such as no **P** or **H**, but significant **T** (evidence of *trans* effects that appear only in the hybrid).

We classified genes by inheritance comparing overall hybrid expression ($F1_{total}$) to parental expression. Note that here we refer to *dominance* only in the phenotypic sense; we do not make inferences about the genetic basis of inheritance of expression phenotypes in the hybrid. We define the following inheritance classes:

A. conserved: no DE between $F1_{total}$ and P_{mel} or P_{sim}.

B. additive: DE between $F1_{total}$ and both P_{mel} and P_{sim}. $F1_{total}$ has an intermediate expression level.

C. *mel* dominant: no DE between $F1_{total}$ and P_{mel}. DE between $F1_{total}$ and P_{sim}.

D. *sim* dominant: no DE between $F1_{total}$ and P_{sim}. DE between $F1_{total}$ and P_{mel}.

E. overdominant: DE between $F1_{total}$ and both P_{mel} and P_{sim}. $F1_{total}$ is expressed at a higher level than both parents.

F. underdominant: DE between $F1_{total}$ and both P_{mel} and P_{sim}. $F1_{total}$ is expressed at a lower level than both parents.

SFPs and AG-biased genes

We define sets of seminal fluid proteins (Sfps) and accessory gland (AG)-biased genes to investigate patterns of DE, regulation, and inheritance in these gene classes. We refer to Wigby et al. (2020) for annotation of Sfps. There are 208 Sfps expressed in our dataset. To annotate AG-biased genes we obtained expression data from FlyAtlas2 (Leader et al. 2018), and calculated the index of tissue specificity, τ (Yanai et al. 2005), for all genes. We define AG-biased genes as those that are more highly expressed in the accessory gland than all other tissues and have $\tau \geq$

0.8. There are 378 AG-biased genes expressed in our dataset, 238 of which are non-Sfps, which we use for further analysis.

GO analysis

We performed gene ontology (GO) enrichment analyses with Enrichr (Kuleshov et al. 2016). We defined background gene sets as all genes expressed in the data ($\log_2$(counts) ≥ 1). We used the Bioconductor *D. melanogaster* annotation (Carlson 2021) and queried terms from all three sub-ontologies (Biological Process, Molecular Function, and Cellular Component).

Upstream sequence analysis

We obtained sequences spanning 1000 bp, 750 bp, and 500 bp upstream of each *D. melanogaster* TSS (Flybase annotation 6.41) and removed any overlapping coding sequence. Then we used BLAST (gap open penalty: 2; gap extension penalty: 1) to find orthologous sequences in the *D. simulans* genome (Princeton University, release 3.0). We discarded sequences with more than one BLAST hit or overlapping alignments. Next we aligned *D. melanogaster* and *D. simulans* sequences with MUSCLE (Edgar 2004). We estimated the Kimura 2-parameter nucleotide substitution rate (Kimura 1980) for each upstream region using EMBOSS distmat (Rice, Longden, and Bleasby 2000). We also trimmed the 500 bp sequences to shorter lengths of 100 bp, 200 bp, and 300 bp upstream of the TSS and repeated estimation of substitution rate.

Protein sequence evolutionary analysis

We analyzed protein sequence evolution by estimating divergence between species and rates of synonymous (dS) and nonsynonymous substitutions (dN). We obtained the longest open reading frame per gene from FlyBase annotations (*D. melanogaster* 6.41, *D. simulans* 2.02), translated nucleotide sequences with EMBOSS transeq (Rice, Longden, and Bleasby 2000), and aligned amino acid sequences with MUSCLE (Edgar 2004). We back-translated to codon alignments with gaps removed using PAL2NAL (Suyama, Torrents, and Bork 2006). For each gene we estimated dN and dS using Goldman and Yang's maximum likelihood codon-based substitution model (codeml; Goldman and Yang 1994). We performed this analysis with PAML (Yang 1997) implemented in BioPython (Cock et al. 2009).

We additionally analyzed adaptive protein evolution in *D. melanogaster* using available population genomics data (Fraïsse, Puixeu Sala, and Vicoso 2019) with pre-computed McDonald-Krietman tests (McDonald and Kreitman 1991), as in our previous work (Majane, Cridland, and Begun 2022- see supplement for detail). We used the summary statistic α, which estimates the proportion of amino acid substitutions due to positive selection. A positive value suggests directional selection on a given gene, with larger values suggesting a greater proportion of adaptive substitutions.

Chromatin state integration

We analyzed chromatin state in *D. melanogaster* and *D. simulans* with ATAC-Seq data from Blair et al. (unpublished), who inferred ATAC-Seq peaks called from three replicates each in the same strains that we used. We used reciprocal BLAST (gap open penalty: 2; gap extension penalty: 1) on peak sequences between species and verified orthology of 1-to-1 best hits using synteny (nearest upstream and downstream annotated exons). We defined conserved peaks as those with a single reciprocal hit in each species and shared synteny. We inferred synteny from the nearest upstream and downstream exons. We defined orphan peaks as those with no BLAST hits. We additionally BLASTed orphan peak sequences to the reciprocal species' genome and filtered out peaks with A) no hits to the genome, B) a hit within 100 bp of any annotated peak, C) multiple hits. It is important to compare truly orthologous regions when we quantify peak accessibility by counting reads. We defined orthologous regions of orphan peaks in reciprocal species as the span of each BLAST hit. We re-annotated conserved peaks by reciprocal BLAST of peaks to each other species' genome and extended the boundaries of each peak to the span of BLAST hits intersecting the original annotation.

To quantify chromatin accessibility in each peak, we counted the number of aligned ATAC-Seq reads intersecting each peak with HTSeq (Anders, Pyl, and Huber 2015). We analyzed count data using DESeq2 similarly to RNA-Seq analysis. We then annotated peaks with the nearest TSS. For each gene, we used our Salmon quantification to select the transcript with highest expression in our sample and chose its annotated TSS. We then selected 1-to-1 peak-to-TSS pairs with the closest or overlapping peak per TSS, removing any duplicate matches from further analysis. We also filtered the data to include only pairs where each orthologous peak (or region in the case of orphans) matched the same gene in both species.

RESULTS

Alignment and identification of allele-specific reads

We performed RNA-Seq on *D. melanogaster*, *D. simulans*, and an interspecific hybrid (P_{mel}, P_{sim}, and F1). We obtained between 25.4 and 30.8M reads per RNA-Seq sample. We aligned each sample to each of two references, *D. melanogaster* and *D. simulans.* Each species aligns at a rate of 94-97% to the matching reference (Table S5). A slightly better alignment rate for P_{sim} is expected, since the *D. simulans* strain we used in our experiment matches the reference strain, while our *D. melanogaster* strain differs from the reference. F1 aligned to the *D. melanogaster* reference at a rate of 61-62%, and to the *D. simulans* reference at a rate of 69-71%. A higher rate of alignment to *D. simulans* is expected given that the hybrid inherits a *D. simulans* X chromosome.

Among F1 reads, ~20% aligned uniquely to *D. melanogaster*, 33-34% aligned uniquely to *D. simulans*, and 46-47% aligned to both references (Table S6). We used a mismatch-based approach to compare alignments and assign non-uniquely aligned reads to each species (Table S7). We were able to assign ~25% of non-uniquely aligned reads (~12% of total aligned reads) to *D. melanogaster*, and 35-36% of non-uniquely aligned reads (16-17% of total aligned reads) to *D. simulans*, leaving 18-19% of total aligned reads of indeterminable origin and removed from our analysis.

We gave parental samples the same treatment as the F1, to A) account for artifactual effects of the procedure on expression analysis, B) establish a ground-truth false-positive rate for species-assignment, and C) identify problematic gene regions with high rates of erroneous species-assignment. ~1.2% of P_{mel} reads uniquely aligned to the *D. simulans* genome, and 0.6-0.7% of P_{sim} reads uniquely aligned to the *D. melanogaster* genome (Table S6). Among non-uniquely aligning parental reads, our assignment algorithm assigned 0.16-0.17% of P_{mel} reads to *D. simulans*, and 0.031-0.035% of P_{sim} reads to *D. melanogaster* (Table S7). We used incorrectly assigned parental reads to identify gene regions with elevated levels of misassignment. Most misassigned reads do not overlap genes (Table S8). We identified 382 genes where misassigned reads significantly impacted estimates of gene expression (Supplemental Data), which we removed from downstream analysis.

We used Salmon (Patro et al. 2017) to quantify gene expression on reads that passed our filtering and species-assignment. For F1 samples we used read whitelists to only quantify those reads with a confident species call. Salmon has advantages over alignment-based approaches for interspecific differential gene expression analysis because it accounts for transcript length and GC content differences, which may vary among orthologs across species due to evolution and/or gene annotation. Salmon mapping rates are as follows: 89-90% of F1 reads of *D. melanogaster* origin (hereafter $F1_{mel}$), 84-85% of F1 reads of *D. simulans* origin (hereafter $F1_{sim}$), ~94% of P_{mel} reads, and 86-87% of P_{sim} reads.

Transcriptome-wide view

We converted quantified expression values from Salmon to estimated counts with tximport (Soneson, Love, and Robinson 2015) and used these counts as the basis of all downstream analyses. Given the different inheritance patterns of the autosomes and X chromosome in the hybrid, we give these gene sets independent treatment for all analyses. Because the hybrid has a single *D. simulans* X chromosome, there is no ASE analysis of X-linked genes. We will refer to total F1 expression (sum of both alleles) as $F1_{total}$, and ASE measures by $F1_{mel}$ and $F1_{sim}$.

Principal component analysis of transcriptome-wide gene expression shows tightly shared variance of within-sample replicates (Figure 6). Among autosomal genes, P_{mel} groups with $F1_{mel}$ and P_{sim} groups with $F1_{sim}$, while $F1_{total}$ groups away from these two clusters. PC1 appears to explain differences between F1 samples and $P_{mel/sim}$, with allele-specific samples lying between $F1_{total}$ and $P_{mel/sim}$, though much closer to $P_{mel/sim}$. PC2 appears to explain differences in expression between *D. melanogaster* and *D. simulans*, with $F1_{total}$ lying roughly halfway between species-specific expression. Among X-linked genes, both PC1 and PC2 appear to explain differences

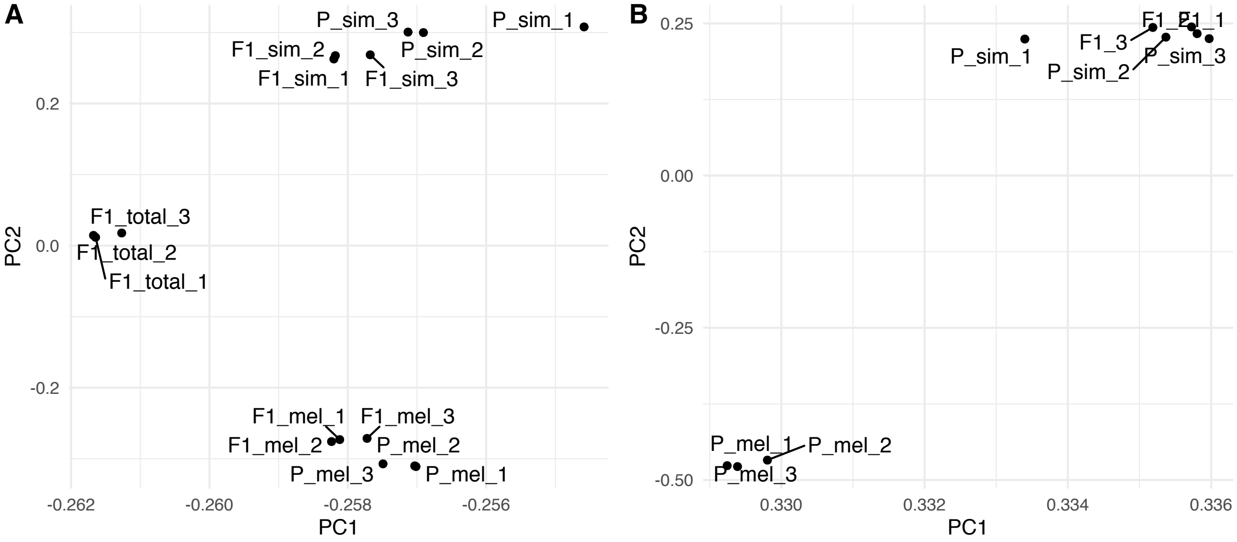

Figure 6. PCA of transcriptome-wide expression (log-transformed counts) showing the first two PCs. A) Autosomal-linked genes expression. Parental expression is similar to parent-specific alleles in the hybrid, with total hybrid expression clustering away from parental and allele-specific expression. B) X-linked gene expression. *D. simulans* expression is similar to hybrid expression.

between *D. melanogaster* and *D. simulans* allele-derived expression, and X-linked F1 expression is tightly grouped with P_{sim} expression. We did not find appreciable differences in distributions of expression variance between samples or allele-specific expression (Fig. S8), aside from a very modest elevation in standard deviation among X-linked genes relative to autosomal genes.

We used correlations of average gene expression to quantify transcriptome-wide divergence. Among autosomal genes, P_{mel} and P_{sim} have a Pearson correlation coefficient r = 0.934 (Figure 7A). $F1_{total}$ expression is more like each parent; expression profiles between $F1_{total}$ and P_{mel} have an r = 0.965 (Figure 7B), while $F1_{total}$ and P_{mel} have an r = 0.967 (Figure 7C). Correlations between allele-specific expression and the same-species parent are strongest: $F1_{mel}$ and P_{mel} have an r = 0.982 (Figure 7D); $F1_{sim}$ and P_{sim} have an r = 0.980 (Figure 7E). Allele-specific expression within hybrids is somewhat more correlated than parental expression profiles are to one another; $F1_{mel}$ and $F1_{sim}$ have an r = 0.945 (Figure 7F). Among X-linked genes, P_{mel} and P_{sim} are similarly correlated as with autosomal genes (r = 0.930, Figure 7G). X-linked gene expression in hybrids is overall very similar to *D. simulans*: $F1_{total}$ and P_{sim} have an r = 0.983 (Figure 7H). Comparing $F1_{total}$ and P_{mel} is very similar to the parental X-linked expression contrast with an r = 0.929 (Figure 7I). Overall, the data suggest widespread additivity for autosomal genes, and *D. simulans*-like expression on the hybrid X chromosome. The strong similarity in expression between the hybrid and the parents, and between hybrid ASE and parents, suggests that the accessory glands of this hybrid are not subject to particularly widespread misexpression.

Differential gene expression

In our analyses of differential gene expression (DE), we define DE in two ways: A) genes with a significant difference in normalized counts (Wald test, adjusted $p < 0.01$), and B) genes with an

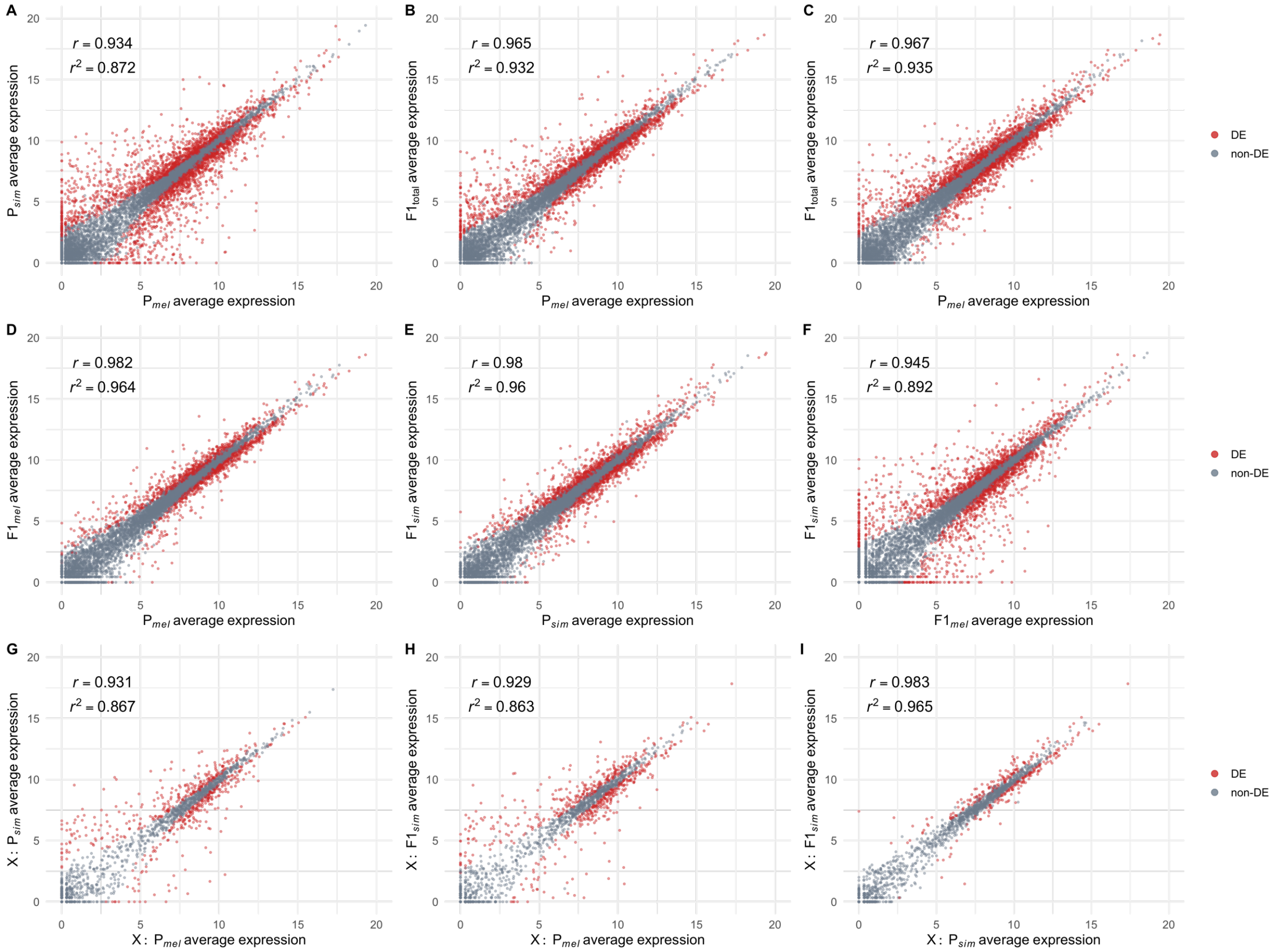

Figure 7. Correlations of average gene expression across the transcriptome. r refers to the Pearson correlation coefficient, r^2 is the coefficient of determination in a simple linear model. (A-C) Comparisons of parental expression and total hybrid expression. (D-F) Comparisons of ASE. (G-I) Comparisons of X-linked gene expression.

adjusted p value < 0.01 and an absolute value of moderated $\log_2$(fold change) > 1. This second class of genes comprises larger effect-size DE which are potentially more biologically relevant.

Without a $\log_2$(fold change) cutoff, we find that 33% of 9,223 total expressed genes are DE between P_{mel} and P_{sim} (Table 2). Fewer are DE in comparison to the hybrid: 25% of expressed genes are DE between $F1_{total}$ and P_{mel}, and 27% between $F1_{total}$ and P_{sim}. Requiring a $\log_2$(fold change) > 1 reduces the number of DE genes, and there are relatively fewer DE genes among contrasts with the hybrid: 17% of genes are DE between P_{mel} and P_{sim}, while 10% are DE between $F1_{total}$ and P_{mel}, and 10% between $F1_{total}$ and P_{sim}.

DE is less frequent between hybrid allele-specific expression and parents relative to total hybrid expression (Table 2): 17% of genes are DE between $F1_{mel}$ and P_{mel} with no $\log_2$(fold change) cutoff, and 19% of genes are DE between $F1_{sim}$ and P_{sim}. DE with a $\log_2$(fold change) > 1 is infrequent; just 5% of genes are DE between $F1_{mel}$ and P_{mel}, and 6% between $F1_{sim}$ and P_{sim}. The percent of genes DE between $F1_{mel}$ and $F1_{sim}$ (allele-specific expression within the hybrid indicative of *cis*-regulatory effects) is 23% without a $\log_2$(fold change) cutoff, and 12% with a cutoff > 1.

Among expressed X-linked genes, rates of DE between P_{mel} and P_{sim} are similar to autosomal genes: 33% with no $\log_2$(fold change) cutoff, and 17% with a cutoff > 1. DE is more common between F1 and P_{mel} with 37% of X-linked genes DE with no cutoff and 18% with a cutoff. 18% of genes are DE between F1 and P_{sim} with no cutoff, indicative of *trans*-regulatory effects on expression given that the hybrid has a *D. simulans* X chromosome. Requiring $\log_2$(fold change) > 1 dramatically reduces the number of genes DE between F1 and P_{sim} to just 4%, however, suggesting that *trans*-regulatory differences are unlikely to lead to large shifts in X-linked gene expression.

Table 2. Differentially expressed genes. In the first three columns, DE is defined as a significant Wald test (Bonferroni adjusted $p < 0.01$). In the last three columns, DE additionally requires a $\log_2$(fold change) value greater than 1.

chromosome	contrast	DE	non-DE	fraction DE	DE: $\log_2$(FC) > 1	non-DE: $\log_2$(FC) > 1	fraction DE: $\log_2$(FC) > 1
autosomes	P_{mel} P_{sim}	3005	6218	0.326	1608	7615	0.174
autosomes	$F1_{total}$ P_{mel}	2290	6933	0.248	883	8340	0.096
autosomes	$F1_{total}$ P_{sim}	2459	6764	0.267	956	8267	0.104
autosomes	$F1_{mel}$ P_{mel}	1585	7638	0.172	430	8793	0.047
autosomes	$F1_{sim}$ P_{sim}	1790	7433	0.194	561	8662	0.061
autosomes	$F1_{mel}$ $F1_{sim}$	2132	7091	0.231	1103	8120	0.120
X	P_{mel} P_{sim}	564	1135	0.332	284	1415	0.167
X	$F1_{sim}$ P_{mel}	631	1068	0.371	303	1396	0.178
X	$F1_{sim}$ P_{mel}	310	1389	0.182	66	1633	0.039

DE among seminal fluid proteins (Sfps) and accessory gland-biased genes

Sfps are known to have very high rates of amino acid substitutions as well as gene turnover between species. Given the observation that rates of protein evolution are often correlated with gene expression evolution, we asked whether expressed Sfps (208 total) were more likely than non-Sfps to be DE between *D. melanogaster* and *D. simulans* (P_{mel} vs P_{sim}) and between hybrid alleles (ASE: $F1_{mel}$ vs $F1_{sim}$). Indeed, 59% of Sfps are DE between *D. melanogaster* and *D. simulans*, compared to 32% of non-Sfps (G test, $p < 0.001$), and 51% of Sfps are DE between hybrid alleles (ASE), compared to 23% of non-Sfps (G test, $p < 0.001$, Table S9). A greater proportion of Sfps are also DE with a $\log_2$(fold change) cutoff > 1: 28% of Sfps are DE between *D. melanogaster* and *D. simulans*, compared to 17% of non-Sfps, and 22% of Sfps are DE between hybrid alleles (ASE), compared to 12% of non-Sfps (Table S10). However, Sfps are a highly expressed class of genes (Fig. S9A-C), making them more likely to be DE. The median $\log_2$(counts) of Sfps is 8.46, while the median of non-Sfps is 4.31. To account for the effect of expression level on the likelihood of DE, we used a multiple logistic regression with average expression and Sfp status as independent variables and DE as the dependent variable (DE ~ $\log_2$(counts) + Sfp). In the parental contrast, average expression significantly predicts DE between P_{mel} and P_{sim} ($\beta = 0.22 \pm 0.01$, $p < 0.001$), however Sfp status does not predict DE ($\beta = -0.04 \pm 0.16$, $p = 0.792$). Considering ASE, average expression significantly predicts DE between $F1_{mel}$ and $F1_{sim}$ ($\beta = 0.21 \pm 0.007$, $p < 0.001$), but Sfp status does not predict DE ($\beta = 0.09 \pm 0.16$, $p = 0.554$). Therefore, we conclude that in our experiment, Sfps are not any more likely to be DE than non-Sfps when accounting for expression level.

If we consider genes to be DE with a $\log_2$(fold change) cutoff > 1 however, we observe a strikingly different result. Highly expressed genes are not much more likely to have large effect-size DE; average expression level has a weak relationship with DE between P_{mel} and P_{sim} ($\beta = 0.04 \pm 0.01$, $p < 0.001$). Sfp status does predict large-effect size DE ($\beta = 0.45 \pm 0.16$, $p = 0.006$). Average expression does not have a relationship with DE between $F1_{mel}$ and $F1_{sim}$ ($\beta = 0.01 \pm$

0.01, p = 0.08), and Sfp status predicts DE (β = 0.53 ± 0.17, p = 0.002). We therefore conclude that Sfps are significantly enriched for large effect size DE events compared to non-Sfps.

In addition to Sfps, other genes characteristic of the accessory gland can be defined with the index of tissue specificity, τ (Yanai et al. 2005). Here we define AG-biased genes as non-Sfps with $\tau > 0.8$, a set of 238 genes. 31.1% of non-biased genes are DE between P_{mel} - P_{sim}, and 21.5% are DE between hybrid alleles. AG-biased genes have an elevated level of DE; 63% are DE between P_{mel} - P_{sim}, and 57% are DE between hybrid alleles (Table S9). AG-biased genes are more highly expressed than non-AG-biased genes, but not as highly expressed as Sfps (Fig. S9D-F). To ask whether AG-biased genes are more likely to be DE, we used a multiple logistic regression on non-Sfps with average expression and AG-bias as independent variables and DE as the dependent variable (DE ~ $\log_2$(counts) + AG-bias). In the parental contrast, average expression significantly predicts DE between P_{mel} and P_{sim} (β = 0.32 ± 0.01, p < 0.001), and AG-bias also predicts DE (β = 0.51 ± 0.15, p < 0.001). For ASE, average expression significantly predicts DE between $F1_{mel}$ and $F1_{sim}$ (β = 0.21 ± 0.007, p < 0.001), and again AG-bias predicts DE (β = 0.71 ± 0.14, p < 0.001).

With a $\log_2$(fold change) cutoff, the AG-biased genes are even more likely to be DE. 16.6% of non-biased genes are DE between P_{mel} - P_{sim}, and 11.3% are DE between hybrid alleles. 37% of AG-biased genes are DE between P_{mel} - P_{sim}, and 28.6% are DE between hybrid alleles (Table S10). In the parental contrast, average expression weakly predicts DE between P_{mel} and P_{sim} (β = 0.03 ± 0.01, p < 0.001), and AG-bias strongly predicts DE (β = 0.95 ± 0.14, p < 0.001). For ASE, average expression does not predict DE between $F1_{mel}$ and $F1_{sim}$ (β = 0.01 ± 0.01, p = 0.40), and AG-bias strongly predicts DE (β = 0.96 ± 0.15, p < 0.001). It is therefore apparent that AG-biased genes are much more likely to be DE than more broadly expressed genes in the accessory gland and are especially enriched for large effect-size DE events.

Gene regulatory divergence classification

We characterized *cis*- and *trans*-regulatory effects for each autosomal gene by comparing ASE in hybrids to expression in each parent species. 2764 genes have evidence of *cis* effects (30% of expressed genes), 3338 have evidence of *trans* effects (36%), and 1601 have evidence of both *cis* and *trans* effects (17%). While there are more genes with significant *trans* effects, the median *cis* effect is significantly larger ($\log_2$(fold change) = 0.92) than the median *trans* effect ($\log_2$(fold change) = 0.64, Wilcoxon rank sum test, $p < 0.001$).

We further classified the regulatory basis of each gene following the algorithm outlined by McManus et al. (2010) (Figure 8, Tables S11, S12). 4062 genes (44%) are conserved, with nosignificant *cis* or *trans* effects. 933 (10.1%) are purely *cis*-regulated, and 912 (9.9%) are purely *trans*-regulated. Genes with both *cis* and *trans* effects are classified into three groups. First, there are 1116 (12.1%) with *cis + trans* regulation, where *cis* and *trans* effects have the same directionality (eg. $P_{mel} > P_{sim}$ and $F1_{mel} > F1_{sim}$). Secondly, just 104 (1.1%) genes have *cis by trans* regulation—*cis* and *trans* effects with opposite directionality (eg. $P_{mel} > P_{sim}$ and $F1_{mel} < F1_{sim}$).

