Comparative single cell analysis of transcriptional bursting reveals the role of genome organization on *de novo* transcript origination

UnJin Lee[1], Cong Li[1], Christopher B. Langer[1], Nicolas Svetec[1], Li Zhao[1,*]

Affiliations:

[1]Laboratory of Evolutionary Genetics and Genomics, The Rockefeller University, New York, NY, USA

[*] Correspondence to: lzhao@rockefeller.edu

Abstract:
Spermatogenesis is a key developmental process underlying the origination of newly evolved genes. However, rapid cell type-specific transcriptomic divergence of the *Drosophila* germline has posed a significant technical barrier for comparative single-cell RNA-sequencing (scRNA-Seq) studies. By quantifying a surprisingly strong correlation between species- and cell type-specific divergence in three closely related *Drosophila* species, we apply a new statistical procedure to identify a core set of 198 genes that are highly predictive of cell type identity while remaining robust to species-specific differences that span over 25-30 million years of evolution. We then utilize cell type classifications based on the 198-gene set to show how transcriptional divergence in cell type increases throughout spermatogenic developmental time. After validating these cross-species cell type classifications using RNA fluorescence in situ hybridization (FISH) and imaging, we then investigate the influence of genome organization on the molecular evolution of spermatogenesis vis-a-vis transcriptional bursting. We first demonstrate how mechanistic control of pre-meiotic transcription is achieved by altering transcriptional burst size while post-meiotic control is exerted via altered bursting frequency. We then report how global differences in autosomal vs. X chromosomal transcription likely arise in a developmental stage preceding full testis organogenesis by showing evolutionarily conserved decreases in X-linked transcription bursting kinetics in all examined somatic and germline cell types. Finally, we provide evidence supporting the cultivator model of *de novo* gene origination by demonstrating how the appearance of newly evolved testis-specific transcripts potentially provides short-range regulation of neighboring genes' transcriptional bursting properties during key stages of spermatogenesis.

Significance Statement:
Understanding the divergence and evolution of novel genes and expression is essential in evolutionary biology. The testis is a hotspot for gene origination and diversification, partly because of strong sexual selection on male reproduction. However, rapid transcriptomic divergence in the germline has significantly hindered comparative single-cell RNA-seq studies and understanding of the rapidly evolving components of the cell types. Our novel strategy, which does not rely on preexisting marker genes for cell type identity, overcame the challenge. We provide evidence supporting the cultivator model of de novo gene origination—a model we

recently proposed to emphasize the genomic environment in shaping novel gene origination—by illustrating how newly evolved transcripts are regulated by the transcriptional bursting of neighboring genes.

Main text:

Introduction

Gametogenesis is an essential process in all sexually reproducing species in nature. It ensures that parents' genetic material is successfully transferred to the next generation. Male gametogenesis, spermatogenesis, is a specialized developmental process that exhibits several key features. First, the function and process of spermatogenesis are remarkably conserved across animals. Second, despite these conserved functions, testis-expressed genes tend to evolve very rapidly, with some genes being newly duplicated or newly originated. Third, although cell types in a lineage in the testis—such as in *Drosophila*—exhibit conserved features, the expression patterns of genes in each cell type can vary drastically between species. Because of this, the expression dynamics of testis-expressed genes are an important issue for both evolutionary biology and developmental biology.

The testis is a hotspot harboring newly evolved genes, including novel protein-coding genes arising *de novo* from non-coding or non-genic sequences, and non-coding RNA genes arising *de novo* from previously transcriptionally inactive sequences. While the origination of *de novo* protein-coding genes has been relatively well-studied (1–3), the study of the evolutionary mechanisms underlying the origination of *de novo* transcripts such as lncRNA in fruit flies has been lacking (1, 4, 5). This gap highlights the importance of comprehensively identifying such RNA genes in multiple *Drosophila* species. Notably, such *de novo* transcripts are often expressed in a tissue-specific manner, typically in the adult testis of *Drosophila* (1, 4, 6). Intriguingly, scRNA-Seq data has shown that *de novo* originated genes, unlike recently evolved duplicate genes, are biased toward expression in mid-spermatogenesis (7). It is not yet clear if this expression pattern is common across multiple *Drosophila* species, and what function this particular expression pattern may hold.