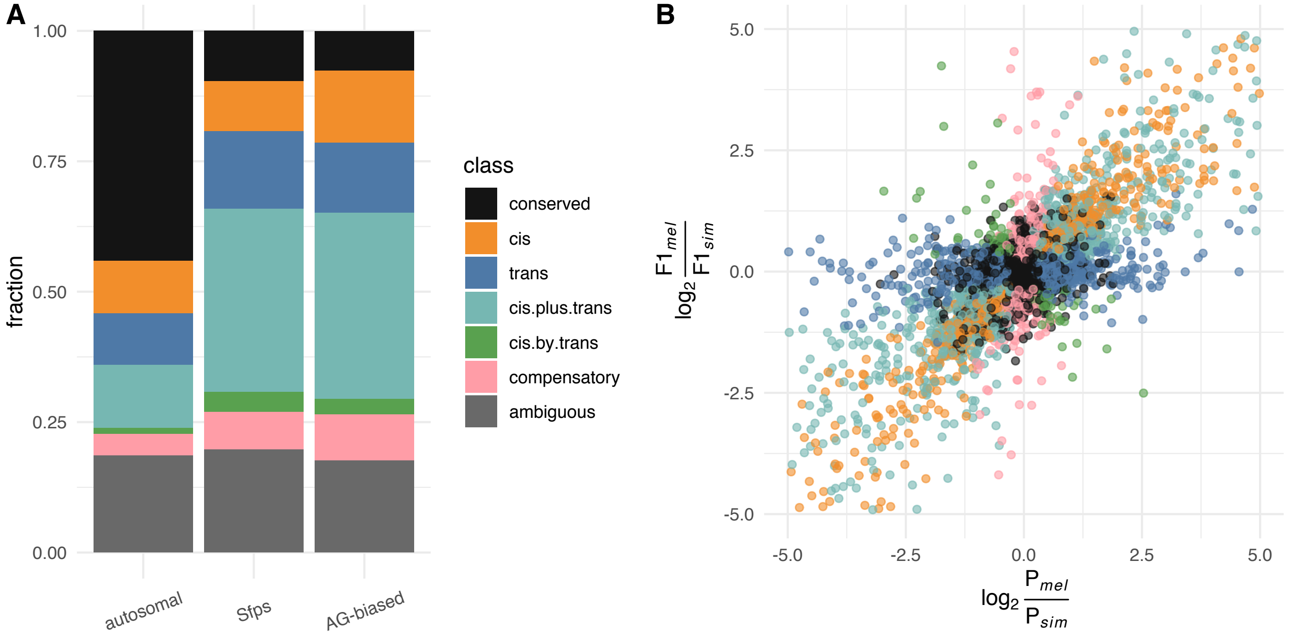

Figure 8. Regulatory classification by cis and trans mechanisms. A) Fraction of genes classified into each regulatory type, with Sfps and AG-biased genes shown separately. B) $\log_2$(fold change) of ASE and parental expression are shown with regulatory types highlighted. Ambiguous genes are removed, and scale is limited for clarity. For full data visualization see Figure S10.

Finally, there are 381 genes (4.1%) with compensatory regulation, such that there is no DE between P_{mel} and P_{sim} despite evidence of *cis* and *trans* regulatory evolution. There are additionally 1715 genes (18.6%) that cannot be classified into any of the above categories and are labeled ambiguous.

Comparing the regulatory classification of all genes to Sfps and AG-biased genes, both these gene classes are much less likely to be conserved (Figure 8A, Table S11), which is expected given their higher rates of divergence overall. Removing conserved and ambiguous classifications from consideration more clearly reveals differences in how Sfps and accessory gland-biased genes are regulated relative to all autosomal genes (Table S12). Both Sfps and accessory gland-biased genes are less likely to be purely *cis* regulated. There are about half as many *cis*-regulated genes among Sfps and a third fewer among accessory gland-biased genes. *cis* + *trans* regulation is particularly more common in Sfps and accessory gland-biased genes. Rates of pure *trans*, *cis* by *trans*, and compensatory regulation are roughly equal among gene sets.

Inheritance classification, misexpression, gain- and loss-of-function phenotypes

We characterized patterns of inheritance of expression phenotypes for autosomal genes by comparing $F1_{total}$ to each parent (Figure 9, Tables S13, S14). Conserved genes exhibit no DE in any comparison, comprising 4886 genes (53% of all expressed genes). 731 genes (7.9%) are additive, where P_{mel} and P_{sim} are DE and $F1_{total}$ has an intermediate expression phenotype. Genes with parental divergence and with $F1_{total}$ expression levels that were not DE relative to either parent are classified as either *mel* dominant or *sim* dominant. Rates are similar: 1403 (15.2%) are *mel* dominant and 1208 (13.1%) are *sim* dominant. Genes that are overexpressed in $F1_{total}$ relative to both parents are overdominant, and underexpressed genes underdominant. There are 448 (4.9%) overdominant and 547 (5.9%) underdominant genes.

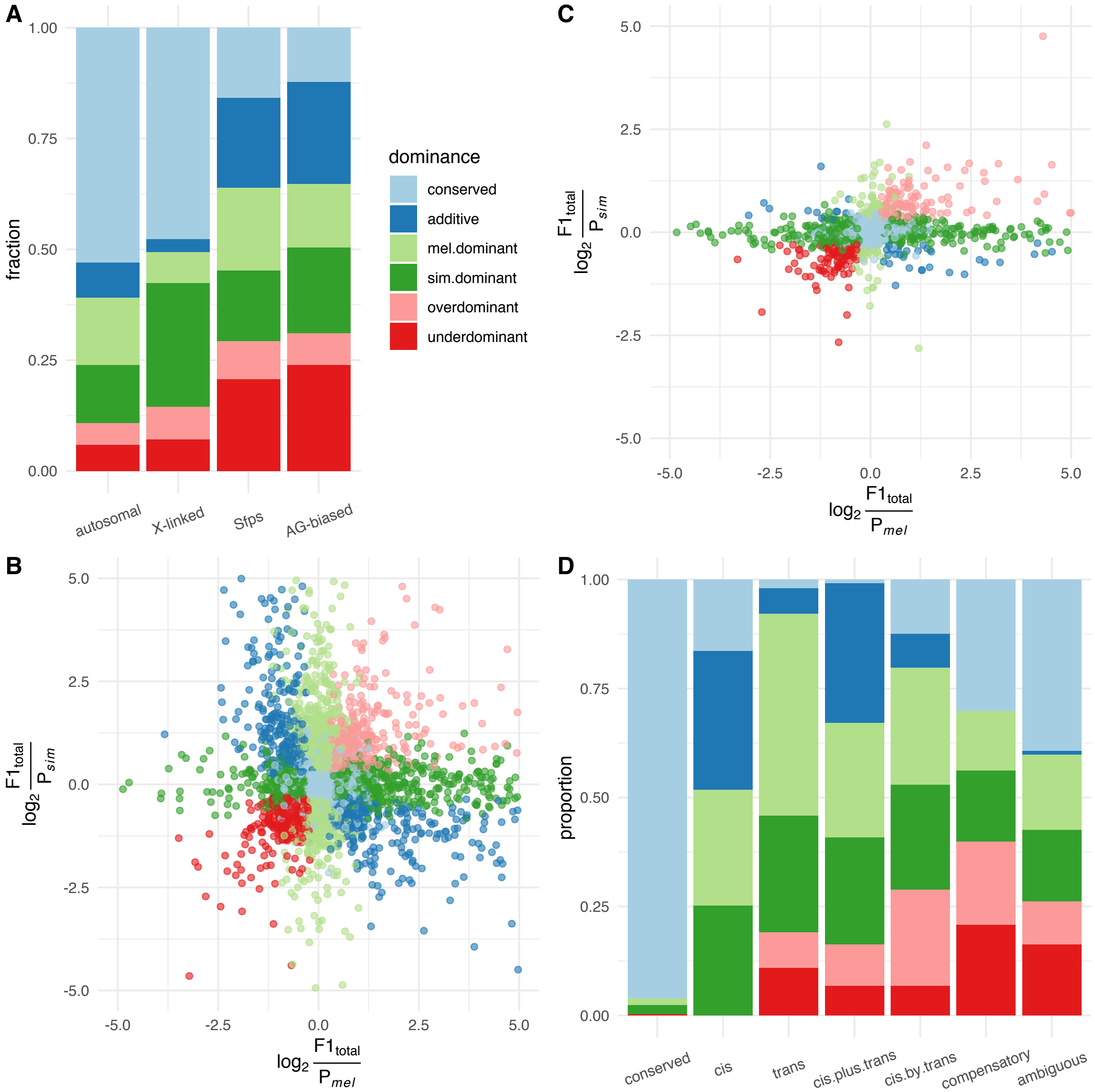

Figure 9. Inheritance classification gene expression phenotypes in hybrid offspring. A) Fraction of inheritance types for autosomal, X-linked, Sfps, and AG-biased genes are each shown independently. B) log2(fold change) of hybrid expression relative to each parent for autosomal genes. C) X-linked genes. Scale is limited in B-C for clarity. For full visualization, see Figure S10. D) Inheritance types among each of the regulatory classifications identified through *cis* and *trans* mechanisms.

We also classified X-linked genes according to phenotypic inheritance patterns (Figure 9A, C, Tables S13, S14). Compared to autosomal genes, X-linked have a similar percentage of conserved genes. As expected, given the hemizygous *sim* X chromosome and lack of *cis* effects in the hybrid, there is little additivity (3%) or *mel* dominant inheritance (7%); X-linked genes have a strong excess of *sim*-dominant phenotypes (28%). X-linked genes are also more likely to be underdominant or overdominant compared to autosomal genes: 14.5% of X-linked genes are misexpressed compared to 10.8% of autosomal genes (G-test, $p < 0.001$), consistent with the faster-X hypothesis (Vicoso and Charlesworth 2006).

As with regulatory classes, Sfps and accessory gland-biased genes are much less likely to be conserved than all genes. Looking at the distributions of non-conserved classes (Table S14), both gene classes are more likely to be additive than *mel* or *sim* dominant. Sfps and accessory gland-biased genes as well as significantly higher levels of underdominance than overdominance, a departure from trends among all autosomal genes.

Beyond misexpression, we also classified genes that are not DE between parents, and have a gain-of-function (GOF, overexpression in hybrids) or loss-of-function (LOF, underexpression in hybrids) expression phenotype (Supplemental Data). There are 58 genes with significant GOF (12% of all overexpressed genes), and 40 with significant LOF (10% of all underexpressed genes)—representing relatively rare events. Further, restricting GOF to cases with insignificant expression in parents ($\log_2$(counts) < 1 in P_{mel} and P_{sim}, $\log_2$(counts) > 1 in $F1_{total}$) leaves five instances, including *prolyl-4-hydroxylase-α MP* and four uncharacterized genes (Table S14). There are two cases of LOF with insignificant hybrid expression ($\log_2$(counts) > 1 in P_{mel} and P_{sim}, $\log_2$(counts) < 1 in $F1_{total}$): *β-Tubulin at 85D* and *Pendulin* (Table S16).

Next, we examined the relationship between regulatory and inheritance classes (Figure 9D, Table S17). We expect that genes with stronger *cis* components would be more likely to have an additive inheritance pattern, on the basis of the relative contributions of each species' allele to total hybrid expression (Lemos et al. 2008; McManus et al. 2010). We found that genes with strong *cis* regulatory components had the highest levels of additivity: 31.8% of *cis* and 31.9% of *cis* + *trans* regulated genes had additive inheritance, compared to just 5.9% of *trans* regulated genes. We expected that genes with antagonistic *cis* and *trans* components would be more likely to lead to misexpression in hybrids, highlighting potential incompatibilities between species. Indeed, we find that *cis*-by-*trans* and compensatory gene classes are more likely to lead to underdominant and overdominant inheritance patterns than other genes classes. *Cis* by *trans* regulated genes have an excess of overdominance (22%) relative to underdominance (7%), but this may be attributable to the small sample size of 104 *cis*-by-*trans* regulated genes. Finally, we observe a strong trend towards *trans* regulated genes being inherited in a *mel*-dominant fashion (46% of *trans* regulated genes, compared to just 27% being *sim*-dominant; *mel*-dominant genes have a significantly greater proportion of *trans* regulation: G-test, p < 0.001).

GO enrichment analysis

We ran a GO analysis to determine significantly enriched terms within regulatory and inheritance classes (Fig. S11, Supplemental Data). To increase the sample of X-linked genes, we used genes overexpressed or underexpressed in the hybrid relative to *D. simulans*, rather than genes strictly classified as overdominant or underdominant. Purely *cis*-regulated, *cis*-by-*trans*, and conserved genes are generally associated with larger p values and/or weakly enriched GO terms. Among our more significant results are 73 terms associated with translation in purely *trans*-regulated genes, including many ribosomal subunit and eukaryotic elongation factor proteins (Fig S11A, Supplemental Data). Translation-related genes are also significantly enriched for *mel*-dominant, and particularly for overdominant inheritance. Among the 40 translation proteins that are *trans* and *mel*-dominant, 32 are more highly expressed in *D. melanogaster* (chi-square test, p = 0.005). The remaining 33 genes with other modes of inheritance are not biased towards either parent (chi-square test, p = 0.82).

Translation-related GO terms are also enriched in overdominant inheritance gene sets on the autosomes and overexpressed genes on the X chromosome (Fig S11B, C, Supplemental Data). There are 61 overdominant translation-related genes on the autosomes. This gene set only partially overlaps with the *trans*-regulated gene set—18 of overdominant translation-related genes are *trans*-regulated, but eight are *cis* + *trans*, six are *cis* by *trans*, 12 are compensatory, and 17 are ambiguous. On the X chromosome, there are an additional 52 overexpressed genes associated with translation. Taken together, the data suggest that translation-related genes are especially likely to be both *trans*-regulated and overdominant, but that overdominance in these genes may be enacted through diverse regulatory mechanisms.

Underdominant inheritance / underexpression is strongly associated with golgi / endoplasmic reticulum vesicle transport GO terms on both the autosomes (137 genes, Fig. S11B) and X chromosome (36 genes, Fig. S11C). Of 137 underdominant transport-related genes on the autosomes, 86 have an ambiguous regulatory classification, while 20 are *trans*, 18 are *cis* + *trans*, and 13 are compensatory. Of the ambiguous terms, all are non-DE between parents, and non-DE between hybrid alleles. Therefore, there is no evidence of *cis* effects in these genes. Underdominance is indicative of *trans* factors, however these effects have not led to divergence between *D. melanogaster* and *D. simulans*, suggesting *trans* effects that occur specifically in the hybrid.

Upstream sequence divergence

We aligned noncoding sequences upstream of orthologous *D. melanogaster* and *D. simulans* TSSs and estimated the nucleotide substitution rate (Kimura 1980) to analyze divergence in putative promoter regions. We analyzed distributions of substitution rate for various upstream sequence lengths (Fig. S12). The 300 bp region captures the highest overall levels of divergence, so we chose this set for further analysis. We observe significant variation in upstream sequence evolution among regulatory and inheritance classes (Kruskal tests, $p < 0.01$). Among regulatory classes, conserved genes have the lowest rate of upstream sequence divergence with a median of 70.1 substitutions / kb (Figure 10A). All other classes except *cis* by *trans* have significantly greater divergence rates (Wilcoxon rank sum tests, p < 0.01). Ambiguous and compensatory genes have the greatest rates at 88.1 and 88.3 substitutions / kb, respectively. Among inheritance classes, additive genes (median 74.4 substitutions / kb) have similar rates to conserved genes (median 73.2 substitutions / kb; Wilcoxon rank sum test p > 0.05). Underdominant genes have much higher rates of upstream sequence divergence with a median of 103.5 substitutions / kb (pairwise Wilcoxon rank sum tests, p < 0.001 vs all other classes). Given the enrichment of underdominant genes for golgi / protein transport-related GO terms, we asked whether those

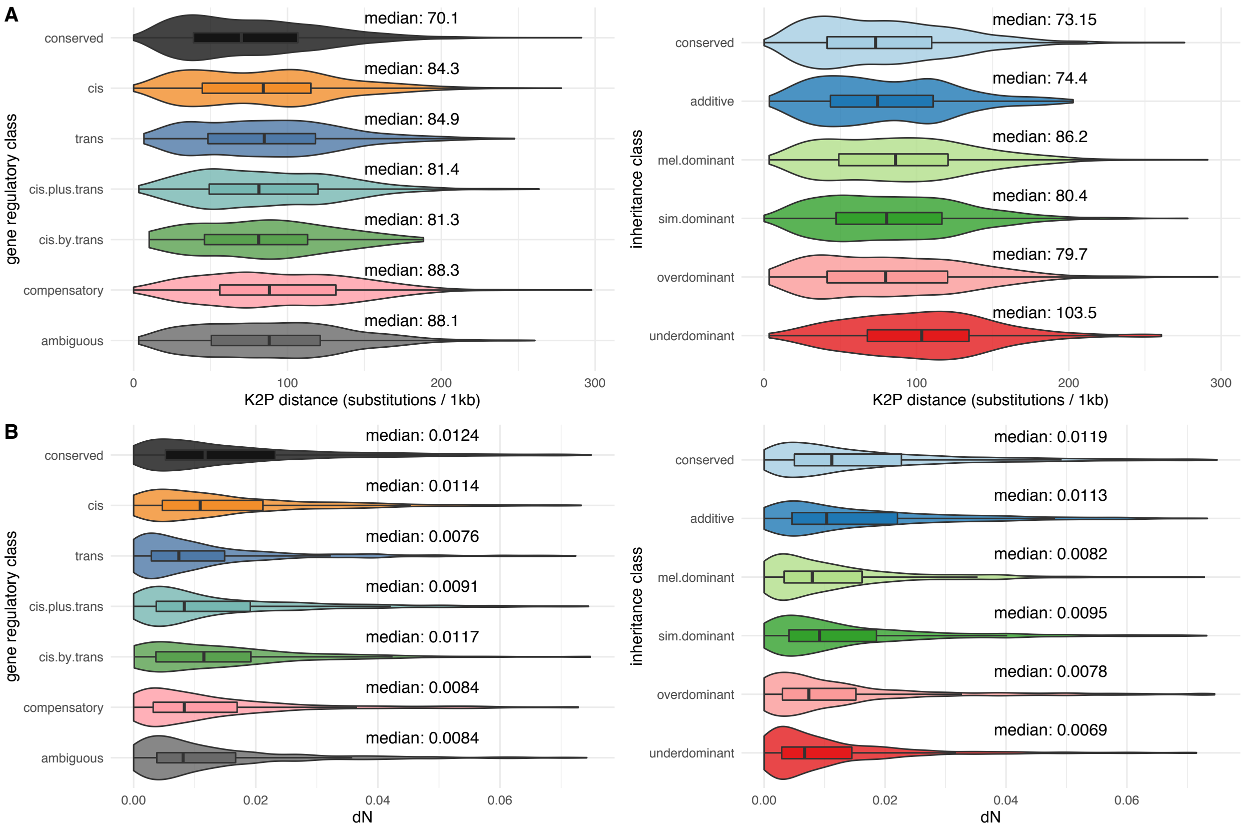

Figure 10. A) Distributions of Kimura-2-parameter estimated substitution rates among regulatory and inheritance classes. B) Distributions of nonsynonomous substitution rate (dN) among regulatory and inheritance classes. K2P distance and dN vary significantly across (Kruskal tests, p < 0.001). Alongside the median, significant differences by pairwise Wilcoxon rank sum tests (Holm-Bonferroni adjusted p < 0.05) are indicated by different letters (a, b, …) across gene sets.

genes were confounded with the elevated level of upstream sequence divergence. Of 451 underdominant genes with upstream sequence information, 109 are associated with golgi / protein transport-related GO terms. If we remove these from the analysis, underdominant genes still have significantly greater upstream sequence divergence than all other classes (median 101 substitutions / kb, pairwise Wilcoxon rank sum tests, $p < 0.001$). Genes with *mel* or *sim* dominance (median 86.2 and 80.4 substitutions / kb) have an intermediate level of upstream sequence divergence. Since *cis*-regulatory evolution could proceed through mutations in promoter regions, we asked whether the magnitude of ASE or parental divergence is correlated to upstream sequence divergence, however, there does not appear to be any relationship (Fig. S13). Upstream sequence divergence in Sfps or AG-biased genes does not differ significantly from non-Sfps / non-AG-biased genes (Wilcoxon rank sum tests, $p = 0.16$, $p = 0.26$, respectively).

Protein sequence evolution

Previous studies have demonstrated positive correlations between gene expression divergence and protein sequence evolution (Warnefors and Kaessmann 2013; Hodgins et al. 2016; Zhong, Lundberg, and Råberg 2021). We asked whether this was the case in our data, and additionally whether rates of protein sequence evolution vary among different regulatory and inheritance classes. There is no evidence of association between the rate of protein sequence evolution (dN) and expression divergence between parents or ASE in the accessory gland (Fig. S14A, B). In contrast, conserved genes tend to have higher dN than other regulatory and inheritance classes (Figure 10B). dN is negatively correlated with expression, as expected (Drummond et al. 2005). However, a multivariate regression of dN by expression level and parental expression divergence suggests that genes with parental conservation do have greater dN than DE genes (average expression: $\beta = 1.3\text{x}10^{-3} \pm 8.1\text{x}10^{-5}$, $p < 0.001$; parental expression DE: $\beta = 1.7\text{x}10^{-3} \pm 6.4\text{x}10^{-4}$, $p = 0.007$). Expression averaged between both parents Conserved genes are followed by *cis* regulated genes and other classes with strong *cis* components. *Trans*, compensatory, and ambiguous genes have relatively lower dN. Conserved and additive classes have higher dN than other inheritance modes. These are followed by *mel* and *sim* dominant genes, with overdominant and underdominant genes having the lowest dN.

We also observed differences in dN among Sfps and AG-biased genes (Fig. S14C, D). As expected given high rates of positive selection (Tsaur, Ting, and Wu 1998; Aguadé 1999; Begun et al. 2000; Holloway and Begun 2004; Begun et al. 2006-3; Schully and Hellberg 2006; Wong et al. 2008; Majane, Cridland, and Begun 2022) or relaxed selective constraint (Dapper and Wade 2020; Patlar et al. 2021) on Sfps, median dN is 3.6 times greater in Sfps than non-Sfps (Wilcoxon rank sum tests, p < 0.001). Among AG-biased non-Sfps, dN is modestly elevated 1.3 times higher, but still significantly different from non-accessory gland-biased genes (Wilcoxon rank sum tests, p = 0.00101, p < 0.001).

We also analyzed adaptive protein substitutions in *D. melanogaster* with McDonald-Kreitman tests and compared the summary statistic α (higher α suggests a greater overall proportion of adaptive amino acid substitutions) among gene classes. We find patterns that are similar to—although weaker than—patterns in dN (Fig. S15A). In both regulatory and inheritance classes, conserved genes have significantly higher median α than some other types, but we do not observe significant differences among other classifications. As with dN, we observe significantly elevated α among Sfps (Fig. S15B): median α among Sfps is 0.256; among non-Sfps, median α = -0.375 (Wilcoxon rank sum test, p < 0.001). Unlike dN, α does not differ significantly among AG-biased and non-biased genes (Fig. S15C, Wilcoxon rank sum test, p = 0.52).

Chromatin state integration

We used ATAC-Seq data (Blair et al., unpublished) from *D. melanogaster* and *D. simulans* to intersect chromatin accessibility with our DE and ASE analyses. We annotated ATAC-Seq

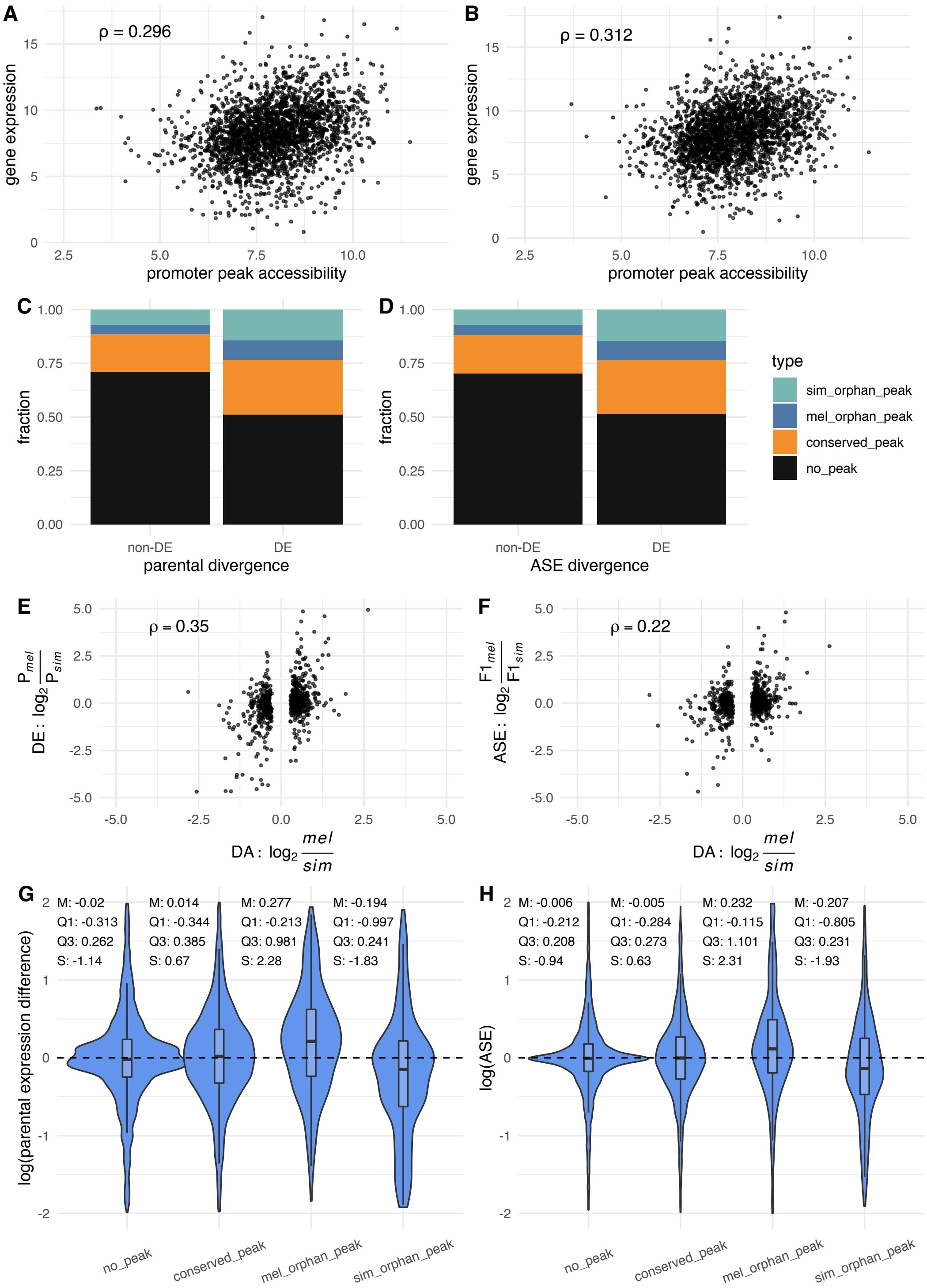

Figure 11. Interfacing promoter region accessibility estimated from ATAC-Seq data with gene expression. Gene expression and conserved peak accessibility have a positive relationship in (A) P_{mel} and (B) P_{sim}. Spearman's rank coefficient ρ is displayed. DE genes are more likely to be associated with orphan peaks in both the (C) parental and (D) ASE contrast. (E) Parental expression divergence and (F) ASE (y-axis) for are plotted against accessibility differences for DA conserved chromatin peaks called by ATAC-Seq (x-axis). (G) Distributions of $\log_2$(fold change) of parental expression difference differ among peak types. M: median, Q1: 1st quartile, Q3: 3rd quartile, S: skewness. Genes associated with conserved peaks have a broader 1st-3rd interquartile range and significantly larger median absolute value than genes without an annotated promoter peak. Orphan peaks are skewed towards greater expression values in the species with a peak. (H) Distributions of $\log_2$(fold change) of ASE show similar patterns to parental divergence.

peaks as conserved, *mel* orphans, or *sim* orphans. Conserved peaks are called in orthologous regions of both species, whereas orphan peaks are called only in one species. We used $\log_2$ of normalized counts of reads overlapping each peak region to define peak accessibility in each species, and quantified differential accessibility (DA) similarly to differential expression. Note that due to the high level of background in ATAC-Seq, there is accessibility even in orphan regions where no peak is called (eg., *D. simulans* will still have appreciable accessibility in the region of a *mel* orphan peak).

In total we annotated 7,416 conserved, 2,370 orphan *sim*, and 1,680 orphan *mel* peaks. We made peak-to-gene associations annotating peaks to the closest / overlapping transcription start site (TSS), which left us with 2,898 conserved, 1,627 orphan *sim*, and 1,232 orphan *mel* peaks with 1-to-1 gene annotations. Among annotated conserved peaks, 88% overlap the TSS. Roughly 6% of non-overlapping peaks are upstream of the TSS, and 6% are downstream. The median width of conserved peaks is 644 bp in both *D. melanogaster* and *D. simulans*, while the mean is 759.5 bp in *melanogaster* and 758 bp in *simulans*. Orphan peaks are much less likely to overlap with the TSS: 16% of orphan *mel* and 19% of orphan *sim* have overlap. Orphan peaks are also more likely to be upstream than downstream. In *D. melanogaster*, 57% of orphan peaks are upstream while 26% are downstream; in *D. simulans*, 50% are upstream and 30% are downstream. Orphan peaks are also smaller than conserved peaks: the median width of *mel* orphans is 399 bp, while the median width of *sim* orphans is 255 bp. PCA of $\log_2$(counts) shows that replicates cluster together by species (Fig. S16A-C), though we note that clustering is not as strong as RNA-Seq data, which is expected due to the background and variance inherent of ATAC-Seq data.

Relative accessibility among species in conserved peaks is normally distributed (Fig. S17A; 25% percentile $\log_2$(*mel* / *sim*) = -0.205; 75% percentile = 0.212), suggesting there is no systematic directionality in chromatin accessibility between species. The distribution of $\log_2$(*mel* / *sim*) for orphan peaks is highly skewed towards each respective species (Fig. S17B, C; *mel* orphans: 25% percentile $\log_2$(*mel* / *sim*) = 0.48; 75% percentile = 1.50; *sim* orphans: 25% percentile $\log_2$(*mel* / *sim*) = -1.48; 75% percentile = -0.48, in line with the expectation that the species with a peak called will have higher accessibility. There are a small number of cases where the species without a peak has higher accessibility (2.3% of *mel* orphans, 4.9% of *sim* orphans). We removed from further consideration A) orphan peaks that are less accessible in the species with a peak present, and B) orphan peaks that are not DA, leaving 74% of *mel* and 72% of *sim* orphans (Fig. S17D, E).

We expect that relatively more accessible chromatin should allow for greater levels of gene expression, so we assessed the relationship between peak accessibility and gene expression. There is a weak positive relationship between accessibility and expression (Figure 11A, B), suggesting that chromatin state and expression are indeed correlated. Given this relationship, we asked whether the presence of chromatin peaks was associated with the likelihood of DE in nearby genes. In both parental and ASE contrasts, genes that are DE are particularly enriched for the presence of nearby orphan peaks relative to non-DE genes (Figure 11C, D, Table S18). We used a multiple logistic regression with average expression and peak status as independent variables and DE as the dependent variable (DE ~ $\log_2$(counts) + peak). In the parental expression contrast, presence of a conserved peak is not related to DE (β = -0.04 ± 0.06, p = 0.512), but presence of orphan peaks are strong predictors of DE (*mel* orphan: β = 0.56 ± 0.10, p < 0.001, *sim* orphan: β = 0.48 ± 0.08, p < 0.001). Similarly, conserved peaks are not related to DE in ASE (β = -0.10 ± 0.07, p = 0.132), but orphan peaks strongly predict DE (*mel* orphan: β = 0.46 ± 0.10, p < 0.001, *sim* orphan: β = 0.49 ± 0.08, p < 0.001). Average gene expression also predicts DE in both contrasts, but the regression coefficients are notably smaller in comparison with orphan peak presence (parental: β = 0.22 ± 0.01, p < 0.001; ASE: β = 0.21 ± 0.01, p < 0.001).

If we define DE genes with $\log_2$(fold change) > 1, there is a stronger relationship between orphan peaks and DE than without a $\log_2$(fold change) cutoff, and a weak negative relationship between conserved peak presence and DE (Table S19). In the parental contrast, presence of a conserved peak negatively predicts DE (β = -0.17 ± 0.08, p = 0.046), and orphan peaks are strong predictors of DE (*mel* orphan: β = 0.78 ± 0.11, p < 0.001, *sim* orphan: β = 0.86 ± 0.09, p < 0.001). Similarly, conserved peaks are negatively related to DE in ASE (β = -0.30 ± 0.10, p = 0.002), and orphan peaks strongly predict DE (*mel* orphan: β = 0.85 ± 0.11, p < 0.001, *sim* orphan: β = 0.77 ± 0.10, p < 0.001).

Next, we asked whether the magnitude of expression differences between species or ASE was correlated with the magnitude of peak accessibility differences among DA peaks. We find a weakly positive relationship between ranks of these measures (Figure 11E, F; conserved peaks, parental divergence: ρ = 0.35; ASE: ρ = 0.22). These correlations are remarkably weaker for orphan peaks (Fig. S18). In summary, it appears that the presence of accessible chromatin increases the likelihood of differential expression of nearby genes, and that quantitative differences in conserved peak accessibility are correlated with concordant differences in gene expression. However, there is not a strong quantitative relationship here regarding orphan peaks.

To ask about the relationship of chromatin accessibility on *cis* and *trans* regulatory divergence, we compared Spearman rank correlations between expression divergence and accessibility divergence among conserved, pure *cis*, pure *trans*, and *cis* + *trans* regulated genes (Table S20). As expected, among conserved genes there is no correlation of accessibility divergence with either parental expression (ρ = -0.01) or ASE divergence (ρ = -0.08). Genes with *trans*-regulatory components have a stronger correlation of accessibility divergence with parental expression divergence (ρ = 0.52 for both *trans* and *cis* + *trans*) than pure *cis*-regulated genes (ρ = 0.41). Genes with *trans*-regulatory components have a no correlation of accessibility divergence with ASE divergence (ρ = 0.14), while *cis* (ρ = 0.42) and *cis* + *trans* (ρ = 0.44) genes have relatively strong correlations. It therefore appears that genes with *cis* and *trans* regulatory divergence have relatively stronger correlations between accessibility and parental expression divergence, and that only *cis*-regulated genes show correlations between accessibility divergence and ASE divergence. Finally, we observed the relative distributions of gene expression differences in genes with nearby 1-to-1 peaks and genes without a nearby peak, by comparing log2(fold changes) (Figure 11G). In the parental contrast (P_{mel} - P_{sim}), genes without a peak annotation have a narrower distribution of log2(P_{mel} / P_{sim}) relative to peaks with a conserved peak nearby, but both gene sets have medians near 0. The median absolute value of $\log_2(P_{mel} / P_{sim})$ in genes with a conserved peak is 0.37, significantly higher than genes with no peak, median = 0.29 (Wilcoxon rank sum test, $p < 0.001$). Genes with a *mel* orphan peak nearby are biased towards positive values of $\log_2(P_{mel} / P_{sim})$ with median = 0.252, and genes with a *sim* orphan peak nearby are biased towards negative values with median = -0.221. The median absolute values of $\log_2(P_{mel} / P_{sim})$ of genes near orphan peaks are significantly greater than genes near a conserved peak or no peak (median *mel* orphan = 0.61; *sim* orphan = 0.59; Kruskal test, $p < 0.001$, pairwise Wilcoxon rank sum tests, $p < 0.001$ in each case). Medians associated with *mel* peaks and *sim* peaks are not significantly different (Wilcoxon rank sum test, $p = 0.30$). We observe the same patterns in ASE among different classes of peaks, but the magnitude of expression differences is smaller compared to parental DE (Figure 11H).

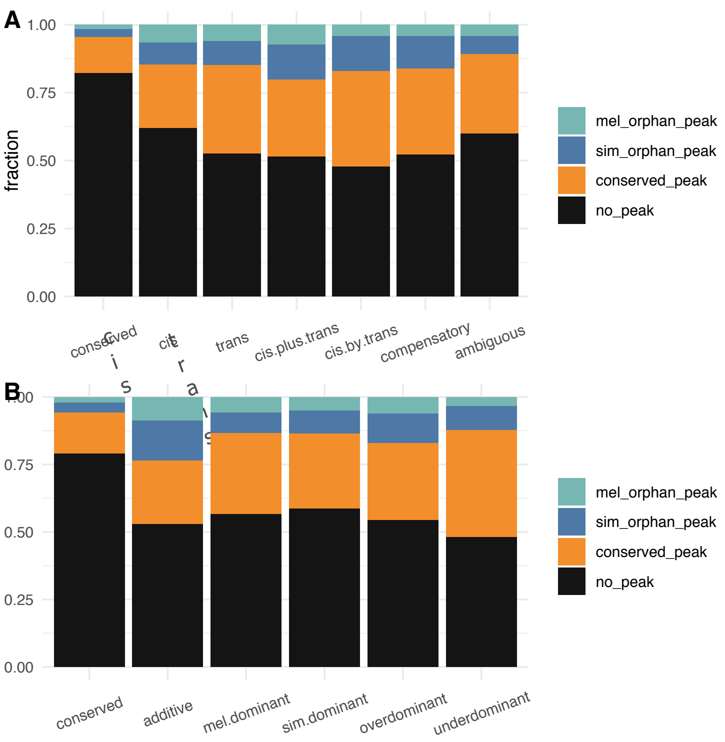

Figure 12. Proportions of peak types among A) regulatory and B) dominance classes.

Regulatory and inheritance classes are associated with different proportions of chromatin peak types (Figure 12, Table S21). As expected, genes that are annotated as conserved in regulation or inheritance are much less likely to be associated with nearby peaks. The set of genes with purely *cis* regulation have a smaller proportion of conserved peaks than genes with significant *trans* factors. Genes with *cis* + *trans* regulation, and genes with additive inheritance, have the highest respective shares of orphan peaks. Underdominant genes have the highest share of conserved peaks among inheritance classes.

DISCUSSION

Autosomal gene expression profiles between *D. melanogaster* and *D. simulans* were overall similar as evidenced by transcriptome-wide expression correlations (*r* = 0.934). The correlation is somewhat stronger than what we observed in our recent bulk RNA-Seq (Cridland et al. 2020) and single-cell RNA-Seq studies (Majane, Cridland, and Begun 2022). Given the myriad technical differences between independent studies, it is difficult to put the strength of these correlations into a broader context, and we therefore cannot make conclusions about the relative level of transcriptome divergence observed here to other Drosophila tissues or species. Hybrid expression profiles are overall more like each parent than the parents are to each other, and appear to be a midpoint between parental expression, as evidenced by PCA. ASE profiles within the hybrid are more similar than parents are to one another, as expected given that parental divergence is the result of *cis* and *trans* effects, while divergence between hybrid alleles is driven only by *cis* effects. Similarly, we observed a higher rate of DE between parents than between hybrid alleles. As we see with transcriptomic correlations, large effect-size DE between hybrid alleles and parent-of-origin is rare (4% of genes), suggesting that *trans*- effects of large effect size are uncommon.

X-linked genes are strikingly more similar between the hybrid and *D. simulans* than to *D. melanogaster* or between the parents, which is expected since, given that the hybrids have a D. *simulans* X chromosome, deviations from *simulans* X-linked expression in the hybrid must be due to *trans*-acting factors. X-linked and autosomal gene expression profiles have very similar correlations between parents. X-linked and autosomal genes also have similar levels of DE, with and without a $\log_2$(fold change) cutoff. These results are unexpected since whole-animal transcriptome data shows a strong faster-X effect with greater divergence between X-linked genes, even among non-male biased genes (Meisel, Malone, and Clark 2012; Kayserili et al. 2012). We observed a modest rate of DE between the hybrid and *D. simulans* X chromosomes, suggesting *trans* effects are not uncommon on the X. However, large effect-size DE in this contrast is particularly rare at just 4% of genes, suggesting that these *trans* effects are unlikely to have a large effect on gene expression.

Since Sfps are known to evolve rapidly at the level of protein sequence and gene turnover, we asked whether they had particularly high rates of expression divergence between species. Sfps are not any more likely to be DE than all autosomal genes. However, Sfps are much more likely to have large effect-size DE. This suggests that in addition to extraordinarily rapid protein divergence, Sfps have also diverged in expression level at an unusually high rate. Accessory gland-biased non-Sfps also have a particularly high rate of expression divergence. Taken together, it therefore appears that genes that comprise the accessory gland's unique transcriptome are diverging at a particularly elevated rate. Given the gland's function and central role in sexual conflict between males and females, adaptive divergence is a plausible explanation for this expression divergence, but future research using phylogenetic methods to model expression phenotypes as evolving traits on a tree is required to distinguish between directional selection and relaxed constraint on expression.

Consistent the observation of relatively little DE between hybrid alleles and parents-of-origin, we find the median *cis* effect is 43% larger than the median *trans* effect, in contrast to McManus et al. (2010), who observed significantly larger *trans* effects in whole female Drosophila hybrids. Despite being smaller, *trans* effects are more common than *cis* effects in our data. While there is a general expectation that *cis* effects accumulate more than *trans* between species (reviewed in Signor and Nuzhdin 2018; Hill, Vande Zande, and Wittkopp 2021), this is not always the case (McManus et al. 2010; Coolon et al. 2014; Sánchez-Ramírez et al. 2021). More work is needed on diverse somatic organs to understand the distribution of *cis* and *trans* effects across tissues, which could help answer a broader question of which factors influence the variation observed in relative levels of interspecific *cis* and *trans* divergence.

We classified genes into regulatory and inheritance classes as originally outlined in McManus et al. (2010). We find that relatively equal proportions of genes have evolved through purely *cis* or trans *regulation*, and a larger proportion evolve through both mechanisms. Opposing directionality of cis and trans evolution (*cis* by *trans* and compensatory classes) is particularly rare. Interestingly, it appears that opposing *cis* and *trans* effects are particularly pronounced in studies with whole adult females or heads (Gibson et al. 2004; Ranz et al. 2004; Landry et al. 2005; Graze et al. 2009; McManus et al. 2010; Coolon et al. 2014). Mouse liver (Goncalves et al. 2012) and testes (Mack, Campbell, and Nachman 2016) also have relatively high levels of opposing *cis* and *trans* effects. Notably, Sánchez-Ramírez et al. (2021) found a higher level of opposing *cis* and *trans* effects among female-biased genes in Caenorhabditis. Cartwright and Lott (2020) found low levels of opposing *cis* and *trans* effects in the early embryo, which increased in later stage embryos. Avian brains (Davidson and Balakrishnan 2016) and Hawai'ian testes (Brill et al. 2016) also had low levels of opposing *cis* and *trans* effects. Thus, there is evidence of variation among species, tissues, developmental stage, and sex. Given the sparsity of interspecific tissue-specific data however, it is difficult to draw general conclusions from the existing studies. Sfps and accessory gland-biased genes appear to accumulate higher levels of cis + trans regulation than all autosomal genes, with purely cis or trans regulation being relatively rare. Additionally, Sfps and accessory gland-biased genes have much more additivity than all genes, as well as elevated misexpression (see below). Levels of opposing cis and trans effects are very similar to all autosomal genes.

Misexpression is relatively rare, with just 10.8% of autosomal genes overdominant or underdominant in the hybrid. Further, just a handful of genes have complete GOF or LOF expression phenotypes. These results suggest that the accessory gland is not prone to widespread dysgenesis between these species, in contrast to results in Drosophila testes or female whole-animal data (Ranz et al. 2004; Haerty and Singh 2006; Moehring, Teeter, and Noor

2007; McManus et al. 2010; Coolon et al. 2014; Cartwright and Lott 2020), Caenorhabditis (Sánchez-Ramírez et al. 2021), mouse liver (Goncalves et al. 2012) and mouse testis (Mack, Campbell, and Nachman 2016). Thus, transcriptomic data are consistent with Stumm-Zollinger and Chen's observations (1988) that these hybrid accessory glands have relatively normal morphology, Sfp protein expression, and ability to induce the female PMR. Other studies have also found limited levels of misexpression in Drosophila larva (Moehring, Teeter, and Noor 2007; Wei, Clark, and Barbash 2014), female heads (Graze et al. 2009), Hawai'ian Drosophila testes (Brill et al. 2016), and avian brains (Davidson and Balakrishnan 2016). Clearly, the level of hybrid dysgenesis in gene expression is highly variable among species and tissues; more work on tissue-specific ASE is needed to fill the gaps and broaden our understanding of hybrid misexpression.

Several studies in Drosophila have found that male-biased genes are prone to being misexpressed in hybrids and are especially likely to be underdominant in whole animals or testis (Haerty and Singh 2006; Michalak and Noor 2003; Moehring, Teeter, and Noor 2007; McManus et al. 2010), a pattern observed in most, but not all Drosophila crosses (Banho et al. 2021). Underdominant male-biased genes are also linked to male sterility (Michalak and Noor 2004). We also found higher levels of misexpression among male biased genes with a particular enrichment for overdominance. Autosomal genes have similar rates of overdominance and underdominance. Sfps have 2.7 times as many misexpressed genes, with a ratio of underdominant to overdominant expression of 2.3; accessory gland-biased non-Sfps are have 2.9 times as many misexpressed genes, with a ratio of underdominant to overdominant expression of 3.4. These data suggest that the widely reported observation of male biased underexpression is not limited just to the testis. Whether this pattern holds for male-biased genes across multiple somatic tissues or only those related to reproduction is an important question for future studies.

Two predictions of the faster-X hypothesis (Vicoso and Charlesworth 2006) are 1) elevated expression divergence among X-linked genes (faster-X divergence) along with 2) increased misexpression in hybrids (faster-X misexpression). Faster-X divergence has been observed in flies (Meisel, Malone, and Clark 2012; Kayserili et al. 2012), but Drosophila hybrids actually have a slower-X misexpression pattern in the testes (Lu et al. 2010; Llopart 2012) or no difference from autosomal misexpression in larvae (Wei, Clark, and Barbash 2014). On the contrary, our data shows no evidence of faster-X divergence, but does show faster-X misexpression, similar patterns to those observed in mice (Good et al. 2010; Larson et al. 2016). Faster-X gene expression patterns may therefore vary among tissues in Drosophila, and future studies of tissue-specific hybrid gene expression are needed to determine the extent and biology underlying these potential differences.

Genes that are autosomal overdominant or X-linked overexpressed are both highly enriched for translation-related genes, including numerous elongation factors and ribosomal subunits. Overexpressed proteins in *D. melanogaster* - *D. simulans* hybrid embryos were enriched for “translation initiation” genes (Bamberger et al. 2018), and misexpressed genes in hybrid house mice testes were also significantly enriched for translation-related GO terms (Mack, Campbell, and Nachman 2016). Therefore, misexpression of translation-related genes could plausibly be related to hybrid incompatibilities across species and developmental stages. Notably, hybrid male sterility evolves very quickly, and the regulation of spermatogenesis occurs primarily at the translational level (Schäfer et al. 1995). Whether sterility and misexpression of translational machinery are linked remains a speculative matter. Translation is also enriched among *trans-*

regulated and *mel*-dominant genes in our data, but these gene sets are only partially overlapping with one another, suggesting that translation-related genes may be regulated and inherited via diverse mechanisms in the accessory gland, and that overdominant translation-related genes are not regulated through a unifying mechanism. Underdominant autosomal and underexpressed X-linked genes are highly enriched for genes related to protein transport, golgi, and endoplasmic reticulum. Notably, ambiguous regulated genes are also enriched for these GO terms, and this gene list substantially overlaps with underdominance. Ambiguous regulation may occur in many ways and is difficult to put into biological context, however in this case, most of these genes are not DE between the parents but do have evidence of *trans*-effects leading to underdominance. This suggests that emergent properties of *trans* factors active specifically in hybrid cells leads to underexpression, potentially indicative of hybrid incompatibilities related to protein transport in the golgi and endoplasmic reticulum.

We analyzed levels of nucleotide divergence in regions upstream of the TSS, which could plausibly affect promoter regions and thereby impact expression divergence. While we do not find a quantitative relationship between expression divergence and upstream sequence divergence, genes that are conserved in their regulation and inheritance tend to have a lower level of divergence. Compensatory, ambiguous, and particularly underdominant genes have elevated levels of upstream sequence divergence. This suggests that underdominance might be arising from incompatibilities between rapidly evolving *cis* loci and *trans* regulatory factors.

We also examined protein sequence in the context of expression evolution. As expected, Sfps have extraordinarily high dN. Sfps also have elevated median α, suggesting overall greater levels of adaptive substitutions in these genes. Accessory gland-biased non-Sfps also have elevated dN, but no significant elevation in α. These results suggest that in addition to Sfps, other proteins characteristic of the accessory gland evolve rapidly. If accessory gland-biased non-Sfps are involved in the production of seminal fluid, which seems likely given that is the gland’s only known function, then it is plausible they may be subject to the same selective forces as Sfps, including positive selection (Tsaur, Ting, and Wu 1998; Aguadé 1999; Begun et al. 2000; Holloway and Begun 2004; Begun et al. 2006-3; Schully and Hellberg 2006; Wong et al. 2008; Majane, Cridland, and Begun 2022) and/or relaxed constraint (Dapper and Wade 2020; Patlar et al. 2021). Unexpectedly, we found that conserved genes have the highest overall levels of dN, ω, and α in the accessory gland, despite Sfps and accessory gland-biased genes having high levels of both forms of divergence. This suggests that for most genes expressed in the accessory gland, there is a modest decoupling between expression evolution in the gland and protein sequence evolution.

Finally, we associated ATAC-Seq peaks with nearest TSS to identify chromatin tied to putative regulatory regions. We found that orphan peaks were highly biased in accessibility towards one species. We observed a modest but significant correlation between gene expression and accessibility among conserved peaks, similar to results of some studies (Nair et al. 2021), but weaker than others (Starks et al. 2019). We observe weaker correlations with orphan peaks, which may be explained by the greater median distance of orphan peaks to the TSS. Many factors contribute to expression level in addition to chromatin state, so we expect to find relatively weak positive relationships. Additionally, the ATAC-Seq data we used here includes only accessory gland tissue, while the RNA-Seq data is from the accessory gland and ejaculatory duct. We therefore expect that our observations relating gene expression and $\log_2$(fold changes) across groups to chromatin state in this study are conservative estimates of true relationships.

We further identified a weakly positive quantitative relationship between DA conserved peak accessibility and the magnitude of DE, in contrast to some other studies that find stronger associations (Racioppi, Wiechecki, and Christiaen 2019; Gontarz et al. 2020; Nair et al. 2021; Sanghi et al. 2021). However, if we limit the analysis to genes with evidence of *cis* or *trans* regulatory divergence, we observe a relatively strong correlations between peak accessibility divergence and expression divergence, suggesting a concordant directionality effect linking chromatin state and expression. Orphan peaks do not appear to have this quantitative relationship, but the presence of an orphan peak very strongly predicts both small- and large-effect size DE in nearby genes. Taken together, the data suggest that presence/absence of chromatin peaks (either by evolutionary gain or loss, which is impossible to determine with this data) likely contributes to gene expression differences between *D. melanogaster* and *D. simulans*—if a peak appears in one species, there is a better chance that the nearest gene will be DE, and more often than not in the direction of the species with the peak—but there is no evidence of a straightforward quantitative relationship, at least that we can detect with this dataset. To overcome some of the technical and biological variables that complicate this analysis, an allele-specific multi-omic gene expression and ATAC-Seq experiment on single cells (Cao et al. 2018; S. Chen, Lake, and Zhang 2019) would provide stronger insights into the relationships between chromatin state, expression, and regulation.

DATA AVAILABILITY

Upon peer-reviewed publication of this work, all scripts and supplemental data will be published at github.com/alexmajane/hybridASE. Sequencing reads will be released on the NCBI SRA. Prior to publication, data and scripts can be made available upon request by emailing Alex Majane (acmajane@gmail.com).

LITERATURE CITED

Aguadé, Montserrat. 1999. “Positive Selection Drives the Evolution of the Acp29AB Accessory Gland Protein in Drosophila.” *Genetics* 152 (2): 543–51.

Andergassen, Daniel, Christoph P. Dotter, Daniel Wenzel, Verena Sigl, Philipp C. Bammer, Markus Muckenhuber, Daniela Mayer, et al. 2017. “Mapping the Mouse Allelome Reveals Tissue-Specific Regulation of Allelic Expression.” *eLife* 6 (August). https://doi.org/10.7554/eLife.25125.

Anders, Simon, and Wolfgang Huber. 2010. “Differential Expression Analysis for Sequence Count Data.” *Genome Biology* 11 (10): R106.

Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber. 2015. “HTSeq--a Python Framework to Work with High-Throughput Sequencing Data.” *Bioinformatics* 31 (2): 166–69.

Assis, Raquel, Qi Zhou, and Doris Bachtrog. 2012. “Sex-Biased Transcriptome Evolution in Drosophila.” *Genome Biology and Evolution* 4 (11): 1189–1200.

Avila, Frank W., Laura K. Sirot, Brooke A. LaFlamme, C. Dustin Rubinstein, and Mariana F. Wolfner. 2011. “Insect Seminal Fluid Proteins: Identification and Function.” *Annual Review of Entomology* 56: 21–40.

Babak, Tomas, Brian DeVeale, Emily K. Tsang, Yiqi Zhou, Xin Li, Kevin S. Smith, Kim R. Kukurba, et al. 2015. “Genetic Conflict Reflected in Tissue-Specific Maps of Genomic Imprinting in Human and Mouse.” *Nature Genetics* 47 (5): 544–49.

Bamberger, Casimir, Salvador Martínez-Bartolomé, Miranda Montgomery, Mathieu Lavallée-Adam, and John R. Yates 3rd. 2018. “Increased Proteomic Complexity in Drosophila Hybrids during Development.” *Science Advances* 4 (2): eaao3424.

Banho, Cecilia A., Vincent Mérel, Thiago Y. K. Oliveira, Claudia M. A. Carareto, and Cristina Vieira. 2021. “Comparative Transcriptomics between Drosophila Mojavensis and D. Arizonae Reveals Transgressive Gene Expression and Underexpression of Spermatogenesis-Related Genes in Hybrid Testes.” *Scientific Reports* 11 (1): 9844.

Begun, David J., and Heather A. Lindfors. 2005. “Rapid Evolution of Genomic Acp Complement in the Melanogaster Subgroup of Drosophila.” *Molecular Biology and Evolution* 22 (10): 2010–21.

Begun, David J., Heather A. Lindfors, Melissa E. Thompson, and Alisha K. Holloway. 2006-3. “Recently Evolved Genes Identified From Drosophila Yakuba and D. Erecta Accessory Gland Expressed Sequence Tags.” *Genetics* 172 (3): 1675–81.

Begun, David J., Penn Whitley, Bridget L. Todd, Heidi M. Waldrip-Dail, and Andrew G. Clark. 2000. “Molecular Population Genetics of Male Accessory Gland Proteins in Drosophila.” *Genetics* 156 (4): 1879–88.

Boorman, E., and G. A. Parker. 1976. “Sperm (ejaculate) Competition in Drosophila Melanogaster, and the Reproductive Value of Females to Males in Relation to Female Age and Mating Status.” *Ecological Entomology* 1 (3): 145–55.

Brawand, David, Magali Soumillon, Anamaria Necsulea, Philippe Julien, Gábor Csárdi, Patrick Harrigan, Manuela Weier, et al. 2011. “The Evolution of Gene Expression Levels in Mammalian Organs.” *Nature* 478 (7369): 343–48.