Another interesting facet of spermatogenesis in *Drosophila* is the transcriptional down-regulation of X-linked genes. However, the molecular mechanisms underlying this down-regulation remain under-studied (8). While prior work has suggested that expression of genes on the X chromosome is globally down-regulated during spermatogenesis in *Drosophila* (9), it is not possible to determine anything beyond the general reduction of expression for X-linked genes (8, 10). While it is unclear whether germline cells have complete dosage compensation, previous single-cell RNA-Seq data has proposed that there is some degree of non-canonical

dosage compensation (11). Alternatively, a decrease in X chromosome expression beyond that which would be expected by dosage has led others to suggest that the entire *Drosophila* X chromosome is inactivated (12).

Transcription is an intrinsically random process which may be modeled stochastically (13, 14). Single molecule observations of transcriptional activity are intrinsically "bursty," consisting of short bouts of transcription that produce discrete quantities of mRNAs ("on" state) followed by periods of inactivity ("off" state) (15–17). Two key parameters controlling the transcriptional kinetics of different genes are burst size, which reflects the number of transcripts produced per burst, and burst frequency, which reflects the number of bursts that occur per unit time (15, 18, 19). Importantly, the analysis of transcriptional bursting kinetics offers unique insight into understanding how certain genes are regulated. The kinetics underlying burst size is a property of core promoter elements, driven by the stoichiometry of RNA polymerase II (pol II) availability, while the kinetics of burst frequency is influenced primarily by enhancer elements, modulated by the probability of transcriptional initiation through the action of transcription factors (20–23). Transcriptional bursting properties have provided a novel perspective into the molecular underpinnings of sex-biased transcription burst amplitude in male and female embryos of *Drosophila melanogaster* (24). In showing that the upregulation of male X chromosome genes is primarily mediated by a higher RNA polymerase II initiation rate and transcriptional burst amplitude (24), this work paved the way for further study of dosage compensation in spermatogenesis.

While the transcriptomes of highly conserved cell types in testis are likely to remain stable over large evolutionary timescales, the rapid evolution of spermatogenesis in *Drosophila* presents significant challenges for comparative analyses of single-cell transcriptomic data, even among closely related *Drosophila* species. For instance, many seemingly conserved marker genes from *D. melanogaster* (25, 26) are lost or not annotated in other *Drosophila* species, which means even marker genes, some of which are transcription factors, are not evolutionarily conserved. Second, the rate of gene expression changes and differences drastically vary in *Drosophila* (27, 28), making it problematic to simply group cell types across different species. Our interest lies not only in the conserved transcriptional programs of cell types across species but also in complete gene expression patterns, including species-specific genes. Therefore, specialized software such as SAMap (29), designed to directly tackle the issue of mapping multi-

species single cell transcriptomic data onto a single, unified projection, is not suitable for our study. Furthermore, the application of standard batch correction and integration techniques such as Seurat's use of 'CCAIntegration' (30), Monocle 3's use of 'batchelor' (31), and the independent Harmony software (32) on single cell testis data utilizing "melanogasterized" transcriptomes often fail to reproduce previous cell-type assignments in *D. melanogaster* samples. These software packages also generate a large number of species-specific cell type clusters that are unlikely to reflect true phenotypic differences in expression.

Within approximately 30 million years of evolution in *Drosophila*, the process of germline differentiation has undergone a large number of phenotypic changes (33–35). Alternatively, a large degree of spermatogenesis (Figure 1a) has remained unaltered through similar evolutionary time scales (36, 37), including gross anatomical phenotypes such as organ structure or the progression of various cell-type specific changes like spermatid differentiation or histone-to-protamine transitions (38, 39). We therefore hypothesized that a high degree of species-specific gene expression evolution within each cell type is the primary factor driving technical difficulties in clustering and identifying cell type differences within comparative studies of *Drosophila* spermatogenesis (Figure S1).

In our study, we undertook a comprehensive analysis of spermatogenesis in three *Drosophila* species, aiming to unravel the complex transcriptional dynamics and novel expression evolution that underpin this developmental process. We first developed a novel approach to study scRNAseq data from rapidly diverging species. We then quantified a strong correlation between species- and cell type-specific transcriptomic divergence across three closely related *Drosophila* species, using this data to identify a core set of genes crucial for predicting cell type identity despite evolutionary differences spanning over approximately 30 million years. Using this gene set, we classified cell types and analyzed how transcriptional divergence progresses throughout spermatogenic development. We further investigated the effects of genome organization on the molecular evolution of spermatogenesis, particularly focusing on transcriptional bursting. We demonstrated variations in transcriptional burst size and frequency before and after meiosis and examined the differences in autosomal versus X chromosomal transcription. Additionally, our findings supported the cultivator model of *de novo* gene origination (40), showing that newly evolved testis-specific transcripts might play a regulatory role in the transcriptional bursting properties of neighboring genes during crucial stages of

spermatogenesis. This work not only deepens our understanding of gene evolution in *Drosophila* but also highlights the interplay between the rapid changes in transcriptional dynamics and conserved cell type identity.