Brill, E., L. Kang, K. Michalak, P. Michalak, and D. K. Price. 2016. “Hybrid Sterility and Evolution in Hawaiian Drosophila: Differential Gene and Allele-Specific Expression Analysis of Backcross Males.” *Heredity* 117 (2): 100–108.

Cao, Junyue, Darren A. Cusanovich, Vijay Ramani, Delasa Aghamirzaie, Hannah A. Pliner, Andrew J. Hill, Riza M. Daza, et al. 2018. “Joint Profiling of Chromatin Accessibility and Gene Expression in Thousands of Single Cells.” *Science* 361 (6409): 1380–85.

Carlson, Marc. 2021. “org.Dm.eg.db: Genome Wide Annotation for Fly.” *R Package*.

Cartwright, Emily L., and Susan E. Lott. 2020. “Evolved Differences in Cis and Trans Regulation Between the Maternal and Zygotic mRNA Complements in the Drosophila Embryo.” *Genetics* 216 (3): 805–21.

Castel, Stephane E., François Aguet, Pejman Mohammadi, GTEx Consortium, Kristin G. Ardlie, and Tuuli Lappalainen. 2020. “A Vast Resource of Allelic Expression Data Spanning Human Tissues.” *Genome Biology* 21 (1): 234.

Chapman, T., D. M. Neubaum, M. F. Wolfner, and L. Partridge. 2000. “The Role of Male Accessory Gland Protein Acp36DE in Sperm Competition in Drosophila Melanogaster.” *Proceedings. Biological Sciences / The Royal Society* 267 (1448): 1097–1105.

Chen, Jenny, Ross Swofford, Jeremy Johnson, Beryl B. Cummings, Noga Rogel, Kerstin Lindblad-Toh, Wilfried Haerty, Federica di Palma, and Aviv Regev. 2019. “A Quantitative Framework for Characterizing the Evolutionary History of Mammalian Gene Expression.” *Genome Research* 29 (1): 53–63.

Chen, Song, Blue B. Lake, and Kun Zhang. 2019. “High-Throughput Sequencing of the Transcriptome and Chromatin Accessibility in the Same Cell.” *Nature Biotechnology* 37 (12): 1452–57.

Clark, A. G., M. Aguadé, T. Prout, L. G. Harshman, and C. H. Langley. 1995. “Variation in Sperm Displacement and Its Association with Accessory Gland Protein Loci in Drosophila Melanogaster.” *Genetics* 139 (1): 189–201.

Clark, A. G., D. J. Begun, and T. Prout. 1999. “Female X Male Interactions in Drosophila Sperm Competition.” *Science* 283 (5399): 217–20.

Cock, Peter J. A., Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, et al. 2009. “Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics.” *Bioinformatics* 25 (11): 1422–23.

Combs, Peter A., Joshua J. Krupp, Neil M. Khosla, Dennis Bua, Dmitri A. Petrov, Joel D. Levine, and Hunter B. Fraser. 2018. “Tissue-Specific Cis-Regulatory Divergence Implicates eloF in Inhibiting Interspecies Mating in Drosophila.” *Current Biology: CB* 28 (24): 3969–75.e3.

Coolon, Joseph D., C. Joel McManus, Kraig R. Stevenson, Brenton R. Graveley, and Patricia J. Wittkopp. 2014. “Tempo and Mode of Regulatory Evolution in Drosophila.” *Genome Research* 24 (5): 797–808.

Cridland, Julie M., Alex C. Majane, Hayley K. Sheehy, and David J. Begun. 2020. “Polymorphism and Divergence of Novel Gene Expression Patterns in Drosophila Melanogaster.” *Genetics* 216 (1): 79–93.

Dapper, Amy L., and Michael J. Wade. 2020. “Relaxed Selection and the Rapid Evolution of Reproductive Genes.” *Trends in Genetics: TIG* 36 (9): 640–49.

Davidson, John H., and Christopher N. Balakrishnan. 2016. “Gene Regulatory Evolution During Speciation in a Songbird.” *G3* 6 (5): 1357–64.

Dickinson, W. J., Robert G. Rowan, and Mark D. Brennan. 1984. “Regulatory Gene Evolution: Adaptive Differences in Expression of Alcohol Dehydrogenase in Drosophila Melanogaster and Drosophila Simulans.” *Heredity* 52 (2): 215–25.

Drummond, D. Allan, Jesse D. Bloom, Christoph Adami, Claus O. Wilke, and Frances H. Arnold. 2005. “Why Highly Expressed Proteins Evolve Slowly.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (40): 14338–43.

Edgar, Robert C. 2004. “MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput.” *Nucleic Acids Research* 32 (5): 1792–97.

Ellegren, Hans, and John Parsch. 2007. “The Evolution of Sex-Biased Genes and Sex-Biased Gene Expression.” *Nature Reviews. Genetics* 8 (9): 689–98.

Fiumera, Anthony C., Bethany L. Dumont, and Andrew G. Clark. 2005. “Sperm Competitive Ability in Drosophila Melanogaster Associated with Variation in Male Reproductive Proteins.” *Genetics* 169 (1): 243–57.

Fraïsse, Christelle, Gemma Puixeu Sala, and Beatriz Vicoso. 2019. “Pleiotropy Modulates the Efficacy of Selection in Drosophila Melanogaster.” *Molecular Biology and Evolution* 36 (3): 500–515.

Gibson, Greg, Rebecca Riley-Berger, Larry Harshman, Artyom Kopp, Scott Vacha, Sergey Nuzhdin, and Marta Wayne. 2004. “Extensive Sex-Specific Nonadditivity of Gene Expression in Drosophila Melanogaster.” *Genetics* 167 (4): 1791–99.

Goldman, N., and Z. Yang. 1994. “A Codon-Based Model of Nucleotide Substitution for Protein-Coding DNA Sequences.” *Molecular Biology and Evolution* 11 (5): 725–36.

Goncalves, Angela, Sarah Leigh-Brown, David Thybert, Klara Stefflova, Ernest Turro, Paul Flicek, Alvis Brazma, Duncan T. Odom, and John C. Marioni. 2012. “Extensive Compensatory Cis-Trans Regulation in the Evolution of Mouse Gene Expression.” *Genome Research* 22 (12): 2376–84.

Gontarz, Paul, Shuhua Fu, Xiaoyun Xing, Shaopeng Liu, Benpeng Miao, Viktoriia Bazylianska, Akhil Sharma, et al. 2020. “Comparison of Differential Accessibility Analysis Strategies for ATAC-Seq Data.” *Scientific Reports* 10 (1): 10150.

Good, Jeffrey M., Thomas Giger, Matthew D. Dean, and Michael W. Nachman. 2010. “Widespread over-Expression of the X Chromosome in Sterile F_1hybrid Mice.” *PLoS Genetics* 6 (9): e1001148.

Graveley, Brenton R., Angela N. Brooks, Joseph W. Carlson, Michael O. Duff, Jane M. Landolin, Li Yang, Carlo G. Artieri, et al. 2011. “The Developmental Transcriptome of Drosophila Melanogaster.” *Nature* 471 (7339): 473–79.

Graze, Rita M., Lauren M. McIntyre, Bradley J. Main, Marta L. Wayne, and Sergey V. Nuzhdin. 2009. “Regulatory Divergence in Drosophila Melanogaster and D. Simulans, a Genomewide Analysis of Allele-Specific Expression.” *Genetics* 183 (2): 547–61, 1SI – 21SI.

Gruber, Jonathan D., Kara Vogel, Gizem Kalay, and Patricia J. Wittkopp. 2012. “Contrasting Properties of Gene-Specific Regulatory, Coding, and Copy Number Mutations in Saccharomyces Cerevisiae: Frequency, Effects, and Dominance.” *PLoS Genetics* 8 (2): e1002497.

Gu, Xun, and Zhixi Su. 2007. “Tissue-Driven Hypothesis of Genomic Evolution and Sequence-Expression Correlations.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (8): 2779–84.

Haerty, Wilfried, Santosh Jagadeeshan, Rob J. Kulathinal, Alex Wong, Kristipati Ravi Ram, Laura K. Sirot, Lisa Levesque, et al. 2007. “Evolution in the Fast Lane: Rapidly Evolving Sex-Related Genes in Drosophila.” *Genetics* 177 (3): 1321–35.

Haerty, Wilfried, and Rama S. Singh. 2006. “Gene Regulation Divergence Is a Major Contributor to the Evolution of Dobzhansky-Muller Incompatibilities between Species of Drosophila.” *Molecular Biology and Evolution* 23 (9): 1707–14.

Hill, Mark S., Pétra Vande Zande, and Patricia J. Wittkopp. 2021. “Molecular and Evolutionary Processes Generating Variation in Gene Expression.” *Nature Reviews. Genetics* 22 (4): 203–15.

Hodgins, Kathryn A., Sam Yeaman, Kristin A. Nurkowski, Loren H. Rieseberg, and Sally N. Aitken. 2016. “Expression Divergence Is Correlated with Sequence Evolution but Not Positive Selection in Conifers.” *Molecular Biology and Evolution* 33 (6): 1502–16.

Hollis, Brian, Mareike Koppik, Kristina U. Wensing, Hanna Ruhmann, Eléonore Genzoni, Berra Erkosar, Tadeusz J. Kawecki, Claudia Fricke, and Laurent Keller. 2019. “Sexual Conflict Drives Male Manipulation of Female Postmating Responses in Drosophila Melanogaster.” *Proceedings of the National Academy of Sciences of the United States of America* 116 (17): 8437–44.

Holloway, Alisha K., and David J. Begun. 2004. “Molecular Evolution and Population Genetics of Duplicated Accessory Gland Protein Genes in Drosophila.” *Molecular Biology and Evolution* 21 (9): 1625–28.

Hopkins, Ben R., Irem Sepil, Marie-Laëtitia Thézénas, James F. Craig, Thomas Miller, Philip D. Charles, Roman Fischer, et al. 2019. “Divergent Allocation of Sperm and the Seminal Proteome along a Competition Gradient in *Drosophila Melanogaster*.” *Proceedings of the National Academy of Sciences of the United States of America* 116 (36): 17925–33.

Hurtado, Juan, Francisca Cunha Almeida, Silvina Anahí Belliard, Santiago Revale, and Esteban Hasson. 2022. “Research Gaps and New Insights in the Evolution of Drosophila Seminal Fluid Proteins.” *Insect Molecular Biology* 31 (2): 139–58.

Imhof, M., B. Harr, G. Brem, and C. Schlötterer. 1998. “Multiple Mating in Wild Drosophila Melanogaster Revisited by Microsatellite Analysis.” *Molecular Ecology* 7 (7): 915–17.

Kayserili, Melek A., Dave T. Gerrard, Pavel Tomancak, and Alex T. Kalinka. 2012. “An Excess of Gene Expression Divergence on the X Chromosome in Drosophila Embryos: Implications for the Faster-X Hypothesis.” *PLoS Genetics* 8 (12): e1003200.

Kimura, M. 1980. “A Simple Method for Estimating Evolutionary Rates of Base Substitutions through Comparative Studies of Nucleotide Sequences.” *Journal of Molecular Evolution* 16 (2): 111–20.

Kryuchkova-Mostacci, Nadezda, and Marc Robinson-Rechavi. 2015. “Tissue-Specific Evolution of Protein Coding Genes in Human and Mouse.” *PloS One* 10 (6): e0131673.

Kuleshov, Maxim V., Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, et al. 2016. “Enrichr: A Comprehensive Gene Set Enrichment Analysis Web Server 2016 Update.” *Nucleic Acids Research* 44 (W1): W90–97.

Landry, Christian R., Patricia J. Wittkopp, Clifford H. Taubes, Jose M. Ranz, Andrew G. Clark, and Daniel L. Hartl. 2005. “Compensatory Cis-Trans Evolution and the Dysregulation of Gene Expression in Interspecific Hybrids of Drosophila.” *Genetics* 171 (4): 1813–22.

Larson, Erica L., Dan Vanderpool, Sara Keeble, Meng Zhou, Brice A. J. Sarver, Andrew D. Smith, Matthew D. Dean, and Jeffrey M. Good. 2016. “Contrasting Levels of Molecular Evolution on the Mouse X Chromosome.” *Genetics* 203 (4): 1841–57.

Leader, David P., Sue A. Krause, Aniruddha Pandit, Shireen A. Davies, and Julian A. T. Dow. 2018. “FlyAtlas 2: A New Version of the Drosophila Melanogaster Expression Atlas with RNA-Seq, miRNA-Seq and Sex-Specific Data.” *Nucleic Acids Research* 46 (D1): D809–15.

Lemos, Bernardo, Luciana O. Araripe, Pierre Fontanillas, and Daniel L. Hartl. 2008. “Dominance and the Evolutionary Accumulation of Cis- and Trans-Effects on Gene Expression.” *Proceedings of the National Academy of Sciences of the United States of America* 105 (38): 14471–76.

Leung, Danny, Inkyung Jung, Nisha Rajagopal, Anthony Schmitt, Siddarth Selvaraj, Ah Young Lee, Chia-An Yen, et al. 2015. “Integrative Analysis of Haplotype-Resolved Epigenomes across Human Tissues.” *Nature* 518 (7539): 350–54.

Liang, Cong, Jacob M. Musser, Alison Cloutier, Richard O. Prum, and Günter P. Wagner. 2018. “Pervasive Correlated Evolution in Gene Expression Shapes Cell and Tissue Type Transcriptomes.” *Genome Biology and Evolution* 10 (2): 538–52.

Llopart, Ana. 2012. “The Rapid Evolution of X-Linked Male-Biased Gene Expression and the Large-X Effect in Drosophila Yakuba, D. Santomea, and Their Hybrids.” *Molecular Biology and Evolution* 29 (12): 3873–86.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12): 550.

Lu, Xuemei, Joshua A. Shapiro, Chau-Ti Ting, Yan Li, Chunyan Li, Jin Xu, Huanwei Huang, et al. 2010. “Genome-Wide Misexpression of X-Linked versus Autosomal Genes Associated with Hybrid Male Sterility.” *Genome Research* 20 (8): 1097–1102.

Mackay, Trudy F. C., Stephen Richards, Eric A. Stone, Antonio Barbadilla, Julien F. Ayroles, Dianhui Zhu, Sònia Casillas, et al. 2012. “The Drosophila Melanogaster Genetic Reference Panel.” *Nature* 482 (7384): 173–78.

Mack, Katya L., Polly Campbell, and Michael W. Nachman. 2016. “Gene Regulation and Speciation in House Mice.” *Genome Research* 26 (4): 451–61.

Majane, Alex C., Julie M. Cridland, and David J. Begun. 2022. “Single-Nucleus Transcriptomes Reveal Evolutionary and Functional Properties of Cell Types in the Drosophila Accessory Gland.” *Genetics* 220 (2). https://doi.org/10.1093/genetics/iyab213.

McDonald, J. H., and M. Kreitman. 1991. “Adaptive Protein Evolution at the Adh Locus in Drosophila.” *Nature* 351 (6328): 652–54.

McManus, C. Joel, Joseph D. Coolon, Michael O. Duff, Jodi Eipper-Mains, Brenton R. Graveley, and Patricia J. Wittkopp. 2010. “Regulatory Divergence in Drosophila Revealed by mRNA-Seq.” *Genome Research* 20 (6): 816–25.

Meiklejohn, Colin D., John Parsch, José M. Ranz, and Daniel L. Hartl. 2003. “Rapid Evolution of Male-Biased Gene Expression in *Drosophila*.” *Proceedings of the National Academy of Sciences of the United States of America* 100 (17): 9894–99.

Meisel, Richard P., John H. Malone, and Andrew G. Clark. 2012. “Faster-X Evolution of Gene Expression in Drosophila.” *PLoS Genetics* 8 (10): e1003013.

Michalak, Pawel, and Mohamed A. F. Noor. 2003. “Genome-Wide Patterns of Expression in Drosophila Pure Species and Hybrid Males.” *Molecular Biology and Evolution* 20 (7): 1070–76.

———. 2004. “Association of Misexpression with Sterility in Hybrids of Drosophila Simulansand D. Mauritiana.” *Journal of Molecular Evolution* 59 (2): 277–82.

Moehring, Amanda J., Katherine C. Teeter, and Mohamed A. F. Noor. 2007. “Genome-Wide Patterns of Expression in Drosophila Pure Species and Hybrid Males. II. Examination of Multiple-Species Hybridizations, Platforms, and Life Cycle Stages.” *Molecular Biology and Evolution* 24 (1): 137–45.

Mueller, J. L., K. Ravi Ram, L. A. McGraw, M. C. Bloch Qazi, E. D. Siggia, A. G. Clark, C. F. Aquadro, and M. F. Wolfner. 2005. “Cross-Species Comparison of Drosophila Male Accessory Gland Protein Genes.” *Genetics* 171 (1): 131–43.

Nair, Venugopalan D., Mital Vasoya, Vishnu Nair, Gregory R. Smith, Hanna Pincas, Yongchao Ge, Collin M. Douglas, Karyn A. Esser, and Stuart C. Sealfon. 2021. “Differential Analysis of Chromatin Accessibility and Gene Expression Profiles Identifies Cis-Regulatory Elements in Rat Adipose and Muscle.” *Genomics* 113 (6): 3827–41.

Pal, Soumitra, Brian Oliver, and Teresa M. Przytycka. 2021. “Modeling Gene Expression Evolution with EvoGeneX Uncovers Differences in Evolution of Species, Organs and Sexes.” *bioRxiv*. https://doi.org/10.1101/2020.01.06.895615.

Parisi, Michael, Rachel Nuttall, Pamela Edwards, James Minor, Daniel Naiman, Jining Lü, Michael Doctolero, et al. 2004. “A Survey of Ovary-, Testis-, and Soma-Biased Gene Expression in Drosophila Melanogaster Adults.” *Genome Biology* 5 (6): R40.

Patlar, Bahar, and Alberto Civetta. 2022. “Seminal Fluid Gene Expression and Reproductive Fitness in Drosophila Melanogaster.” *BMC Ecology and Evolution* 22 (1): 20.

Patlar, Bahar, Vivek Jayaswal, José M. Ranz, and Alberto Civetta. 2021. “Nonadaptive Molecular Evolution of Seminal Fluid Proteins in Drosophila.” *Evolution; International Journal of Organic Evolution* 75 (8): 2102–13.

Patro, Rob, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. 2017. “Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression.” *Nature Methods* 14 (4): 417–19.

Poiani, Aldo. 2006. “Complexity of Seminal Fluid: A Review.” *Behavioral Ecology and Sociobiology* 60 (3): 289–310.

Rabinow, L., and W. J. Dickinson. 1981. “A Cis-Acting Regulator of Enzyme Tissue Specificity in Drosophila Is Expressed at the RNA Level.” *Molecular & General Genetics: MGG* 183 (2): 264–69.

Racioppi, Claudia, Keira A. Wiechecki, and Lionel Christiaen. 2019. “Combinatorial Chromatin Dynamics Foster Accurate Cardiopharyngeal Fate Choices.” *eLife* 8 (November). https://doi.org/10.7554/eLife.49921.

Ranz, José M., Kalsang Namgyal, Greg Gibson, and Daniel L. Hartl. 2004. “Anomalies in the Expression Profile of Interspecific Hybrids of Drosophila Melanogaster and Drosophila Simulans.” *Genome Research* 14 (3): 373–79.

Ravi Ram, K., and Mariana F. Wolfner. 2007. “Sustained Post-Mating Response in Drosophila Melanogaster Requires Multiple Seminal Fluid Proteins.” *PLoS Genetics* 3 (12): e238.

Rice, P., I. Longden, and A. Bleasby. 2000. “EMBOSS: The European Molecular Biology Open Software Suite.” *Trends in Genetics: TIG* 16 (6): 276–77.

Romero, Irene Gallego, Ilya Ruvinsky, and Yoav Gilad. 2012. “Comparative Studies of Gene Expression and the Evolution of Gene Regulation.” *Nature Reviews. Genetics* 13 (7): 505–16.

Ronald, James, and Joshua M. Akey. 2007. “The Evolution of Gene Expression QTL in Saccharomyces Cerevisiae.” *PloS One* 2 (7): e678.

Ronald, James, Rachel B. Brem, Jacqueline Whittle, and Leonid Kruglyak. 2005. “Local Regulatory Variation in Saccharomyces Cerevisiae.” *PLoS Genetics* 1 (2): e25.

Sánchez-Ramírez, Santiago, Jörg G. Weiss, Cristel G. Thomas, and Asher D. Cutter. 2021. “Widespread Misregulation of Inter-Species Hybrid Transcriptomes due to Sex-Specific and Sex-Chromosome Regulatory Evolution.” *PLoS Genetics* 17 (3): e1009409.

Sanghi, Akshay, Joshua J. Gruber, Ahmed Metwally, Lihua Jiang, Warren Reynolds, John Sunwoo, Lisa Orloff, Howard Y. Chang, Maya Kasowski, and Michael P. Snyder. 2021. “Chromatin Accessibility Associates with Protein-RNA Correlation in Human Cancer.” *Nature Communications* 12 (1): 5732.

Schäfer, M., K. Nayernia, W. Engel, and U. Schäfer. 1995. “Translational Control in Spermatogenesis.” *Developmental Biology* 172 (2): 344–52.

Schully, Sheri Dixon, and Michael E. Hellberg. 2006. “Positive Selection on Nucleotide Substitutions and Indels in Accessory Gland Proteins of the Drosophila Pseudoobscura Subgroup.” *Journal of Molecular Evolution* 62 (6): 793–802.

Signor, Sarah A., and Sergey V. Nuzhdin. 2018. “The Evolution of Gene Expression in Cis and Trans.” *Trends in Genetics: TIG* 34 (7): 532–44.

Sirot, Laura K., Mariana F. Wolfner, and Stuart Wigby. 2011. “Protein-Specific Manipulation of Ejaculate Composition in Response to Female Mating Status in Drosophila Melanogaster.” *Proceedings of the National Academy of Sciences of the United States of America* 108 (24): 9922–26.

Sirot, Laura K., Alex Wong, Tracey Chapman, and Mariana F. Wolfner. 2014. “Sexual Conflict and Seminal Fluid Proteins: A Dynamic Landscape of Sexual Interactions.” *Cold Spring Harbor Perspectives in Biology* 7 (2): a017533.

Soneson, Charlotte, Michael I. Love, and Mark D. Robinson. 2015. “Differential Analyses for RNA-Seq: Transcript-Level Estimates Improve Gene-Level Inferences.” *F1000Research* 4 (1521): 1521.

Starks, Rebekah R., Anilisa Biswas, Ashish Jain, and Geetu Tuteja. 2019. “Combined Analysis of Dissimilar Promoter Accessibility and Gene Expression Profiles Identifies Tissue-Specific Genes and Actively Repressed Networks.” *Epigenetics & Chromatin* 12 (1): 16.

St Pierre, Celine L., Juan F. Macias-Velasco, Jessica P. Wayhart, Li Yin, Clay F. Semenkovich, and Heather A. Lawson. 2022. “Genetic, Epigenetic, and Environmental Mechanisms Govern Allele-Specific Gene Expression.” *Genome Research* 32 (6): 1042–57.

Stumm-Zollinger, E., and P. S. Chen. 1988. “Gene Expression in Male Accessory Glands of Interspecific Hybrids of Drosophila.” *Journal of Insect Physiology* 34 (1): 59–74.

Sturtevant, A. H. 1920. “Genetic Studies on DROSOPHILA SIMULANS. I. Introduction. Hybrids with DROSOPHILA MELANOGASTER.” *Genetics* 5 (5): 488–500.

Suyama, Mikita, David Torrents, and Peer Bork. 2006. “PAL2NAL: Robust Conversion of Protein Sequence Alignments into the Corresponding Codon Alignments.” *Nucleic Acids Research* 34 (Web Server issue): W609–12.

Swanson, Willie J., and Victor D. Vacquier. 2002. “The Rapid Evolution of Reproductive Proteins.” *Nature Reviews. Genetics* 3 (2): 137–44.

Tsaur, S. C., C. T. Ting, and C. I. Wu. 1998. “Positive Selection Driving the Evolution of a Gene of Male Reproduction, Acp26Aa, of Drosophila: II. Divergence versus Polymorphism.” *Molecular Biology and Evolution* 15 (8): 1040–46.

Vicoso, Beatriz, and Brian Charlesworth. 2006. “Evolution on the X Chromosome: Unusual Patterns and Processes.” *Nature Reviews. Genetics* 7 (8): 645–53.

Wagstaff, Bradley J., and David J. Begun. 2005. “Molecular Population Genetics of Accessory Gland Protein Genes and Testis-Expressed Genes in Drosophila Mojavensis and D. Arizonae.” *Genetics* 171 (3): 1083–1101.

Wang, Mi, Severin Uebbing, and Hans Ellegren. 2017. “Bayesian Inference of Allele-Specific Gene Expression Indicates Abundant Cis-Regulatory Variation in Natural Flycatcher Populations.” *Genome Biology and Evolution* 9 (5): 1266–79.

Warnefors, Maria, and Henrik Kaessmann. 2013. “Evolution of the Correlation between Expression Divergence and Protein Divergence in Mammals.” *Genome Biology and Evolution* 5 (7): 1324–35.

Wei, Kevin H-C, Andrew G. Clark, and Daniel A. Barbash. 2014. “Limited Gene Misregulation Is Exacerbated by Allele-Specific Upregulation in Lethal Hybrids between Drosophila Melanogaster and Drosophila Simulans.” *Molecular Biology and Evolution* 31 (7): 1767–78.

Whittle, Carrie A., and Cassandra G. Extavour. 2019. “Selection Shapes Turnover and Magnitude of Sex-Biased Expression in Drosophila Gonads.” *BMC Evolutionary Biology* 19 (1): 60.

Wigby, Stuart, Nora C. Brown, Sarah E. Allen, Snigdha Misra, Jessica L. Sitnik, Irem Sepil, Andrew G. Clark, and Mariana F. Wolfner. 2020. “The Drosophila Seminal Proteome and Its Role in Postcopulatory Sexual Selection.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 375 (1813): 20200072.

Wilson, C., A. Leiblich, D. C. I. Goberdhan, and F. Hamdy. 2017. “Chapter Eleven: The Drosophila Accessory Gland as a Model for Prostate Cancer and Other Pathologies.” *Current Topics in Developmental Biology* 121: 339–75.

Wittkopp, Patricia J., Belinda K. Haerum, and Andrew G. Clark. 2004. “Evolutionary Changes in Cis and Trans Gene Regulation.” *Nature* 430 (6995): 85–88.

———. 2008. “Regulatory Changes Underlying Expression Differences within and between Drosophila Species.” *Nature Genetics* 40 (3): 346–50.

Wittkopp, P. J. 2005. “Genomic Sources of Regulatory Variation in Cis and in Trans.” *Cellular and Molecular Life Sciences: CMLS* 62 (16): 1779–83.

Wong, Alex, Michael C. Turchin, Mariana F. Wolfner, and Charles F. Aquadro. 2008. “Evidence for Positive Selection on Drosophila Melanogaster Seminal Fluid Protease Homologs.” *Molecular Biology and Evolution* 25 (3): 497–506.

Yanai, Itai, Hila Benjamin, Michael Shmoish, Vered Chalifa-Caspi, Maxim Shklar, Ron Ophir, Arren Bar-Even, et al. 2005. “Genome-Wide Midrange Transcription Profiles Reveal Expression Level Relationships in Human Tissue Specification.” *Bioinformatics* 21 (5): 650–59.

Yang, Z. 1997. “PAML: A Program Package for Phylogenetic Analysis by Maximum Likelihood.” *Computer Applications in the Biosciences: CABIOS* 13 (5): 555–56.

Zhong, Xiuqin, Max Lundberg, and Lars Råberg. 2021. “Divergence in Coding Sequence and Expression of Different Functional Categories of Immune Genes between Two Wild Rodent Species.” *Genome Biology and Evolution* 13 (3). https://doi.org/10.1093/gbe/evab023.

Zhu, Anqi, Joseph G. Ibrahim, and Michael I. Love. 2019. “Heavy-Tailed Prior Distributions for Sequence Count Data: Removing the Noise and Preserving Large Differences.” *Bioinformatics* 35 (12): 2084–92.

Supplemental materials for Chapter I: Single-nucleus transcriptomes reveal functional and evolutionary properties of cell types in the Drosophila accessory gland

1. Supplemental results
2. Supplemental methods
3. Literature cited
4. Supplemental figures and tables

SUPPLEMENTAL RESULTS

Using all annotated *mel* genes, marker genes for each *mel* cell type reveal both expected and novel markers (Dataset S1). In MC we identify many expected Sfps including *SP* (Fig 1*D*), *Acp36DE*, *Acp26Aa*, and *Acp95EF,* and relatively uncharacterized Sfps including *Obp22a* (Fig. *1E*). The top non-Sfp markers of MC are generally functionally uncharacterized: *CG42852*, *CG43254*, *CG42481*, *CG43392*, *lncRNA:CR43146*, *lncRNA:CR45013*, *CG34041, lncRNA:TS14* (Fig. 1*E*), and the genes *CG44388* and *lncRNA:CR44389*, which are neighbors. Despite its annotation as a lncRNA, *CR44389* possesses a 41 amino acid ORF strongly predicted to have a signal sequence, suggesting it could be a secreted protein. *Ugt50B3,* a UDP-glycosyltransferase, is another strong marker of MC.

Among the 10 Sfps identified as SC markers (Table S1), three were previously known to be SC-specific: *Acp32CD*, *lectin-46Ca* and *lectin-46Cb* (Maeda et al. 2018)*.* Previous work with MC-null mutants identified *Acp32CD* as expressed in SC (Swanson et al. 2001), and here we additionally show that it exhibits very low expression in MC. The Sfps *CG17575, CG3349, CG9029*, *CG13695* (Fig. 1*E*), and *mfas* have also been previously identified as SC-expressed (Gligorov et al. 2013; Sitnik et al. 2016; Immarigeon et al. 2021). Here we show that these Sfps show very low expression in MC and EDC. We also identify the Sfp *Pgant9* as a novel SC marker. We additionally recovered expected non-Sfp markers: *lncRNA:iab8*, *abd-A (Maeda et al. 2018)*, *abd-B*, and *defective proventriculus* (*dve*) (Minami et al. 2012). We also identify non-Sfp SC markers *stranded-at-second* (*sas*) (Fig. 1*E*), *musashi* (*msi*), *form3*, *nahoda*, *CG31121*, *CG4629*, and *CG46430*. Additionally, we discovered that the unannotated transcript *DN2695* (see Identification of unannotated candidate genes in the AG, Table 1, Fig. S6) is a strong SC marker.

We identified 24 Sfp EDC markers (Table S1). Of these, 1 had previously been identified as EDC-enriched: *Dup99B*, *Obp51a*, *Spn77Bc*, *Spn77Bb*, *Est-6*, *Gld*, *Anp*, *CG18258*, *CG5162*, *CG17242*, *CG5402, CG34034,* and *CG31704* (Takemori and Yamamoto 2009; Sepil et al. 2018; Samakovlis et al. 1991; Cavener 1985; Saudan et al. 2002). The remainder have not been previously identified as EDC-specific Sfps: *Treh*, *betaggt-I*, *Sfp93F* (Fig. 1*E*)*, 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* (*vvl*) (Junell et al. 2010) and *Abd-B (Gligorov et al. 2013)*. Novel EDC markers are *anion exchanger 2* (*Ae2*) (Fig. 1*E*), *axundead* (*axed*), *single-minded* (*sim*), *CG7720*, *CG43101*, *CG7342*, and *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.

SUPPLEMENTAL METHODS

Fly stocks and reproductive tract dissection

We used the following sequenced stocks to compare AG transcriptomes between 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*). All animals were raised on a cornmeal-molasses-agar medium at 25°C and 60% relative humidity, on a 12:12 light/dark cycle. For snRNA-Seq experiments, virgin male flies were collected and placed into fresh vials of food in groups of five males per vial. On the day of the experiment, 2-3 day old virgin males were anaesthetized with CO_2 and their accessory glands plus anterior ejaculatory duct were dissected in cold 1X PBS and moved to 1X PBS + 2% BSA (Sigma SRE0036) on ice. Animals were dissected between Zeitgeber time (ZT) 0:30 and 2:30. Five animals from each species (derived from a single vial of food) were dissected before moving to the next species, in a repeating pattern, to prevent biasing interspecific sampling due to potential effects of circadian rhythms or the dissection protocol. Together, 23 *mel* (13 three day old, 10 two day old), 26 *sim* (15 three day old, 11 two day old), and 24 *yak* (14 three day old, 10 two day old) were dissected, and tissue from all three species was pooled together into a single microfuge tube.

Nuclear isolation and purification

Our nuclear isolation protocol is based on Luciano Martelotto's (2019) with some modification. For all steps we used non-stick, nuclease-free polypropylene tubes and pipette tips. All plastics, glassware, and filters were rinsed with PBS + 2% BSA before use. Pooled reproductive tract tissue was washed once with PBS + 2% BSA, then resuspended in 1 mL of lysis buffer (10 mM Tris-HCl pH 7.4, 10mM NaCl, 3 mM $MgCl_2$) with 0.1% NP40 (Sigma 9002-93-1). The tissue and lysis buffer were moved to a glass dounce homogenizer on ice and incubated for five minutes, with intermittent low-speed vortexing. The tissue was then dounce homogenized with 25 strokes with pestle A (loose fit) and incubated for an additional 10 minutes with intermittent low-speed vortexing. The tissue was next dounce homogenized with 25 strokes with pestle B (tight fit). 0.5 mL of PBS + 2% BSA was added to the homogenate, and the mixture was triturated 15 times with a silanized fire-polished glass pasteur pipette (BrainBits FPP). The homogenate was filtered through a 35 micron mesh (Falcon 352235) to remove unlysed tissue, and an additional 100 uL of PBS + 2% BSA was rinsed through the filter afterwards. The filtrate was centrifuged at 500 rcf for eight minutes at 4°C. The supernatant was removed, 200 uL of PBS + 2% BSA was added to the pellet, and the pellet was incubated on ice for five minutes. The pellet was then resuspended before an additional 800 uL of PBS + 2% BSA was mixed into the solution. Nuclei were centrifuged again at 500 rcf for eight minutes at 4°C, the supernatant removed, and the pellet was incubated in 200 uL of PBS + 2% BSA with 10 ug/mL DAPI on ice for five minutes before being resuspended.

FACS and snRNA-Seq

Following isolation of semi-pure nuclei, we used Fluorescence Activated Cell Sorting (FACS) to further purify singlet nuclei from clumps and cell debris. FACS was gated using DAPI and singlet nuclei were sorted into a new tube. Library preparation and sequencing was performed by the UC Davis Genome Center DNA Technologies & Expression Analysis Core Laboratory. The 10X Genomics 3' Single Cell (v2) kit was used to create libraries for snRNA-Seq on purified nuclei, and libraries were then sequenced on one lane of an Illumina HiSeq4000.

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

We used custom perl scripts (github.com/alexmajane/AG_single_nucleus) to identify reads originating from droplets containing nuclei vs. those that were empty and therefore composed entirely of ambient background RNA. For 10X Chromium v3, the R1 fastq contains droplet barcode and UMI data, while the R2 fastq contains cDNA sequence (Zheng et al. 2017). We used the R1 fastq file to parse counts of unique molecular identifiers (UMIs) per droplet barcode. After gathering this count data, we examined the distribution of UMIs per barcode in descending rank-order, to identify the ‘knee’ inflection point separating true cell-containing barcodes from barcodes associated with empty droplets (Macosko et al. 2015). For initial alignment of reads we selected all barcodes above the inflection point, which we expected was an overestimate of the true number of singlet nuclei, prior to further downstream filtering of low-UMI nuclei and multiplet-containing droplets. Given that single-nucleus RNA-Seq typically has a higher background RNA content than single-cell RNA-Seq (Alvarez et al. 2020), we wanted to profile this RNA background for later downstream bioinformatic correction. To profile background RNA, we selected 1,000 empty droplets at random from rank-orders below the inflection point.

We used an alignment-based approach to determine species-of-origin for each cell. Our goal was to use natural genetic variation among the three species. Our approach does not attempt to assign species identity at the level of individual reads, rather, we assign species identity to each nucleus, considered as a population of reads, summarized by the overall alignment rate to each species genome of the reads corresponding to each nucleus. While many individual reads might, considered in isolation, not be assignable to species, since we consider the entire set of reads in a nucleus together for species identification, we do not have to filter out any individual reads within a nucleus based on sequence divergence. Divergence was sufficient to effectively assign species identity while filtering out 4.8% of cells due to insufficient alignment bias. For each cell we aligned all reads to each species’ genome (*D. melanogaster* version 6.19, *D. simulans versio*n 2.02, *D. yakuba* version 1.05) respectively, using Hisat2 v2.1.0 (D. Kim et al. 2019). We sampled five million aligned reads and retained those with a best alignment match to a single species, dropping reads that aligned best equally well to two or more species. We used this subset of aligned reads to determine the percentage of reads originating from each species for each barcode. We selected those barcodes where ≥

50% of the reads aligned best to a single genome and categorized them as barcodes for that species. Next, we selected the R1 and R2 reads corresponding to barcodes associated with putative nuclei, and a new set of fastq files was created for each barcode, removing duplicated UMIs from these files. For empty droplets, we selected the R1 and R2 reads corresponding to each of the randomly sampled barcodes and created new fastq file pairs for each empty droplet. We aligned raw R2 reads to the appropriate species' genome (*D. melanogaster* version 6.33, *D. simulans versio*n 2.02, *D. yakuba* version 1.05) using STAR v2.7.5a (Dobin et al. 2013) with default parameters. We removed all transcripts of the gene *mod(mdg4)* from the *D. melanogaster* GTF file used for alignment, as this gene has trans-spliced transcripts without meaningful strand data, which causes a STAR error. We parsed the STAR logfile of each cell for percent of reads unmapped, percent of reads multi-mapped, and number of total mapped reads. We then used an R script to visualize these data and choose cut-offs to filter probable multiplets, as well as filtering low-quality/high background nuclei from the data.

In our first pass analysis of the data, we discovered an apparent lack of *Abd-B* expression in secondary cells of *sim* and *yak*, which was highly unexpected due to the central role of *Abd-B* in *mel* secondary cell development. Investigating possible technical explanations, we found discrepancies in the completeness of exon annotation across the three species, with the annotated *Abd-B* orthologues in *sim* and *yak* significantly shorter than that of *mel*. Since the Chromium library prep produces reads beginning around 300-400 bases from 3' end of expressed transcripts, rather than across the length of the transcript as in typical bulk mRNA-Seq library preps, differences in exon annotation of orthologous genes, especially at the 3' end, may lead to artifactual inference of DE across species. To investigate this possibility we used de novo transcriptome assemblies from paired-end bulk tissue mRNA-Seq samples to improve the annotation of all AG-expressed transcripts, prior to counting features in our aligned single-nucleus data. Transcripts for *D. simulans* (*w^{501}*) and *D. yakuba* (Tai18E2) were generated using Trinity v2.11.0 (Grabherr et al. 2011). We then aligned these transcripts back to the appropriate

species genome and transcriptome to identify assembled transcripts that aligned to existing genes. Records for these new transcripts were added to the species' GTF files from FlyBase. Custom GTF files with updated *sim* and *yak* exon annotation can be found at github.com/alexmajane/AG_single_nucleus. Following this improved annotation, count data exhibited similar *Abd-B* expression in secondary cells across species, as expected.
Next, we counted features from BAM files using HTSeq-count v0.12.3 (Anders, Pyl, and Huber 2015) with default parameters. Given that intronic reads may be included in single-nucleus RNA-Seq (Lake et al. 2016), we counted reads mapping to either exons or introns. Empty droplets were treated similarly to nuclei and were independently processed using each of the three species' annotations. Although the "true" background RNA profile is not expected to vary by the species of the nucleus contained in each droplet, the alignment and feature counting steps will filter out some number of reads of discordant species origin that are sufficiently divergent, so background profiles at this step will depend on the species-specific genome and gene models used.

To remove background RNA from nuclei, we used the R package SoupX v1.4.8 (Young and Behjati 2020). We performed preliminary cell type clustering of uncorrected data to identify marker genes in Seurat v3.2.2 (Stuart et al. 2019), the basis for the background correction algorithm used by SoupX. We also did additional filtering of high UMI-count cells to remove outliers (potential doublets). This final filtering left us with 1167 *mel*, 2115 *sim*, and 989 *yak* nuclei. We performed preliminary Seurat analysis independently for each species using its full set of genes (not limited to 1-to-1-to-1 orthologues). The set of empty drops aligned to each respective species was used as the basis for estimation of the background transcriptome, and background correction was independently performed for each species.

For comparative analyses we created a set of 1-to-1-to-1 orthologues (11,481 genes) using the *D. melanogaster* orthologue table from Flybase (2020 version 2). We used an R script to identify the set of *mel* orthologues with single orthologues identified in *sim* and *yak* respectively. To maximize the number of Sfps included in our comparative analysis we compiled the unique set of Sfps published from two proteomic studies, Findlay *et al*. (2009), and Sepil *et al*. (2018). Of these 264 Sfps, 77 were not in our set of annotated 1-to-1-to-1 orthologues. Of these, we were able to manually curate 25 novel 1-to-1-to-1 orthologous genes (Table S6) using tblastx (Camacho et al. 2009) and Ensembl Metazoa genome browsers (Howe et al. 2020) to confirm synteny.

Marker gene identification and differential expression among species

All analyses of single-nucleus gene expression data were performed in R v3.6.1 using Seurat v3.2.2 (Satija et al. 2015; Butler et al. 2018; Stuart et al. 2019). We used two parallel approaches. We did an integrated analysis of the data across species, using our set of *mel*, *sim*, and *yak* 1-to-1-to-1 orthologues. We also performed an independent analysis of *mel* using all annotated genes to get a fuller picture of differences in gene expression among cell types. For both datasets we used a fairly standard Seurat workflow. We log-normalized UMI counts and z-score transformed counts on a gene-by-gene basis. We selected the top 2000 variable genes for downstream analysis by ranking their dispersion values. We chose not to use the usual variance-stabilizing transformation (VST) method, which normalizes dispersion by absolute expression to account for noise inherent in single-cell RNA-Seq (Hafemeister and Satija 2019). Our dataset has one dominant cell type (MC), and two rare cell types (SC and EDC). The nature of the data means that the most highly expressed, specific markers to MC tend not to show a very high variability over the entire dataset. As expected, VST did not include several important MC marker genes including *SP* and *Acp95EF*, while the dispersion method includes them. We clustered cells with the Shared Nearest Neighbor (SNN) algorithm and used UMAP dimensionality reduction (McInnes et al. 2018) to visualize clustering. 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 logFC requirement of 0.25. We filtered marker genes to those with Bonferroni-corrected p-values less than 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 percent 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 sampled only MC and estimated pseudotime using Slingshot (Street et al. 2018) and identified dynamically differentially expressed genes with tradeSeq (Van den Berge et al. 2020). Further details of the specific parameters and methods used to process data in Seurat can be found in our R scripts. We performed GO enrichment analyses for non-Sfp markers of each *mel* cell type (SC, EDC, MCsp1, MCsp2, and the complete population of MC considered jointly) using DAVID (Huang et al. 2009) with all expressed non-Sfps as a background gene list. We considered only non-Sfps because of the overwhelming impact of Sfps as a class on GO enrichment terms.

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. To compare the rate of qualitative expression divergence across cell types, we calculated ratios of DE genes at various $\log_2$ fold change (logFC) cut-offs across the three cell types, for each of the three 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 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 differentially expressed genes exhibiting a logFC greater than one in at least one cell type specific pairwise species contrast (out of a total of nine contrasts: three cell types X three species). We then calculated pairwise Pearson 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 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 (2019) from *D. melanogaster* (Fraïsse, Puixeu Sala, and Vicoso 2019) with *D. simulans* as the outgroup. The McDonald-Kreitman test (McDonald and Kreitman 1991) compares the ratios of polymorphic and fixed synonymous amino acid substitutions to nonsynonymous substitutions. The summary statistic α (Smith and Eyre-Walker 2002) represents an estimate of excess fixed amino acid substitutions relative to the expectation under strict neutrality, describing the predicted proportion of amino acid substitutions resulting from directional selection. A positive value of α suggests directional selection acting on a given gene. Among positive α values, a greater value for a given set of genes suggests a greater proportion of amino acid substitutions fixed under directional selection. We considered two summaries of the role of adaptation in protein divergence: 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 adaptive protein substitutions, 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 enrichment of expression in the AG and values of $\tau > 0.9$, indicative of highly AG-specific expression, $\alpha > 0.5$, and at least five 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

We first identified the set of cell barcodes corresponding to each of the three cell types (MC, SC, EDC) based on marker gene identification as described earlier. Next, we concatenated raw fastq files corresponding to each barcode together to create a set of reads originating from each population of cells. We used TrimGalore! v0.6.5 (github.com/FelixKrueger/TrimGalore) to prep the raw reads, followed by de novo transcriptome assembly using Trinity v2.11.0 (Grabherr et al. 2011). Since transcript content is only contained in the R2 read of 10X Chromium libraries, our data is effectively single-end, so we used the '--single' and '--SS_lib_type F' options to Trinity.

We used a BLAST-based strategy (Camacho et al. 2009) to identify candidate unannotated transcripts in *D. melanogaster*. First, using blastn we aligned all transcripts to two custom databases, A) the sequences of all annotated genes in *mel*, *sim*, and *yak*, and B) the whole-genome sequences of all three species. We aligned the *mel* de novo transcriptome to all three species' concatenated databases to avoid having background transcripts from *sim* or *yak* identified as unannotated in *mel*. We then took the set of transcripts that had at least one BLAST hit to the genomic database, but no BLAST hits to the database of annotated genes. In the course of this analysis we discovered template-switching oligo (TSO) concatemer occurrence at the 5' end of some transcripts. TSO concatemers in 10X Chromium libraries have been previously documented (Svensson 2017). We removed the sequence (AAGCAGTGGTATCAACGCAGAGTACATGGG) from the 5' end of assembled transcripts. We also checked the raw data for TSO occurrence and found that these sequences exclusively occur at the 5' end, precluding the possibility of artifactual chimeric transcripts formed during Trinity assembly. Additionally, these concatemers are not expected to affect alignment of reads to the genome since STAR automatically trims non-aligning adapter sequences (Dobin et al. 2013).

We used counts from Salmon v0.12.0 (Patro et al. 2017) to filter the set of assembled transcripts to those expressed at a level of counts greater than 10% of the total number of cells in at least one cell type, to filter out very-lowly expressed transcripts. We removed poly-A sequence from the transcripts, and we aligned these filtered transcripts back to a blastn database containing solely the *mel* genome to remove background transcripts from other species. We converted tabular BLAST output to a GTF file using a python script and used HTSeq to do feature counting on individual cell transcriptomes aligned with STAR (described earlier), as well as background drop transcriptomes. We used SoupX to remove background contamination as described above, with the global contamination fraction estimated from earlier all-genes analysis.

Given the single-ended, 3'-biased nature of the 10X library prep, full-length transcript assemblies cannot be confidently constructed. Therefore, we augmented this analysis by making a de novo transcriptome assembly for the FlyAtlas2 (Leader et al. 2018) accessory gland data using Trinity (Grabherr et al. 2011). We identified homologous FlyAtlas2 transcripts using BLAST. These transcripts were used to improve the completeness of our candidate transcripts and eliminate a few candidates that overlapped annotated genes. Similarly, for SC-biased transcripts we used bulk RNA-Seq data from FACS-sorted secondary cells (Immarigeon et al. 2021) to improve our annotations. Both supplementary RNA-Seq datasets are unstranded, so we obtained strand information from our 10X-based transcript assemblies. We used FlyAtlas2 data to improve the annotation of *DN8354, DN35169, DN10930, DN16089, DN2736, DN5813,* and *DN818*. We found no matching FlyAtlas2 transcripts for *DN4707*, *DN11110*, and *DN10097*. We found a single, very lowly expressed FlyAtlas2 transcript homologous to *DN2695*, but the 10X transcript was longer than the FlyAtlas2 transcript. We searched for transcripts homologous to *DN2695* and *DN10097* in FACS-sorted secondary cell RNA-Seq data (Immarigeon et al. 2021). We found matching transcripts in this dataset expressed at significant levels: *DN2695* at a mean 20.8 TPM and *DN100097* at a mean 1.12 TPM. This improved our assembled *DN2965* transcript length from 463 to 2,176 bp, and took *DN10097* from 260 to 353 bp.

We used Ensembl Metazoa BLAST and the *mel* genome browser (Howe et al. 2020) to identify gene coordinates, strand, and neighboring annotated genes, and to verify that these candidate genes do not overlap with any existing annotated features. Though we selected our candidate gene sequences based on expression, Trinity groups sets of transcripts together based on shared k-mers—many of which are very lowly expressed. For each of the 11 candidates, we additionally verified that none of the clustered transcripts within the gene “family” overlapped with annotated features.

For cell type-specific analysis of unannotated gene expression, we added counts to the broader *mel* dataset post-hoc. We used Seurat’s FindAllMarkers() method to identify cell type bias of unannotated genes. To adjust the test for identification of cell type specific bias, rather than identification of canonical cell type markers, we relaxed the default thresholding requirements, including the requirement that a gene be expressed in at least 25% of focal cluster cells, and the minimal logFC requirement of 0.25. Significance of expression bias was assessed using a Wilcoxon Rank Sum test with Bonferroni multiple test correction. 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. We additionally assessed coding potential of our transcripts using CPAT v2.0.0 (L. Wang et al. 2013).

LITERATURE CITED

Almagro Armenteros, José Juan, Konstantinos D. Tsirigos, Casper Kaae Sønderby, Thomas Nordahl Petersen, Ole Winther, Søren Brunak, Gunnar von Heijne, and Henrik Nielsen. 2019. "SignalP 5.0 Improves Signal Peptide Predictions Using Deep Neural Networks." Nature Biotechnology 37 (4): 420–23.

Alvarez, Marcus, Elior Rahmani, Brandon Jew, Kristina M. Garske, Zong Miao, Jihane N. Benhammou, Chun Jimmie Ye, et al. 2020. "Enhancing Droplet-Based Single-Nucleus RNA-Seq Resolution Using the Semi-Supervised Machine Learning Classifier DIEM." *Scientific Reports* 10 (1): 11019.

Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber. 2015. "HTSeq--a Python Framework to Work with High-Throughput Sequencing Data." *Bioinformatics* 31 (2): 166–69.

Avila, Frank W., Allie B. Cohen, Fatima S. Ameerudeen, David Duneau, Shruthi Suresh, Alexandra L. Mattei, and Mariana F. Wolfner. 2015. "Retention of Ejaculate by Drosophila Melanogaster Females Requires the Male-Derived Mating Plug Protein PEBme." *Genetics* 200 (4): 1171–79.

Benjamini, Yoav, and Yosef Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 57 (1): 289–300.

Butler, Andrew, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. 2018. "Integrating Single-Cell Transcriptomic Data across Different Conditions, Technologies, and Species." *Nature Biotechnology* 36 (5): 411–20.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10 (December): 421.

Cavener, D. R. 1985. "Coevolution of the Glucose Dehydrogenase Gene and the Ejaculatory Duct in the Genus Drosophila." *Molecular Biology and Evolution* 2 (2): 141–49.

Cridland, Julie M., Alex C. Majane, Hayley K. Sheehy, and David J. Begun. 2020. "Polymorphism and Divergence of Novel Gene Expression Patterns in Drosophila Melanogaster." *Genetics* 216 (1): 79–93.

Dobin, Alexander, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. 2013. "STAR: Ultrafast Universal RNA-Seq Aligner." *Bioinformatics* 29 (1): 15–21.

Findlay, Geoffrey D., Michael J. MacCoss, and Willie J. Swanson. 2009. "Proteomic Discovery of Previously Unannotated, Rapidly Evolving Seminal Fluid Genes in Drosophila." *Genome Research* 19 (5): 886–96.

Fraïsse, Christelle, Gemma Puixeu Sala, and Beatriz Vicoso. 2019. "Pleiotropy Modulates the Efficacy of Selection in Drosophila Melanogaster." *Molecular Biology and Evolution* 36 (3): 500–515.

Gligorov, Dragan, Jessica L. Sitnik, Robert K. Maeda, Mariana F. Wolfner, and François Karch. 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 Genetics* 9 (3): e1003395.

Grabherr, Manfred G., Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, et al. 2011. “Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome.” *Nature Biotechnology* 29 (7): 644–52.

Hafemeister, Christoph, and Rahul Satija. 2019. “Normalization and Variance Stabilization of Single-Cell RNA-Seq Data Using Regularized Negative Binomial Regression.” *Genome Biology* 20 (1): 296.

Howe, Kevin L., Bruno Contreras-Moreira, Nishadi De Silva, Gareth Maslen, Wasiu Akanni, James Allen, Jorge Alvarez-Jarreta, et al. 2020. “Ensembl Genomes 2020-Enabling Non-Vertebrate Genomic Research.” *Nucleic Acids Research* 48 (D1): D689–95.

Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki. 2009a. “Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources.” *Nature Protocols* 4 (1): 44–57.

———. 2009b. “Bioinformatics Enrichment Tools: Paths toward the Comprehensive Functional Analysis of Large Gene Lists.” *Nucleic Acids Research* 37 (1): 1–13.

Immarigeon, Clément, Yohan Frei, Sofie Y. N. Delbare, Dragan Gligorov, Pedro Machado Almeida, Jasmine Grey, Léa Fabbro, et al. 2021. “Identification of a Micropeptide and Multiple Secondary Cell Genes That Modulate Drosophila Male Reproductive Success.” *Proceedings of the National Academy of Sciences of the United States of America* 118 (15).

Junell, Anna, Hanna Uvell, Monica M. Davis, Esther Edlundh-Rose, Asa Antonsson, Leslie Pick, and Ylva Engström. 2010. “The POU Transcription Factor Drifter/Ventral Veinless Regulates Expression of Drosophila Immune Defense Genes.” *Molecular and Cellular Biology* 30 (14): 3672–84.

Kim, Daehwan, Joseph M. Paggi, Chanhee Park, Christopher Bennett, and Steven L. Salzberg. 2019. “Graph-Based Genome Alignment and Genotyping with HISAT2 and HISAT-Genotype.” *Nature Biotechnology* 37 (8): 907–15.

Kruskal, William H., and W. Allen Wallis. 1952. “Use of Ranks in One-Criterion Variance Analysis.” *Journal of the American Statistical Association* 47 (260): 583–621.

Lake, B. B., R. Ai, G. E. Kaeser, N. S. Salathia, Y. C. Yung, R. Liu, A. Wildberg, et al. 2016. “Neuronal Subtypes and Diversity Revealed by Single-Nucleus RNA Sequencing of the Human Brain.” *Science* 352 (6293): 1586–90.

Leader, David P., Sue A. Krause, Aniruddha Pandit, Shireen A. Davies, and Julian A. T. Dow. 2018. “FlyAtlas 2: A New Version of the Drosophila Melanogaster Expression Atlas with RNA-Seq, miRNA-Seq and Sex-Specific Data.” *Nucleic Acids Research* 46 (D1): D809–15.

Lu, Shennan, Jiyao Wang, Farideh Chitsaz, Myra K. Derbyshire, Renata C. Geer, Noreen R. Gonzales, Marc Gwadz, et al. 2020. “CDD/SPARCLE: The Conserved Domain Database in 2020.” *Nucleic Acids Research* 48 (D1): D265–68.

Mackay, Trudy F. C., Stephen Richards, Eric A. Stone, Antonio Barbadilla, Julien F. Ayroles, Dianhui Zhu, Sònia Casillas, et al. 2012. “The Drosophila Melanogaster Genetic Reference Panel.” *Nature* 482 (7384): 173–78.

Macosko, Evan Z., Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, et al. 2015. “Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets.” *Cell* 161 (5): 1202–14.

Maeda, Robert K., Jessica L. Sitnik, Yohan Frei, Elodie Prince, Dragan Gligorov, Mariana F. Wolfner, and François Karch. 2018. “The lncRNA Male-Specific Abdominal Plays a Critical Role in Drosophila Accessory Gland Development and Male Fertility.” *PLoS Genetics* 14 (7): e1007519.

Martelotto, Luciano. 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.

McDonald, J. H., and M. Kreitman. 1991. “Adaptive Protein Evolution at the Adh Locus in Drosophila.” *Nature* 351 (6328): 652–54.

Minami, Ryunosuke, Miyuki Wakabayashi, Seiko Sugimori, Kiichiro Taniguchi, Akihiko Kokuryo, Takao Imano, Takashi Adachi-Yamada, Naoko Watanabe, and Hideki Nakagoshi. 2012. “The Homeodomain Protein Defective Proventriculus Is Essential for Male Accessory Gland Development to Enhance Fecundity in Drosophila.” *PloS One* 7 (3): e32302.

Patro, Rob, Geet Duggal, Michael I. Love, Rafael A. Irizarry, and Carl Kingsford. 2017. “Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression.” *Nature Methods* 14 (4): 417–19.

Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. “Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies.” *Nucleic Acids Research* 43 (7): e47.

Samakovlis, C., P. Kylsten, D. A. Kimbrell, A. Engström, and D. Hultmark. 1991. “The Andropin Gene and Its Product, a Male-Specific Antibacterial Peptide in Drosophila Melanogaster.” *The EMBO Journal* 10 (1): 163–69.

Satija, Rahul, Jeffrey A. Farrell, David Gennert, Alexander F. Schier, and Aviv Regev. 2015. “Spatial Reconstruction of Single-Cell Gene Expression Data.” *Nature Biotechnology* 33 (5): 495–502.

Saudan, Philippe, Klaus Hauck, Matthias Soller, Yves Choffat, Michael Ottiger, Michael Spörri, Zhaobing Ding, et al. 2002. “Ductus Ejaculatorius Peptide 99B (DUP99B), a Novel Drosophila Melanogaster Sex-Peptide Pheromone.” *European Journal of Biochemistry / FEBS* 269 (3): 989–97.

Sepil, Irem, Ben R. Hopkins, Rebecca Dean, Marie-Laëtitia Thézénas, Philip D. Charles, Rebecca Konietzny, Roman Fischer, Benedikt Kessler, and Stuart Wigby. 2018. “Quantitative Proteomics Identification of Seminal Fluid Proteins in Male Drosophila Melanogaster.” *Molecular & Cellular Proteomics: MCP*, October.

Sitnik, Jessica L., Dragan Gligorov, Robert K. Maeda, François Karch, and Mariana F. Wolfner. 2016. “The Female Post-Mating Response Requires Genes Expressed in the Secondary Cells of the Male Accessory Gland in Drosophila Melanogaster.” *Genetics* 202 (3): 1029–41.

Smith, Nick G. C., and Adam Eyre-Walker. 2002. “Adaptive Protein Evolution in Drosophila.” *Nature* 415 (6875): 1022–24.

Sokal, Robert R., and F. James Rohlf. 2012. *Biometry: The Principles and Practice of Statistics in Biological Research, 4th Ed*. New York, NY: Freeman and Co.

Street, Kelly, Davide Risso, Russell B. Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. 2018. “Slingshot: Cell Lineage and Pseudotime Inference for Single-Cell Transcriptomics.” *BMC Genomics* 19 (1): 477.

Stuart, Tim, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M. Mauck 3rd, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. 2019. “Comprehensive Integration of Single-Cell Data.” *Cell* 177 (7): 1888–1902.e21.

Svensson, Valentine. 2017. “Low Mapping Rate 3 - TSO Concatemers — What Do You Mean ‘Heterogeneity’?” September 8, 2017. http://www.nxn.se/valent/2017/9/8/low-mapping-rate-3-tso-concatemers.

Swanson, W. J., A. G. Clark, H. M. Waldrip-Dail, M. F. Wolfner, and C. F. Aquadro. 2001. “Evolutionary EST Analysis Identifies Rapidly Evolving Male Reproductive Proteins in Drosophila.” *Proceedings of the National Academy of Sciences of the United States of America* 98 (13): 7375–79.

Takemori, Nobuaki, and Masa-Toshi Yamamoto. 2009. “Proteome Mapping of the Drosophila Melanogaster Male Reproductive System.” *Proteomics* 9 (9): 2484–93.

Van den Berge, Koen, Hector Roux de Bézieux, Kelly Street, Wouter Saelens, Robrecht Cannoodt, Yvan Saeys, Sandrine Dudoit, and Lieven Clement. 2020. “Trajectory-Based Differential Expression Analysis for Single-Cell Sequencing Data.” *Nature Communications* 11 (1): 1201.

Wang, Liguo, Hyun Jung Park, Surendra Dasari, Shengqin Wang, Jean-Pierre Kocher, and Wei Li. 2013. “CPAT: Coding-Potential Assessment Tool Using an Alignment-Free Logistic Regression Model.” *Nucleic Acids Research* 41 (6): e74.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer.

Yanai, Itai, Hila Benjamin, Michael Shmoish, Vered Chalifa-Caspi, Maxim Shklar, Ron Ophir, Arren Bar-Even, et al. 2005. “Genome-Wide Midrange Transcription Profiles Reveal Expression Level Relationships in Human Tissue Specification.” *Bioinformatics* 21 (5): 650–59.

Young, Matthew D., and Sam Behjati. 2020. “SoupX Removes Ambient RNA Contamination from Droplet-Based Single-Cell RNA Sequencing Data.” *GigaScience* 9 (12). https://www.ncbi.nlm.nih.gov/pubmed/33367645.

Zheng, Grace X. Y., Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, et al. 2017. “Massively Parallel Digital Transcriptional Profiling of Single Cells.” *Nature Communications* 8 (January): 14049.

TABLES S1 – S4

Table S1. Sfp marker genes in *mel* SC and EDC, as described in Dataset S1. pct.1 refers to the fraction of focal cells expressing the marker, pct.2 refers to the fraction on non-focal cells expressing the marker. *p* is the result of a Wilcoxon rank-sum test with Bonferroni correction.

gene	cluster	avg logFC	pct.1	pct.2	*p*
CG14913	SC	3.109	0.780	0.001	5.37E-189
CG17575	SC	4.598	1.000	0.021	3.71E-172
CG3349	SC	1.711	0.320	0.001	6.09E-72
CG13965	SC	3.692	1.000	0.139	6.00E-68
lectin-46Ca	SC	4.504	1.000	0.169	2.11E-60
CG9029	SC	4.328	1.000	0.186	1.60E-56
Acp32CD	SC	4.606	1.000	0.207	3.82E-53
lectin-46Cb	SC	4.527	1.000	0.237	6.51E-49
mfas	SC	1.186	1.000	0.813	5.12E-16
Pgant9	SC	0.619	0.720	0.383	0.00102
Dup99B	EDC	6.457	1.000	0.002	1.32E-243
Obp51a	EDC	4.210	1.000	0.003	2.87E-239
Anp	EDC	4.879	1.000	0.003	3.40E-239
Spn77Bc	EDC	5.066	1.000	0.006	8.92E-226
CG18258	EDC	2.809	0.797	0.001	8.95E-194
CG43101	EDC	4.484	0.983	0.020	1.11E-181
CG17549	EDC	3.056	0.831	0.013	4.41E-161
CG42782	EDC	3.457	0.966	0.030	1.67E-150
CG33290	EDC	2.687	0.712	0.006	2.85E-150
CG5162	EDC	3.741	0.983	0.042	5.18E-142
CG17242	EDC	3.371	0.966	0.042	1.45E-136
CG5402	EDC	3.561	0.966	0.049	5.77E-129
CG15394	EDC	2.950	0.898	0.040	1.79E-124
Spn77Bb	EDC	5.154	1.000	0.084	8.22E-105
Est-6	EDC	3.979	0.983	0.082	1.59E-101
Treh	EDC	2.358	0.847	0.055	2.51E-93
Gld	EDC	2.441	0.763	0.046	4.62E-86
betaggt-I	EDC	2.132	0.559	0.017	1.14E-83
CG34034	EDC	4.023	1.000	0.175	3.77E-68
Sfp93F	EDC	3.023	0.966	0.161	8.77E-65
CG11590	EDC	1.99	0.508	0.023	1.92E-64
NT5E-2	EDC	1.118	0.254	0.009	7.80E-33
CG31704	EDC	1.937	0.983	0.450	1.13E-32
trx	EDC	0.491	0.356	0.106	0.00106

Table S2. p values (Benjamini-Hochberg corrected) of pairwise G-tests for percentage of expressed genes DE among cell types and species contrasts. ms = mel vs sim, my = mel vs yak, sy = sim vs yak; MC = main cells, SC = secondary cells, EDC = ejaculatory duct cells. Significant values in bold face.

	ms_MC	ms_SC	ms_ED	my_MC	my_SC	my_ED	sy_MC	sy_SC
ms_SC	0.750	-	-	-	-	-	-	-
ms_ED	**0.001**	**0.009**	-	-	-	-	-	-
my_MC	**0.027**	0.099	0.276	-	-	-	-	-
my_SC	0.061	0.171	0.272	0.872	-	-	-	-
my_ED	**0.000**	**0.000**	**0.001**	**0.000**	**0.000**	-	-	-
sy_MC	0.352	0.662	**0.018**	0.193	0.301	**0.000**	-	-
sy_SC	**0.027**	0.090	0.423	0.839	0.760	**0.000**	0.171	-
sy_ED	**0.000**	**0.000**	0.303	**0.031**	**0.037**	**0.031**	**0.000**	0.087

Table S3. Genes annotated as non-SFPs that are narrowly expressed in the AG and have strong evidence of adaptive protein substitutions between *mel* and *sim*. All are MC-biased in expression. Substitution counts are from Fraïsse et al. (2019). Expression values are from Leader et al. (2018), with τ calculated in Majane et al. (2020). *tau am* refers to τ in adult males. *ag exp* is the mean AG expression expressed in FPKM. *next tissue* is the adult male tissue with the next-greatest expression, and *next exp* is the value in FPKM. *max af* and *max larval* indicate the greatest level of expression in any adult female or larval tissue, respectively.

gene	pS	pN	dS	dN	alpha	tau am	ag exp	next tissue	next exp	max af	max larval
CG42852	5	1	1	4	0.950	0.995	158055	testis	7262	107	28
CG42789	9	3	6	5	0.600	0.997	933	testis	25	0.5	0.2
CAH16	17	7	28	37	0.688	0.998	144	testis	1.6	0.3	0
CG42470	14	10	2	14	0.898	0.998	269	fat body	6.1	0.2	0.2
CG34299	36	25	12	18	0.537	0.996	458	testis	14	0.2	0.1
CG43829	11	17	2	9	0.657	0.998	60	fat body	4.7	0.3	0.1
CG43354	4	2	2	6	0.833	0.991	3119	testis	259	3.3	1.2
CG43396	14	11	10	17	0.538	0.995	194	testis	4.7	2.2	0.4
CG4271	38	12	16	18	0.719	0.998	311	testis	4.3	55	0
CG31493	8	2	16	17	0.765	0.998	273	testis	3.1	0.3	0
CG43050	20	10	17	34	0.750	0.985	301	testis	45	0.5	0.7
CG42869	6	6	2	14	0.857	0.996	1101	testis	43	0.5	0

Table S4. Unannotated candidate genes expressed in the *D. melanogaster* accessory gland. Coordinates are from BLAST results to *D. melanogaster* version 6.36. Length refers to the span of BLAST coordinates. MC, SC, and EDC refer to the fraction of genes in each cell type with expression, respectively. Average logFC is the cell type with highest fraction of expression compared to the other two cell types. *p* is the result of a Wilcoxon Rank Sum test with Bonferroni correction.

transcript	coordinates	strand	length	neighbor genes (dist., strand)	expression bias	MC	SC	EDC	logFC	*p*
DN4707	**3R**:29346322-29346674	F	352	*lncRNA:CR45547* (9.4 kb, R); *Cnx99A* (26.7 kb, R)	broad	0.121	0.060	0.050	0.544	0.309
DN8354	**2R**:20149969-20150499	R	530	*lncRNA:CR45229* (0.2 kb, F); *CG1041* (3.2 kb, F)	broad	0.074	0.060	0.017	0.255	1
DN35169	**3R**:26745224-26745903	R	630	*CG14247* (6.6 kb, F); *lncRNA:CR45562* (11.4 kb, R)	broad / MC	0.141	0.060	0.050	0.595	0.087
DN10930	**3R**:13056814-13057676	R	863	*pic* (0.1 kb, F); *sim* (0.1 kb, F)	EDC	0.000	0.000	0.150	0.750	<0.001
DN16089	**3R**:14099143-14099715	R	572	*PK1*-R (0.1 kb, R); *CG9920* (1.1 kb, R)	EDC	0.000	0.000	0.183	0.718	<0.001
DN11110	**X**:9316379-9316731	F	352	*lncRNA:CR44534* (1.2kb, F); *c1.21* (1.2kb, R)	MC	0.149	0.020	0.017	0.923	0.001
DN2736	**2L**:9747046-9747785	R	739	*Cyp4e3* (0.05 kb, R); *Nckx30C* (0.5 kb, R)	MC	0.177	0.060	0.050	0.856	0.006
DN5813	**2R**:16922917-16924195	F	1278	*Pkc53E* (0.5 kb, F); *inaC* (21 kb, F)	MC	0.160	0.000	0.017	0.981	<0.001
DN818	**3R**:7341782-7344884	R	3102	*lncRNA:CR44337* (9.7 kb, F); *CG31496* (0.2 kb, F)	MC	0.353	0.080	0.167	1.170	<0.001
DN10097	**2L:**7597302-7597655	R	353	*lncRNA:CR45370* (6.6 kb, F); *Spn28B* (14 kb, F)	SC	0.002	0.100	0.000	0.826	<0.001
DN2695	**2L:**7601554-7603730	R	2176	*lncRNA:CR45370* (0.5 kb, F); *Spn28B* (18 kb, F)	SC	0.000	0.460	0.000	2.130	<0.001

FIGURES S1 – S7

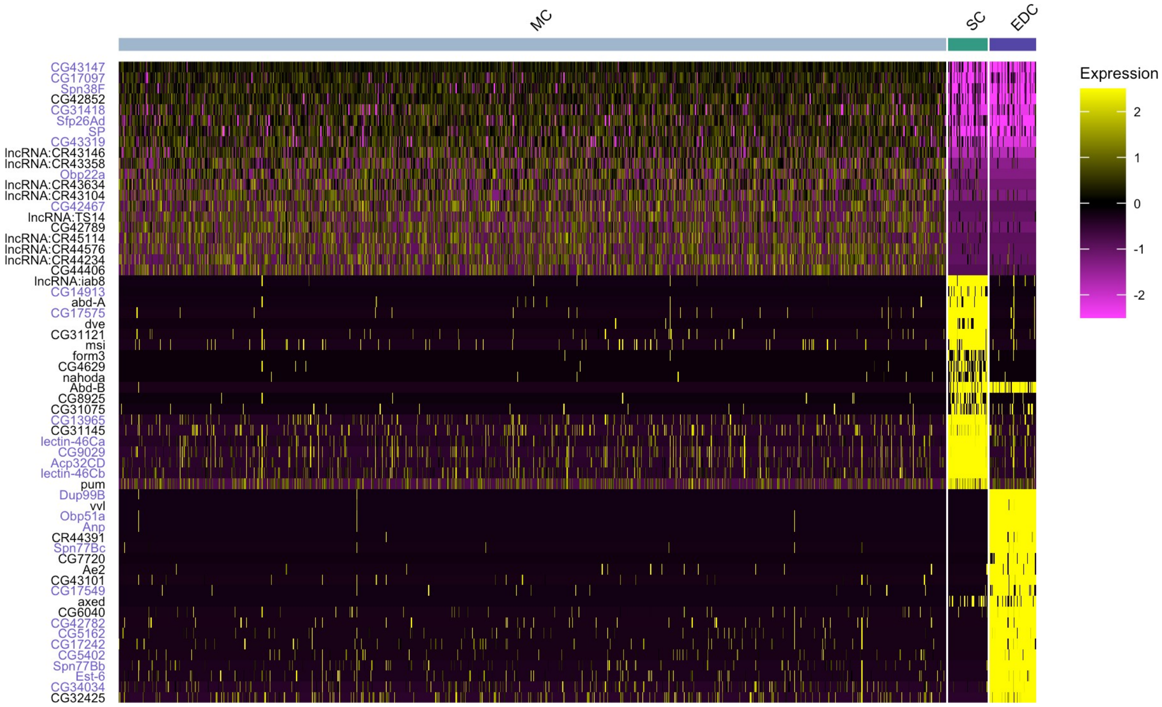

Figure S1. Heatmap showing scaled expression values for the complete set of *D. melanogaster* cells. Note that MC dominate the dataset (> 90% of total cells), as expected given the composition of the male reproductive tract. Since expression values are scaled cell-wise, the distributions of scaled expression for MC markers are strikingly different from SC or EDC markers. This is an artifact of the scaling process, which is resolved by downsampling to comparable numbers of nuclei across cell types, as we have done in Fig. 1*F*.

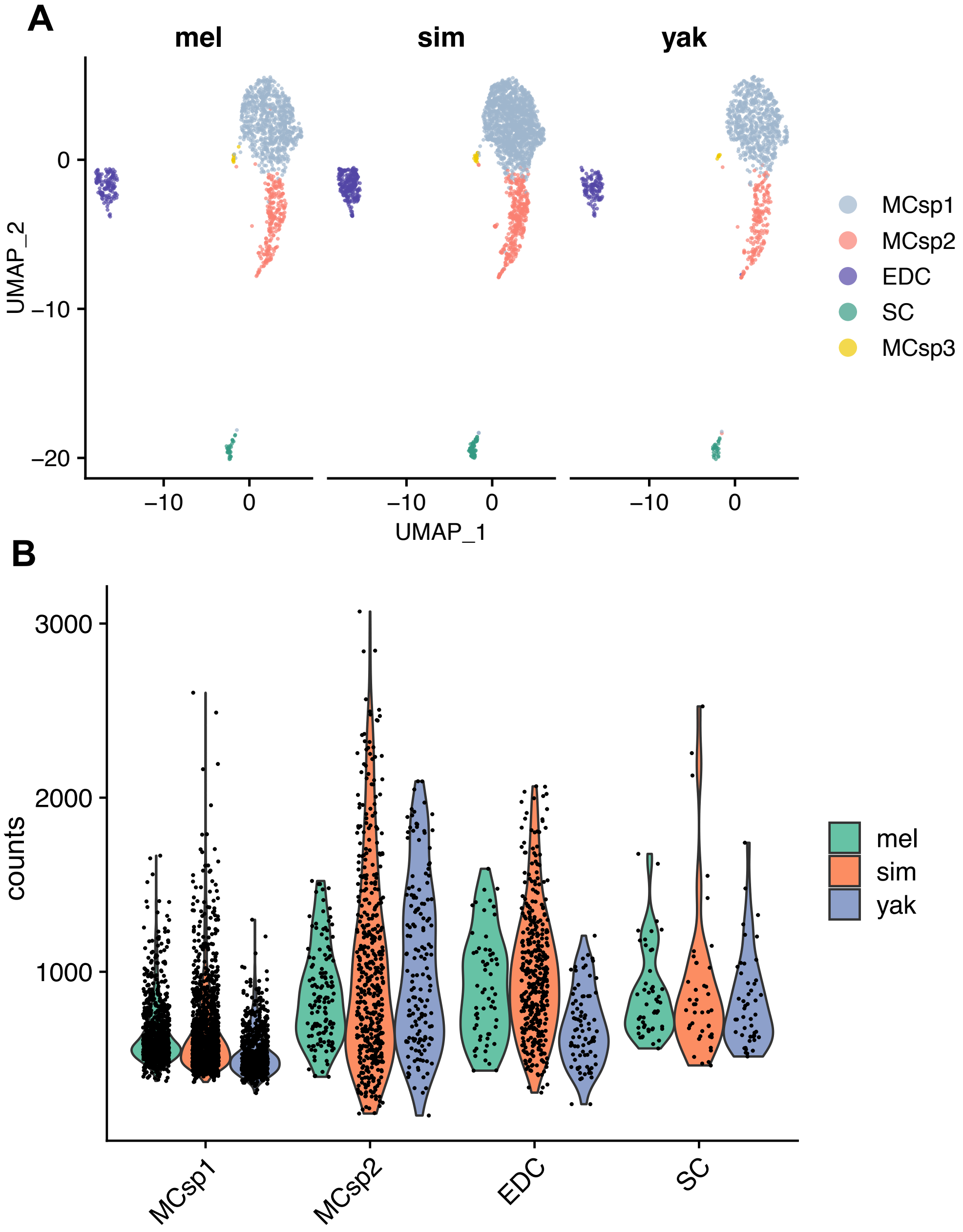

Figure S2. (*A*) Subpopulations of MC appear conserved among *mel*, *sim*, and *yak*. MCsp2, explored in depth in *melanogaster*, is conserved with roughly equal proportionality among species. The data suggest an additional subpopulation, MCsp3, but given its relatively weak support, it is not analyzed further here. (*B*) Decreased counts per nucleus in MCsp1 relative to MCsp2, SC and EDC are observed for all three species (Kruskall and pairwise Wilcoxon Rank Sum tests, $p < 0.001$). We have excluded MCsp3 here given the low number of observations.

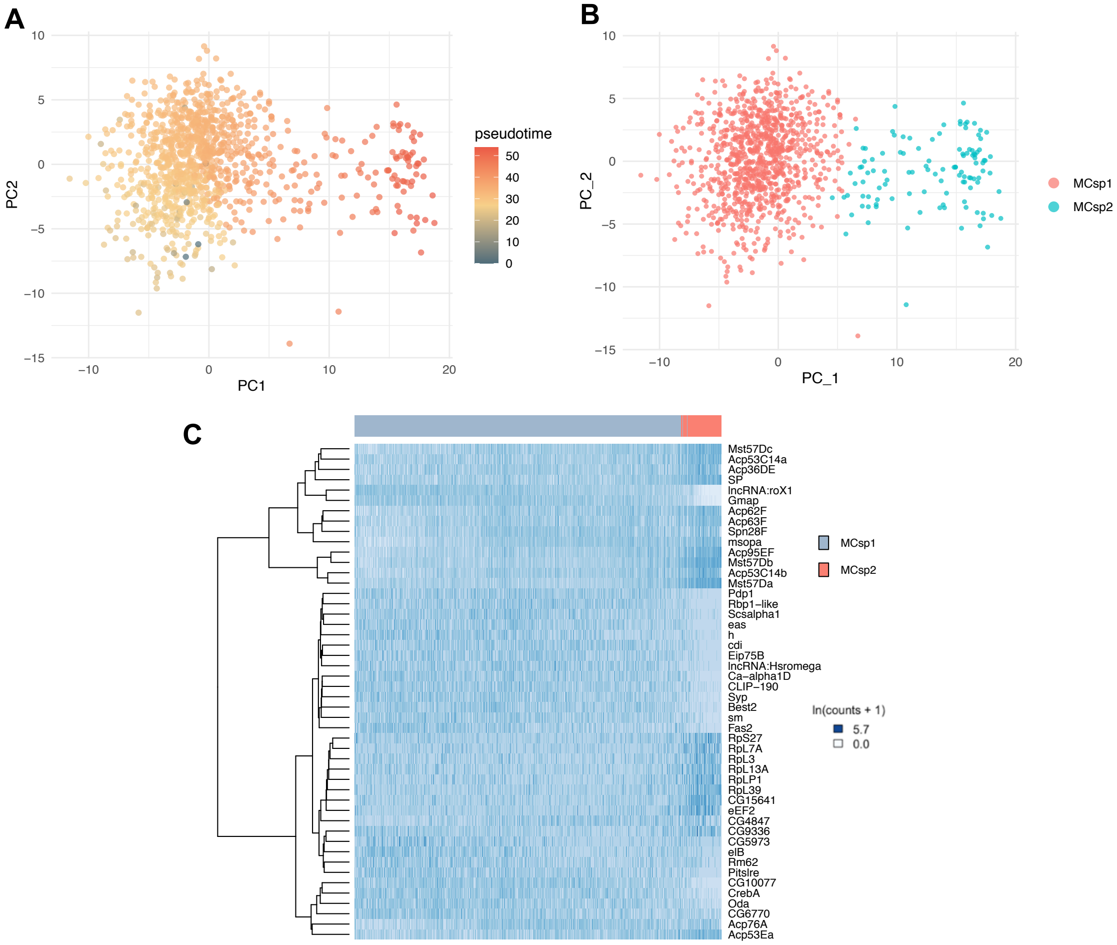

Figure S3. PCA plots of MC subpopulations, colored by (*A*) pseudotime, and (*B*) color estimated in Slingshot. (*C*) The top 50 dynamically DE genes identified by tradeSeq. Expression is displayed as ln(counts + 1), with intensity of blue color indicating relative expression; darker cells are more highly expressed. The dendogram indicates Euclidean distance among genes. Nuclei are ordered by pseudotime from left to right column-wise, and cluster identity is indicated in the bar at the top of the plot.

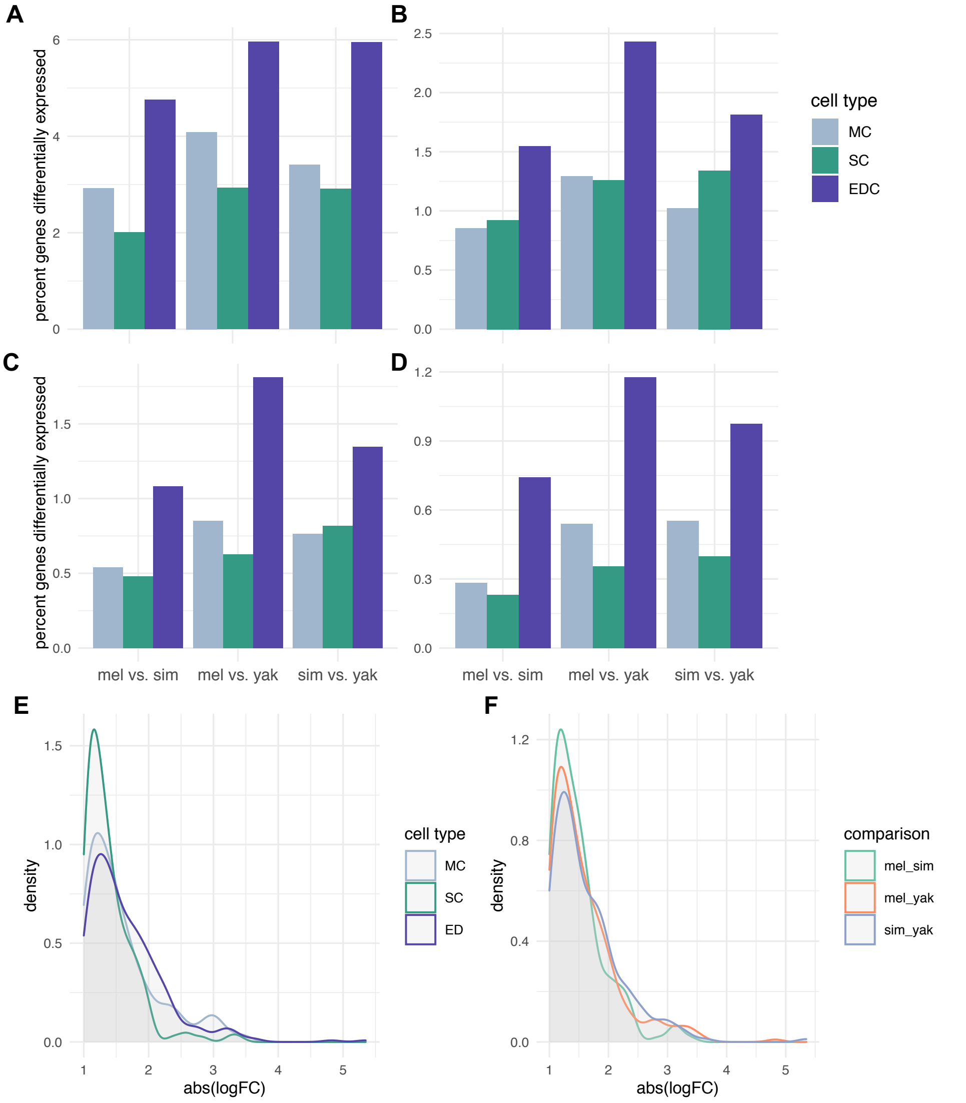

Figure S4. Patterns of DE among cell types are robust to a range of logFC cutoffs. (*A*): logFC ≥ 0.5, (*B*): ≥ 1, (*C*): ≥ 1.25, (*D*): ≥ 1.5. While there is a trend towards MC enrichment at the lowest and highest cutoffs, MC and SC are not significantly different in any intraspecific comparison (Wilcoxon rank sum tests, p > 0.05). The magnitude of DE does not vary among (*E*) cell types or (*F*) species contrasts.

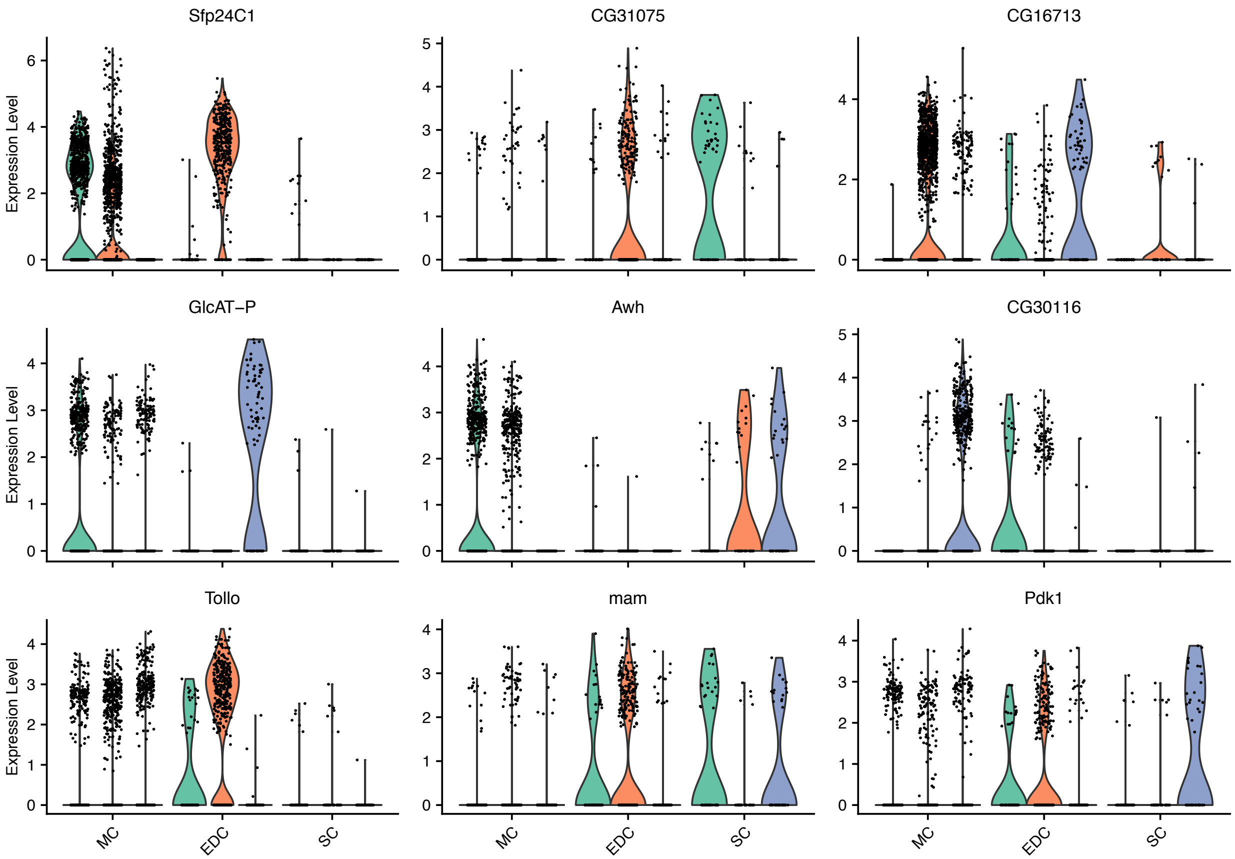

Figure S5. Nine genes with at least one shift in marker status among species. Expression is in log-normalized counts. Species are grouped into threes within cell types, with *mel*, *sim*, and *yak* going from left to right.

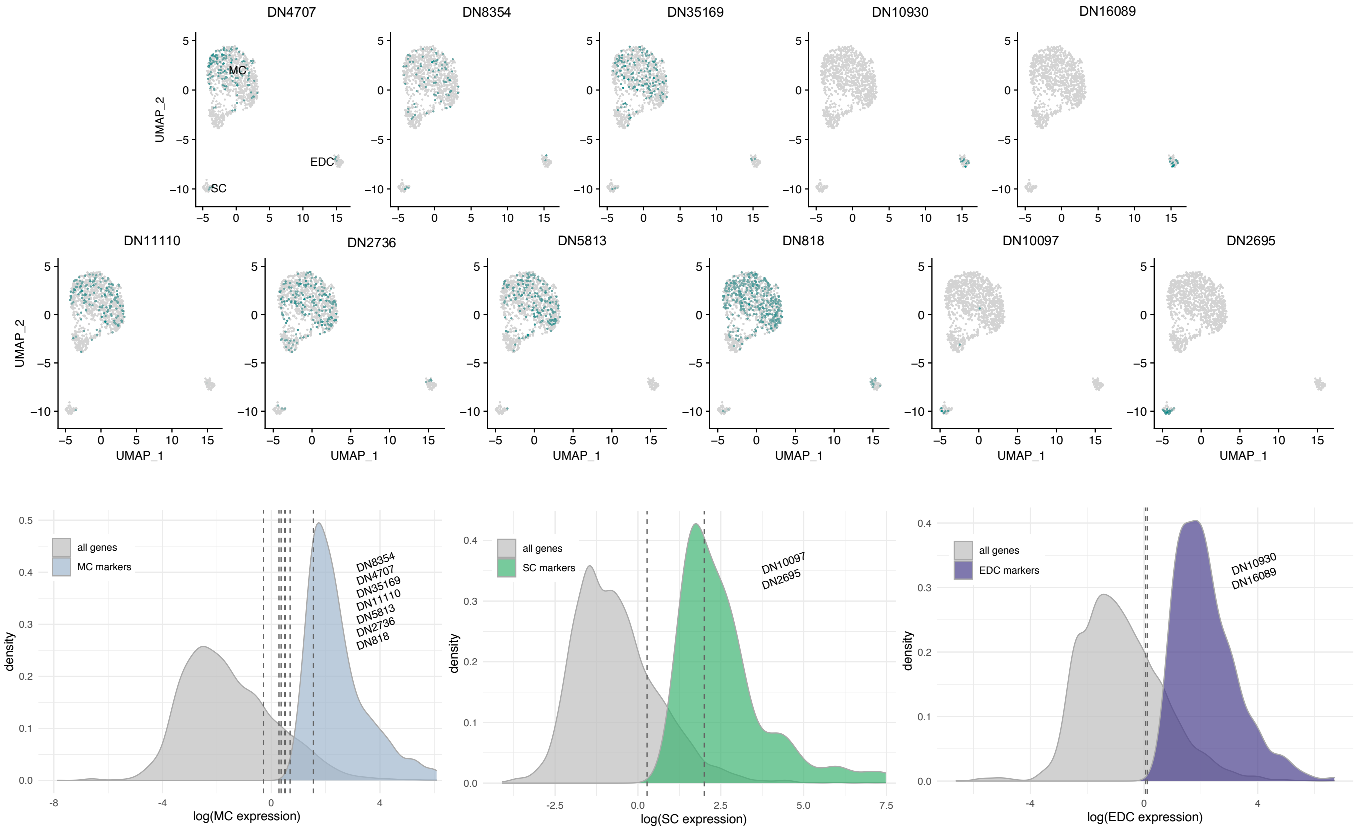

Figure S6. (*A*) UMAP showing three cell types in *mel*, along with expression of unannotated candidate genes (Table 1, Table S4) indicated in teal. (*B*) Distributions of gene expression among all expressed genes (grey) and marker genes (color), for MC, SC, and EDC. Expression values for unannotated genes are overlayed. These data indicate that most unannotated genes are expressed at a relatively high level among all expressed genes, but a very low level among marker genes. Two exceptions to this trend are *DN818* in MC and *DN2695* in SC—the only two unannotated genes that are also characterized as marker genes.

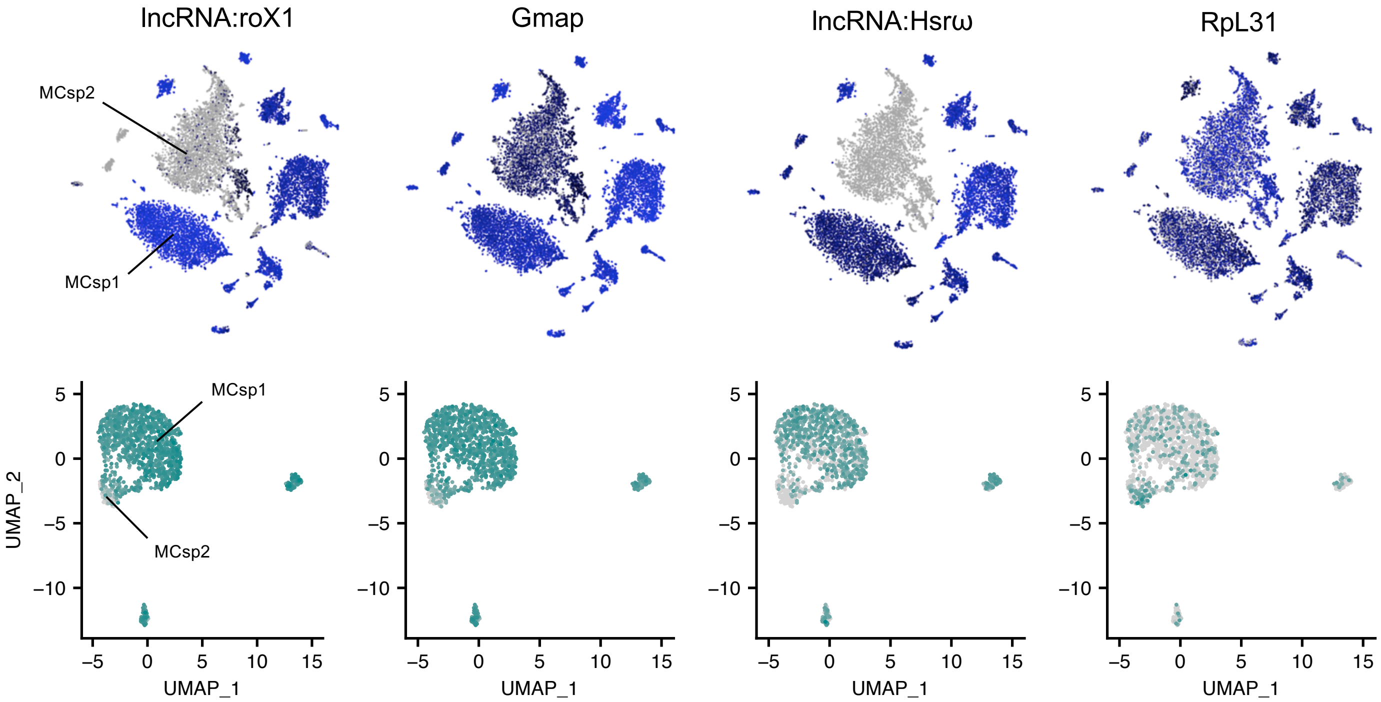

Figure S7. Fly Cell Atlas (Li et al. 2021) shows the same major subpopulations of MC as our data. UMAP plots for male reproductive tract data show evidence of MCsp1 and MCsp2 (for complete cluster information, see Fig. S14 of Li et al. 2021). Top marker genes from our data (Fig. 3*D*) show the same patterns in the Fly Cell Atlas. Bright blue indicates greater expression, dark blue indicates less, while gray indicates lack of expression. The same marker genes in our data are displayed below each Fly Cell Atlas plot, with darker cyan indicating greater expression. *roX1*, *Gmap*, and *Hsrω* have decreased expression in MCsp2 relative to MCsp1. *RpL31*, *eEF2*, and the Sfps *SP*, *Mst57Db*, and *Acp36DE* all show increased MCsp2 expression. Plots obtained from the SCope Fly Cell Atlas Database, 10X stringent dataset (https://scope.aertslab.org/#/FlyCellAtlas/*/welcome).

Supplemental materials for Chapter II: Regulatory basis of gene expression evolution in the Drosophila accessory gland

TABLES S5 – S18

Table S5. Overall alignment rates of reads to each reference genome: *D. melanogaster* (*mel*) and *D. simulans* (*sim*).

sample	replicate	*mel* alignment (%)	*sim* alignment (%)
D. melanogaster	1	94.7	43.0
D. melanogaster	2	94.3	44.4
D. melanogaster	3	94.4	44.3
D. simulans	1	37.9	96.5
D. simulans	2	41.0	96.9
D. simulans	3	39.2	97.0
hybrid	1	62.3	71.0
hybrid	2	62.8	69.0
hybrid	3	61.7	70.6

Table S6. Uniquely mapping and shared reads among references. M = millions of reads.

sample	replicate	*mel* unique	*mel* unique (%)	*sim* unique	*sim* unique (%)	shared	shared (%)
D. melanogaster	1	11.93M	45.4	0.32M	1.20	14.1M	53.4
D. melanogaster	2	11.51M	43.6	0.33M	1.23	14.6M	55.1
D. melanogaster	3	12.9M	44.1	0.36M	1.22	16.0M	54.7
D. simulans	1	0.16M	0.62	13.1M	52.3	11.8M	47.1
D. simulans	2	0.17M	0.69	11.8M	50.9	12.4M	48.4
D. simulans	3	0.17M	0.63	13.8M	50.6	13.3M	48.7
hybrid	1	5.18M	20.1	8.68M	33.7	11.9M	46.7
hybrid	2	5.56M	20.5	9.02M	33.3	12.5M	46.1
hybrid	3	5.01M	19.9	8.35M	33.1	11.9M	47.0

Table S7. Species-of-origin assignments to reads that mapped to both genomes.

sample	replicate	assigned *mel* (%)	assigned *sim* (%)	undetermined (%)
D. melanogaster	1	62.9	0.17	36.9
D. melanogaster	2	61.9	0.20	37.2
D. melanogaster	3	61.5	0.16	38.4
D. simulans	1	0.035	64.9	35.1
D. simulans	2	0.035	65.2	34.8
D. simulans	3	0.031	65.0	35.0
hybrid	1	24.7	34.8	40.4
hybrid	2	25.1	35.0	39.8
hybrid	3	25.6	35.6	38.7

Table S8. Percentage of misassigned parental reads that do not overlap any gene features (corresponding to intergenic regions). Uniquely misassigned reads originate in the step of aligning reads to both genomes. Algorithmically misassigned genes aligned to both genomes but were misassigned when assigning species-of-origin.

sample	replicate	uniquely misassigned with no feature (%)	algorithmically misassigned with no feature (%)
D. melanogaster	1	73.7	39.5
D. melanogaster	2	72.0	41.1
D. melanogaster	3	73.9	41.7
D. simulans	1	53.6	45.8
D. simulans	2	54.1	45.4
D. simulans	3	53.1	44.6

Table S9. Number of DE genes among different classes. *AG-biased* and *non-biased* gene sets exclude Sfps. P refers to the P_{mel} - P_{sim} contrast while ASE refers to the $F1_{mel}$ - $F1_{sim}$ contrast.

gene class	contrast	DE genes	non-DE genes	% DE genes
non-Sfp	P	2883	6132	32.0
Sfp	P	122	86	58.7
non-biased	P	2733	6044	31.1
AG-biased	P	150	88	63.0
non-Sfp	ASE	2025	6990	22.5
Sfp	ASE	107	101	51.4
non-biased	ASE	1889	6888	21.5
AG-biased	ASE	136	102	57.1

Table S10. Number of DE genes among different classes, with $\log_2$(fold change) > 1. *AG-biased* and *non-biased* gene sets exclude Sfps. P refers to the P_{mel} - P_{sim} contrast while ASE refers to the $F1_{mel}$ - $F1_{sim}$ contrast.

gene class	contrast	DE genes	non-DE genes	% DE genes
non-Sfp	P	1549	7466	17.2
Sfp	P	59	149	28.4
non-biased	P	1461	7316	16.6
AG-biased	P	88	150	37.0
non-Sfp	ASE	1057	7958	11.7
Sfp	ASE	46	162	22.1
non-biased	ASE	989	7788	11.3
AG-biased	ASE	68	170	28.6

Table S11. Regulatory classifications among gene sets.

regulatory class	*cis*	*trans*	*cis* + *trans*	*cis* by *trans*	compensatory	conserved	ambiguous
autosomal	933	912	1116	104	381	4062	1715
autosomal (%)	10.12	9.89	12.1	1.13	4.13	44.04	18.59
Sfps	20	31	73	8	15	20	41
Sfps (%)	9.62	14.9	35.1	3.85	7.21	9.62	19.71
AG-biased	33	32	85	7	21	18	42
AG-biased (%)	13.87	13.45	35.71	2.94	8.82	7.56	17.65

Table S12. Regulatory classifications (except conserved and ambiguous) among gene sets

regulatory class	*cis*	*trans*	*cis* + *trans*	*cis* by *trans*	compensatory
autosomal	933	912	1116	104	381
autosomal (%)	27.07	26.47	32.39	3.02	11.06
Sfps	20	31	73	8	15
Sfps (%)	13.61	21.09	49.66	5.44	10.20
AG-biased	33	32	85	7	21
AG-biased (%)	18.54	17.98	47.75	3.93	11.80

Table S13. Inheritance classifications among gene sets.

inheritance class	additive	*mel* dominant	*sim* dominant	overdominant	underdominant	conserved
autosomal	731	1403	1208	448	547	4886
autosomal (%)	7.93	15.21	13.10	4.86	5.93	52.98
X-linked	51	119	473	125	121	810
X-linked (%)	3.00	7.00	27.84	7.36	7.12	47.68
Sfps	42	39	33	18	43	33
Sfps (%)	20.19	18.75	15.87	8.65	20.67	15.87
AG-biased	55	34	46	17	57	29
AG-biased (%)	23.11	14.29	19.33	7.14	23.95	12.18

Table S14. Inheritance classifications (except conserved) among gene sets.

inheritance class	additive	*mel* dominant	*sim* dominant	overdominant	underdominant
autosomal	731	1403	1208	448	547
autosomal (%)	16.85	32.35	27.85	10.33	12.61
X-linked	51	119	473	125	121
X-linked (%)	5.74	13.39	53.21	14.06	13.61
Sfps	42	39	33	18	43
Sfps (%)	24	22.29	18.86	10.29	24.57
AG-biased	55	34	46	17	57
AG-biased (%)	26.32	16.27	22.01	8.13	27.27

Table S15. Gain-of-function in hybrids with insignificant parental expression. Average expression is $\log_2$(normalized counts)

gene	average expression in *D. melanogaster*	average expression in *D. simulans*	average expression in hybrid
prolyl-4-hydroxylase-α MP	0.000	0.000	2.681
CG9395	0.407	0.202	2.362
CG8500	0.705	0.706	2.324
CG30090	0.000	0.178	1.929
CG34316	0.795	0.380	2.501

Table S16. Loss-of-function with insignificant hybrid expression. Average expression is $\log_2$(normalized counts).

gene	average expression in *D. melanogaster*	average expression in *D. simulans*	average expression in hybrid
β-Tubulin at 85D	2.229	1.576	0.000
Pendulin	2.227	2.952	0.314

Table S17. Percentage of inheritance classifications among regulatory types.

regulation	conserved	additive	*mel* dominant	*sim* dominant	overdominant	underdominant
cis	16.4	32.8	26.6	25.2	0	0
trans	2	5.9	45.3	26.8	8.2	10.9
cis + *trans*	0.9	32	26.3	24.5	9.6	6.7
cis by *trans*	12.5	7.7	26.9	24	22.1	6.7
compensatory	30.2	0	13.6	16.3	19.2	20.7
conserved	96.1	0	1.6	2.2	0	0.2
ambiguous	39.3	0.9	17.3	16.4	9.9	16.3

Table S18. Percentage of chromatin peak types associated with DE genes.

contrast	type	no peak	conserved	*mel* orphan	*sim* orphan
P_{mel} P_{sim}	non-DE	74.0	19.0	2.6	4.5
P_{mel} P_{sim}	DE	55.8	27.7	6.4	10.1
$F1_{mel}$ $F1_{sim}$	non-DE	73.2	19.6	2.8	4.5
$F1_{mel}$ $F1_{sim}$	DE	56.1	27.1	6.2	10.6

Table S19. Percentage of chromatin peak types associated with DE genes, with $\log_2$(fold change) > 1.

contrast	type	no peak	conserved	*mel* orphan	*sim* orphan
P_{mel} P_{sim}	non-DE	74.0	19.0	2.6	4.5
P_{mel} P_{sim}	DE	55.8	27.7	6.4	10.1
$F1_{mel}$ $F1_{sim}$	non-DE	73.2	19.6	2.8	4.5
$F1_{mel}$ $F1_{sim}$	DE	56.1	27.1	6.2	10.6

Table S20. Correlations of expression divergence and accessibility divergence among regulatory classes, with differentially abundant, conserved ATAC-Seq peaks (as in Figure 6E,F). n is the sample size of genes in each comparison. P_{mel} P_{sim} ρ is the Spearman rank coefficient of determination between parental expression divergence and chromatin accessibility divergence. P_{mel} P_{sim} p is the p value where the null hypothesis is that $\rho = 0$. Similarly, $F1_{mel}$ $F1_{sim}$ ρ and p represent the Spearman rank coefficient of determination and p value for the correlation between ASE divergence and chromatin accessibility divergence.

regulatory class	n	P_{mel} P_{sim} ρ	P_{mel} P_{sim} p	$F1_{mel}$ $F1_{sim}$ ρ	$F1_{mel}$ $F1_{sim}$ p
conserved	145	-0.01	0.954	-0.08	0.364
cis	70	0.41	< 0.001	0.42	< 0.001
trans	97	0.52	< 0.001	0.14	0.162
cis + *trans*	140	0.52	< 0.001	0.44	< 0.001

Table S21. ATAC-Seq peak classes among regulatory types

regulation	conserved peak	conserved peak (%)	*mel* orphan	*mel* orphan (%)	*sim* orphan	*sim* orphan (%)	no peak	no peak (%)
cis	209	22.4	83	8.9	99	10.6	542	58.1
trans	276	30.3	75	8.2	117	12.8	444	48.7
cis + *trans*	303	27.2	103	9.2	173	15.5	537	48.1
cis by *trans*	34	32.7	7	6.7	17	16.3	46	44.2
compensatory	111	29.1	27	7.1	57	15.0	186	48.8
conserved	527	13.0	111	2.7	182	4.5	3242	79.8
ambiguous	479	27.9	104	6.1	159	9.3	973	56.7

FIGURES S8 – S19

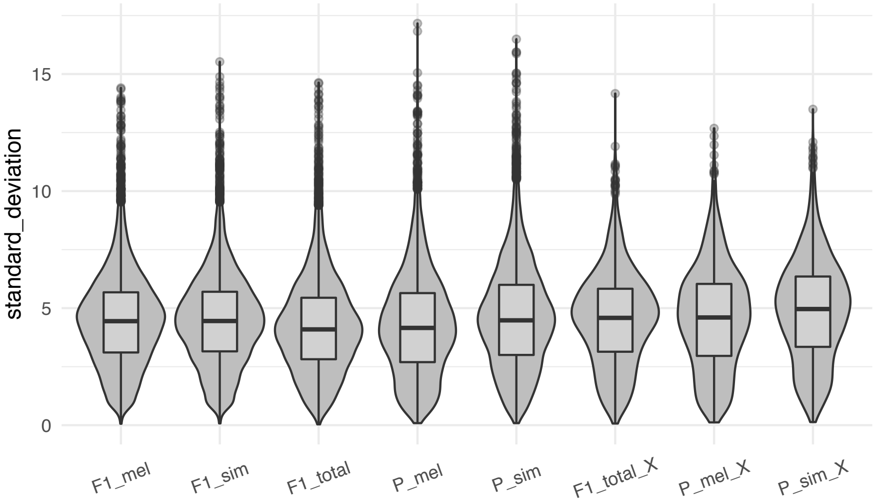

Figure S8. Distributions of observed standard deviation of expression among replicates. The first five depict autosomal-linked genes and the last three depict X-linked genes.

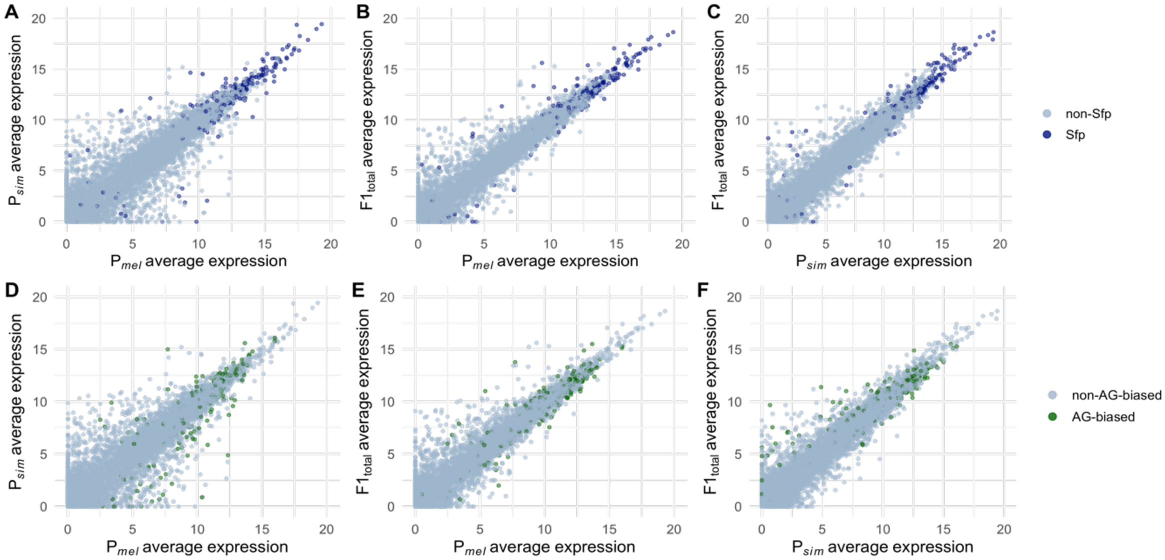

Figure S9. Correlations of average gene expression, as in Figure 7, with Sfps (A-C) and AG-biased genes (D- F) highlighted

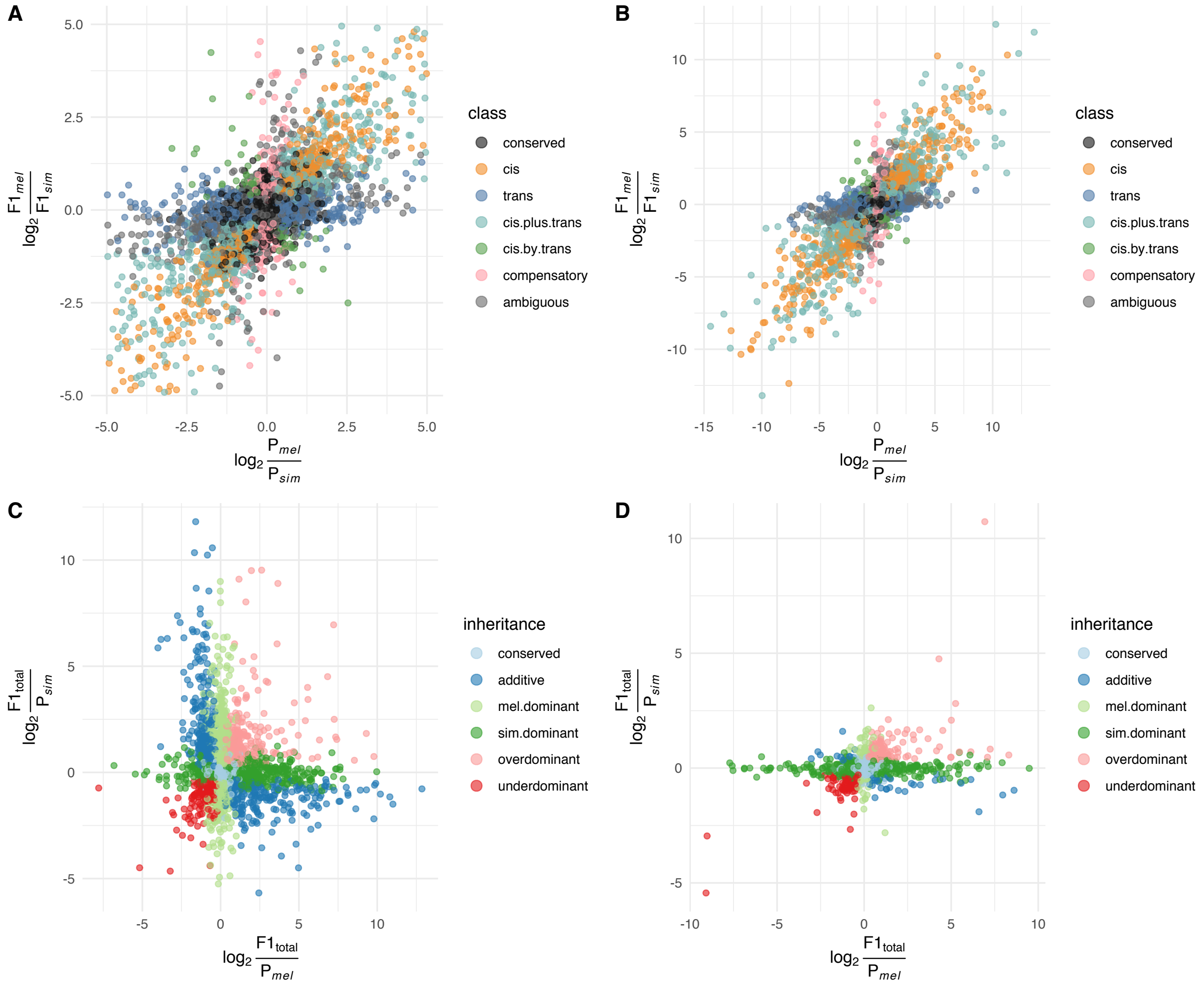

Figure S10. A) Regulatory classification as in Figure 3 with the addition of ambiguous genes. B) Full-scale log$_2$(fold change) regulatory plot. C) Full-scale log$_2$(fold change) inheritance plot for autosomal-linked genes. D) Full-scale log$_2$(fold change) inheritance plot for X-linked genes.

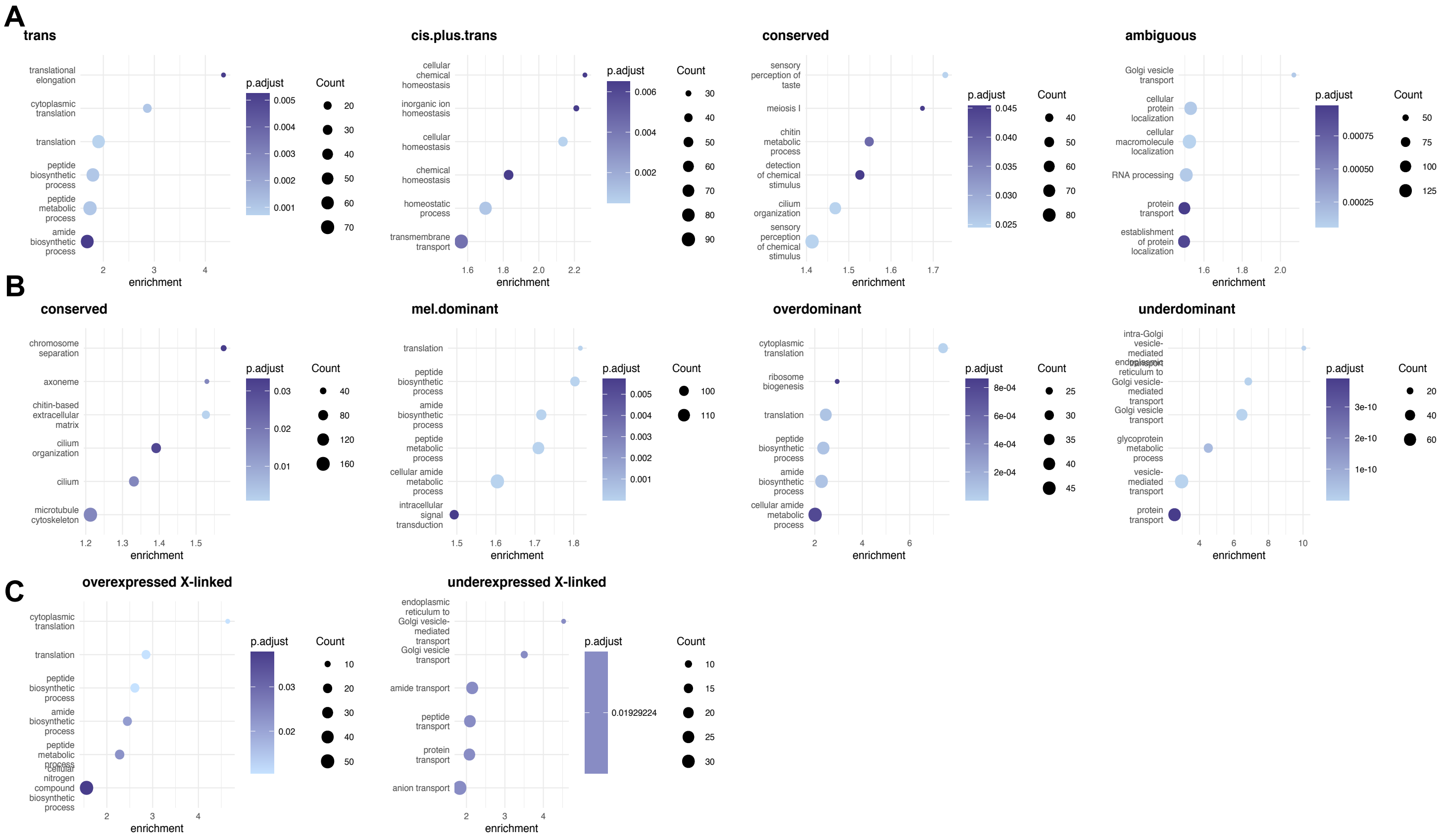

Figure S11. GO terms associated with regulatory and inheritance types. Enrichmentis the ratio of terms in the test gene set compared to the background gene set. A) Regulatory types; B) inheritance types in autosomes; C) over- or underexpressed X-linked genes. Terms associated with *cis* and *cis* by *trans* genes are particularly weakly significant and not shown here (see Supplemental Data).

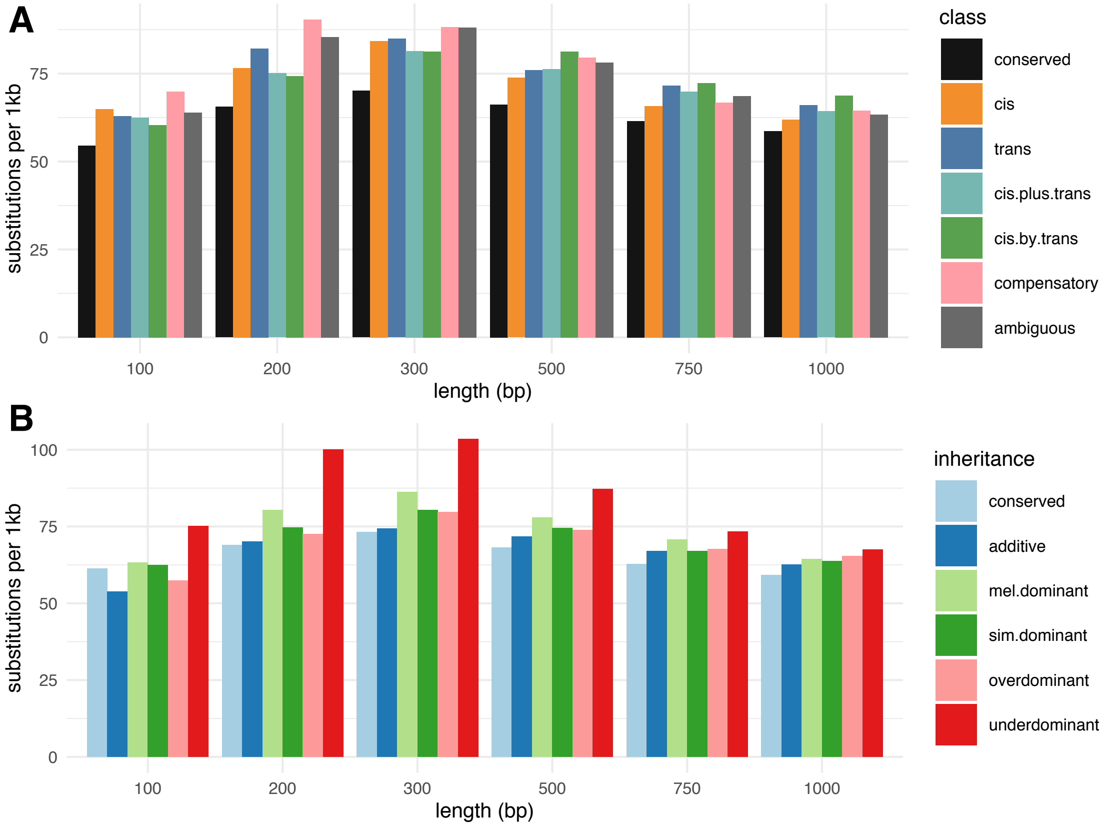

Figure S12. Kimura-2-parameter estimated substitution rates in upstream sequences between *D. melanogaster* and *D. simulans* at various lengths from the TSS. A) regulatory classes; B) inheritance classes.

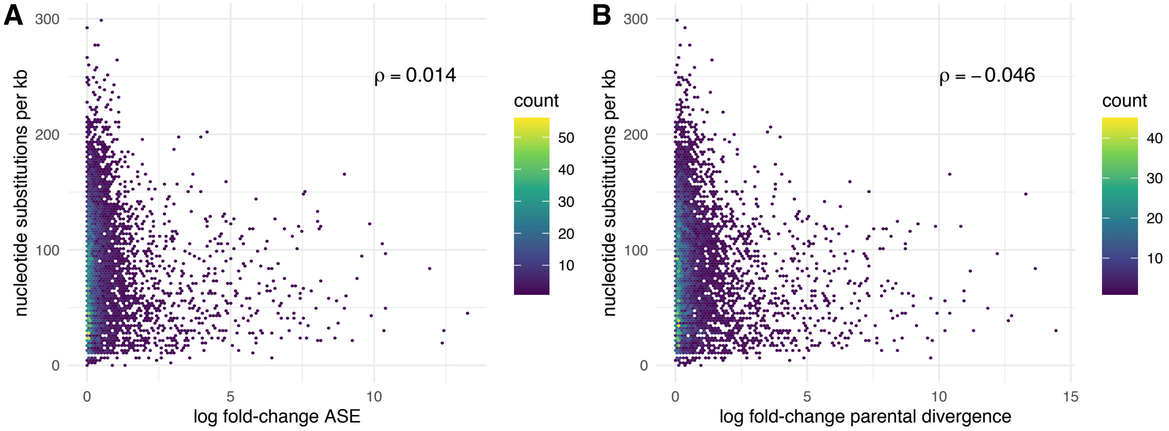

Figure S13. Kimura-2-parameter estimated substitution rates in upstream sequences between *D. melanogaster* and *D. simulans* plotted against A) allele-specific and B) parental expression divergence. Spearman's rank correlation coefficient ρ is shown on each panel.

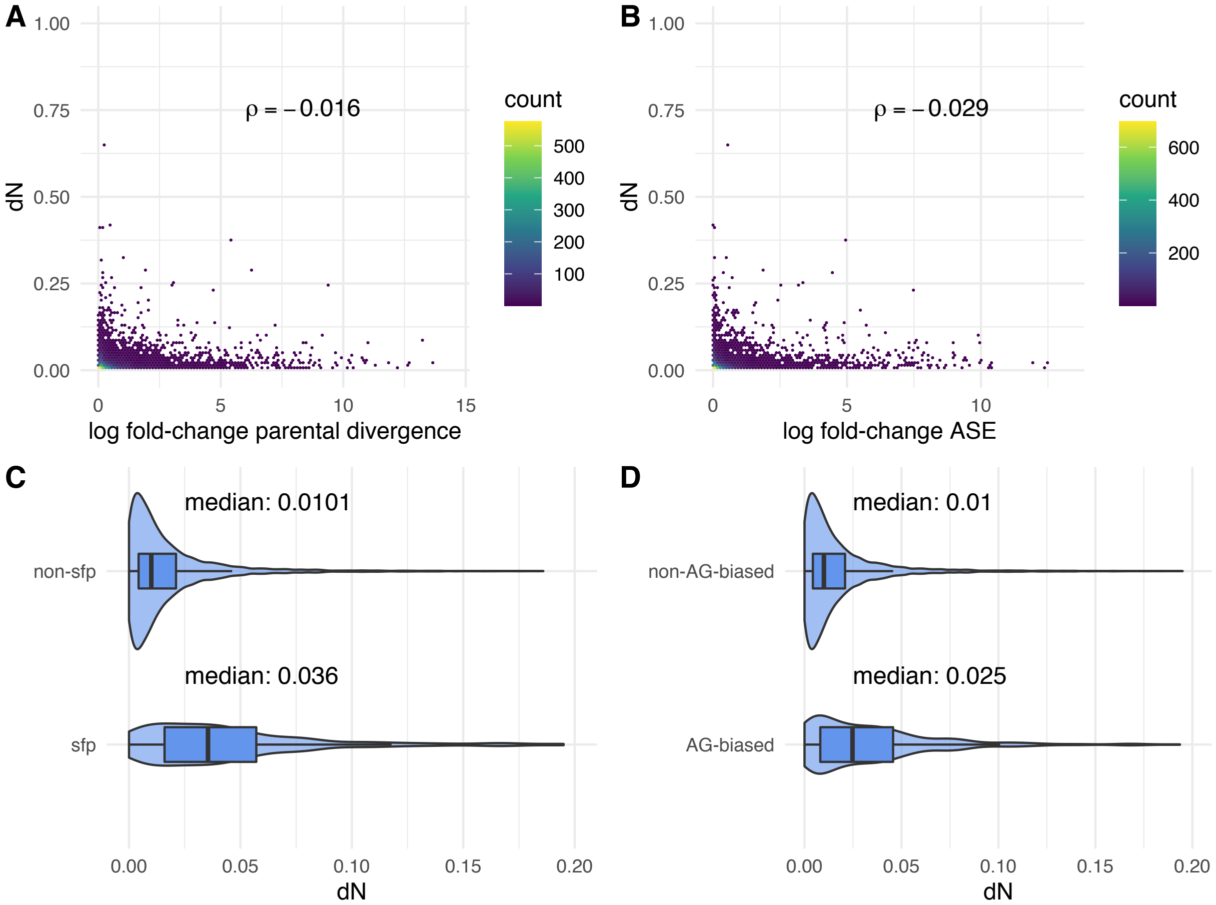

Figure S14. dN between *D. melanogaster* and *D. simulans* plotted against A) parental divergence and B) ASE divergence. Spearman's rank correlation coefficient ρ is shown in each panel. C) Distributions of dN between Sfps and non-Sfps (Wilcoxon rank sum test, $p < 0.001$). D) Distributions of dN between accessory gland-biased and non-accessory gland-biased genes (Wilcoxon rank sum test, $p = 0.00101$).

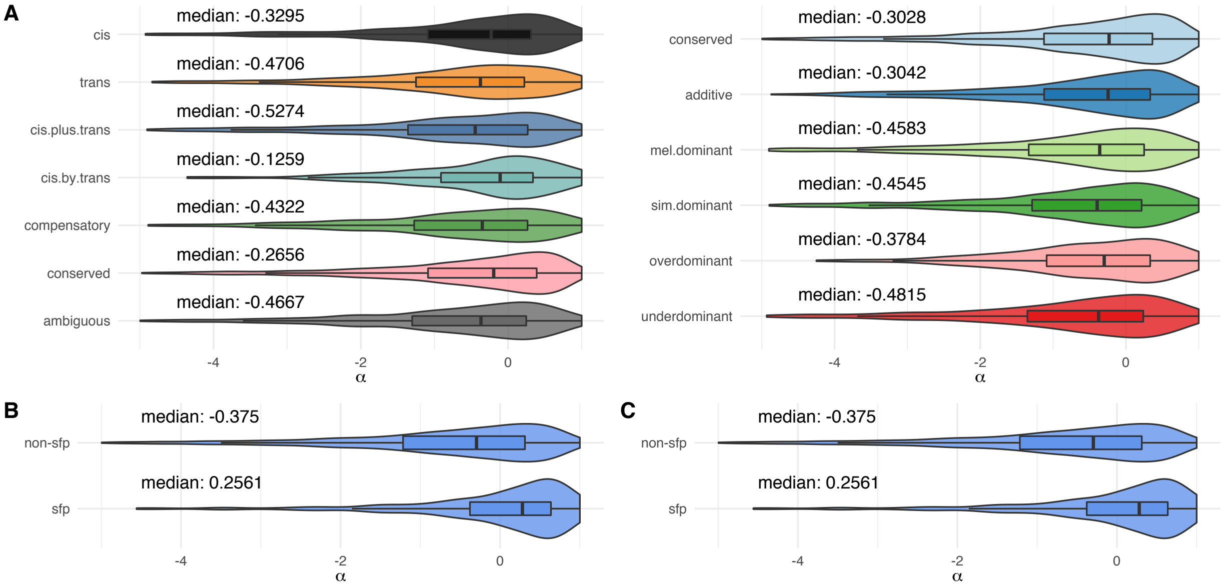

Figure S15. A) Distributions of α among regulatory and inheritance classes. Genes with conserved regulation have significantly greater median α than *trans* (Wilcoxon rank sum test, p = 0.003), *cis* + *trans* (p < 0.001), and ambiguous genes (p < 0.001); all other comparisons of inheritance types are not significantly different (p > 0.05). Genes with conserved inheritance have significantly greater median α than *mel* dominant (p = 0.001), *sim* dominant (p = 0.007), and underdominant genes (p = 0.021); all other comparisons of inheritance types are not significantly different (p > 0.05). B) Sfps have significantly greater median α than non-Sfps (p < 0.001). C) Distributions of α between AG-biased and non-AG-biased genes are not significantly different (p = 0.52).

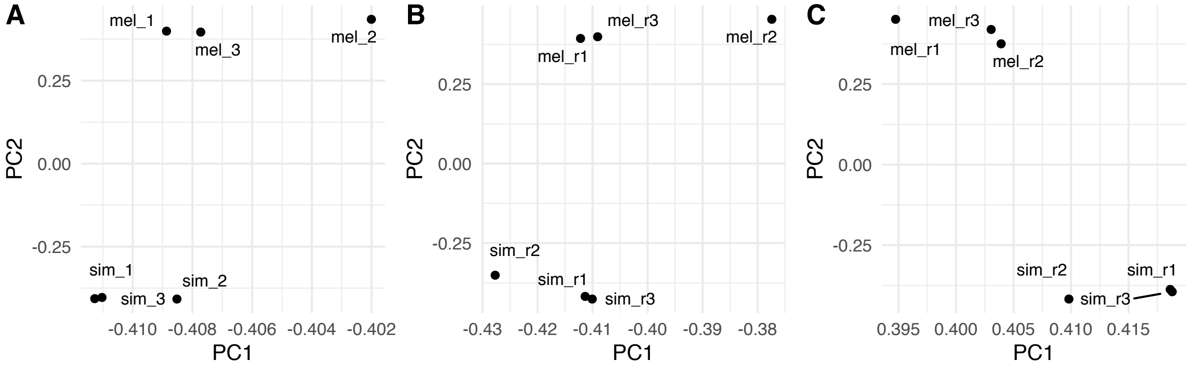

Figure S16. PCA of peak accessibility (log$_2$ transformed counts) with the first two PCs shown: (A) conserved peaks, (B) *mel* orphan peaks, (C) *sim* orphan peaks.

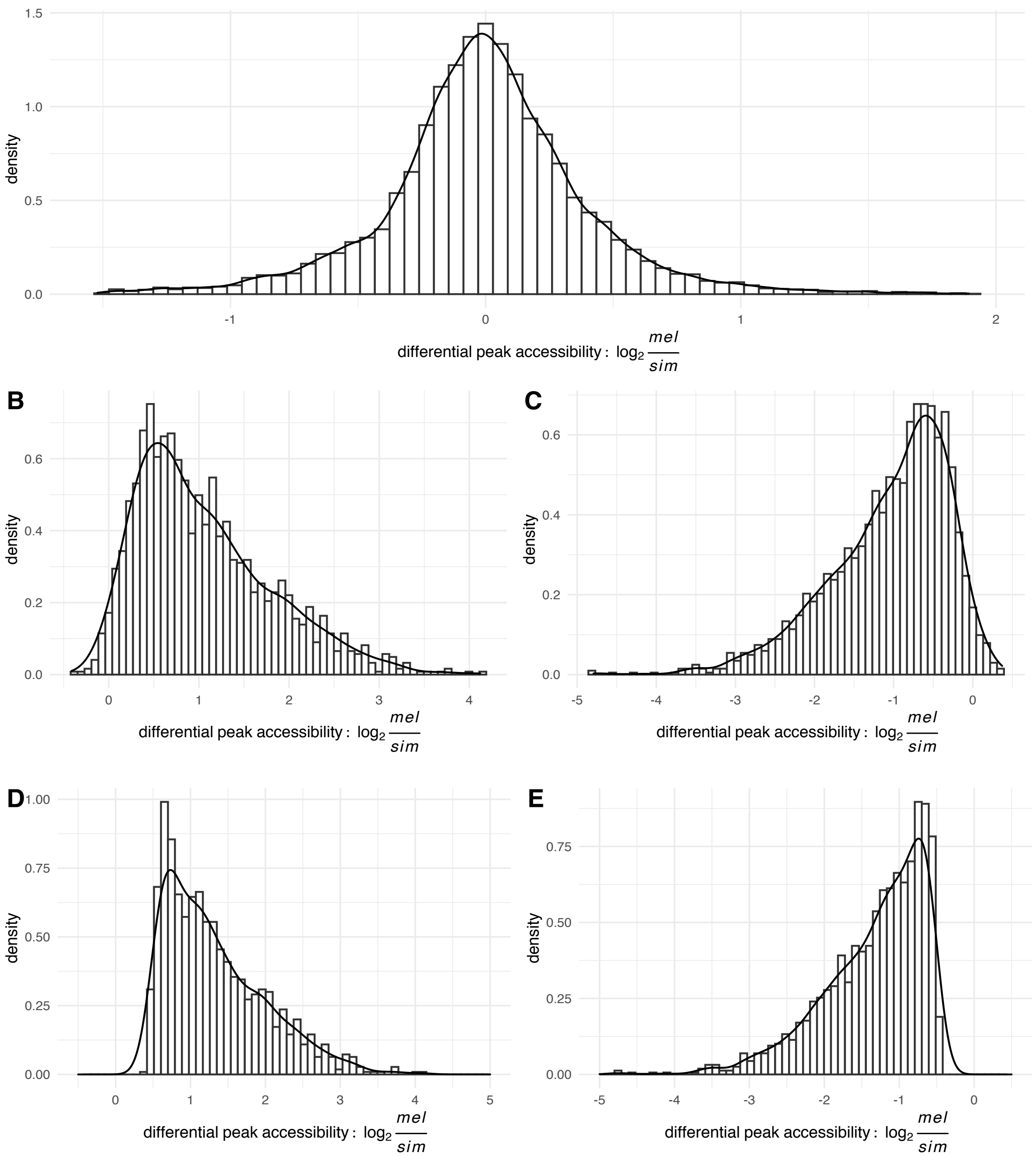

Figure S17. Differential accessibility (DA) of chromatin peaks called by ATAC-Seq, defined as the $\log_2$-transformed ratio of counts *D. melanogaster* counts to *D. simulans* counts. (A) Conserved peaks; (B) *mel* orphan peaks; (C) *sim* orphan peaks. We further filtered orphan peaks to just those that are DA and more highly expressed in the species with a peak called: (D) filtered *mel* orphan peaks; (E) filtered *sim* orphan peaks.

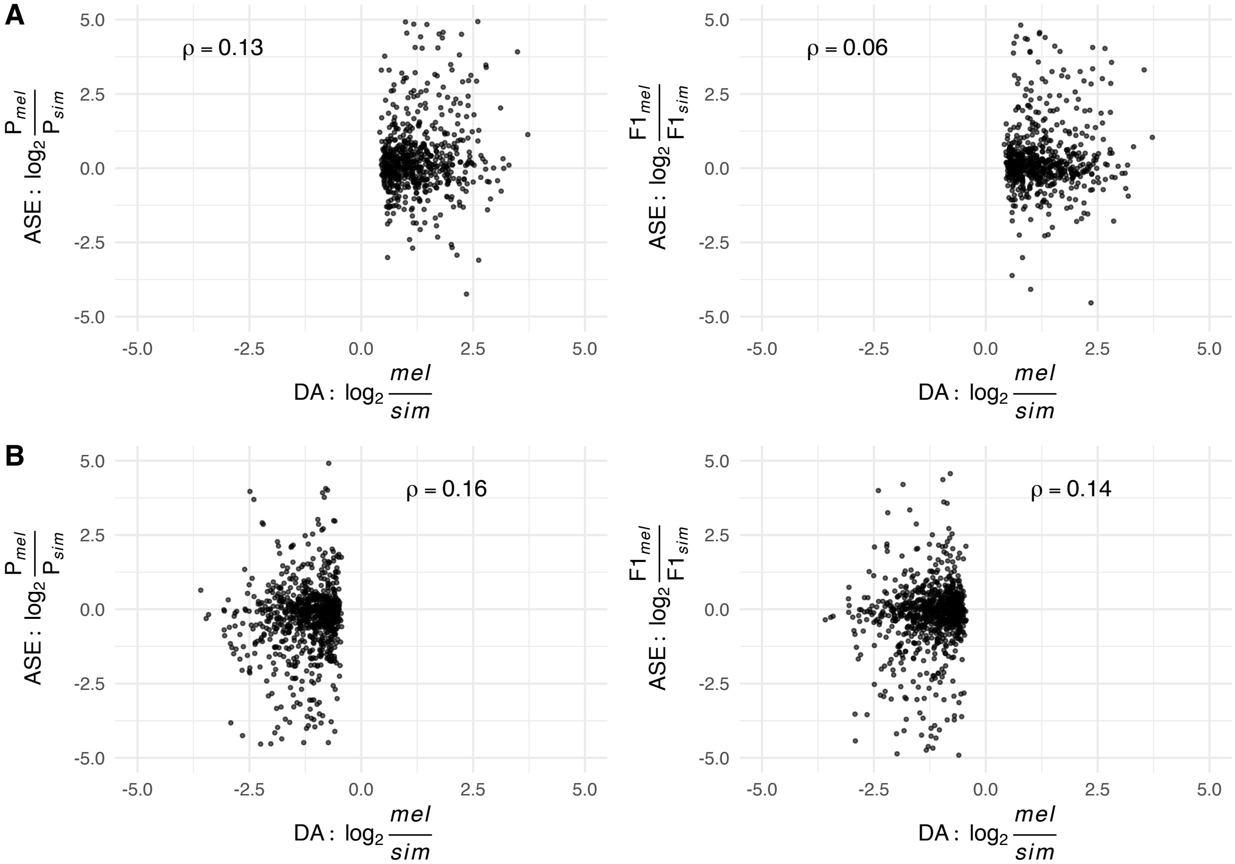

Figure S18. Parental expression divergence or ASE (y-axis) plotted against differential accessibility (DA) of chromatin peaks called by ATAC-Seq (x-axis). (A) *mel* orphan peaks; (B) *sim* orphan peaks. Spearman's rank coefficient ρ is shown on each plot.