Results

A simple ANOVA-based statistical procedure identifies gene lists predictive of cell type but not species identity

We used several commonly available software suites to analyze scRNA-seq data from three species of testes but were unable to obtain high-quality cell type assignments and correlate similar cell types among different species (Figure S1). Widely used markers in *D. melanogaster* did not robustly classify cell type identities in species such as *D. ananassae*. We thus hypothesize that species-specific cell type transcriptomic divergence significantly hinders cell type classification. To determine this, we generated a simulated data set consisting of nine unique cell types evolving in three different species (Supplementary Methods, Figure S2a), to simulate a scenario in which cell type differences are partly masked by species-level differences. We then applied a simple methodology to calculate ANOVA F-statistic scores for cell type and species (Figure S2b). A good cell type marker gene for multiple species should exhibit a large F for cell type and a small F for species identity (bottom-right quadrant in Figure S2b). We selected the top 200 genes showing high F_i (cell type) and low F_i (species identity), and found the genes identified through our methodology effectively captured the most informative genes that preferentially contain cell type information over species identity information (Figure S2c).

Uncorrected simulated data resulted in non-overlapping cell types, as they were clustered based on both cell type and species identity (Figure S2d). When standard batch correction techniques, such as Monocle3's implementation of batchelor via 'align_cds()' (31, 41) , are applied to remove species-specific effects, UMAP and PCA projection results show that these effects have been entirely removed with cell type assignments overlapping entirely despite species-specific differences in cell type transcriptome expression (Figure S2e). This stands in contrast to the removal of species-specific changes using these same batch correction techniques in real scRNA-Seq data derived from testis tissue (Figure S1), as the remaining species-specific differences are still strong enough to hinder the identification and clustering of cell types in these data sets. This is likely due to the excessively simplistic and additive nature of the simulations in

contrast to the more complex, non-linear nature of cell-type specific evolution. Regardless, the application of standard batch correction techniques remained insufficient for cell type identification in real data.

Projection and clustering using the top 200 most informative gene list in the simulated data showed distinct cell type clusters that still maintained species-specific differences (Figure S2f, g). Cell type identities are readily recovered in this visualization, while species-specific divergence within these cell types is also clearly visualized. These show that our method identifies gene lists that are predictive of cell type but not species identity, which is important for clustering cell types across multiple species.

Application of the ANOVA-based statistical procedure identifies 198 gene list in related Drosophila species.

We then applied our non-parametric statistical procedure on our scRNA-Seq data set derived from testis tissue from three related species, *D. melanogaster*, *D. yakuba*, and *D. ananassae*. The divergence time between *D. melanogaster* and *D. yakuba* has been estimated to be about 7 million years ago while the divergence time between *D. melanogaster* and *D. ananassae* has been estimated to be about 25-30 million years ago (42)(Figure 1b). Previously published data and cell type labels from *D. melanogaster* were obtained from (7), while 5000 cells derived from *D. yakuba* and 5000 cells derived from *D. ananassae* were sequenced (c.f. Methods and Material). Raw data was aligned in cellranger to reference genomes and annotations obtained from FlyBase (dmel-r6.44) to generate read counts. Further orthology information between the three species' annotations was obtained using FlyBase (dmel-r6.44) and used to generate a "melanogasterized" *D. yakuba* (dyak-r1.05) and *D. ananasssae* (dana-r1.06) genomes. Specifically, each genome was reduced to a core set of single-copy genes that possessed strict one-to-one orthology between the three species. After processing in Monocle3 (41), data from one-to-one orthologs were combined together and retained for further analyses. A summary of key processing parameters is presented in Table S1.

The distributions of F_i (cell type) and F_i (species identity) were calculated using this data (Figure 1c, Dataset S1). Interestingly, the joint distribution of log(F_i (cell type)) and log(F_i (species identity)) showed an exceedingly large correlation (Pearson $\rho = 0.698$, $p < 2x10^{-16}$, Figure 1c). This result is on par with the correlation found in simulated data (Figure S2b,

Pearson $\rho = 0.74$, $p < 2x10^{-16}$), suggesting a high degree of species-specific cell type evolution in *Drosophila* testis tissue. The medians of the independent distributions F_i (cell type) and F_i (species identity) were chosen as $p_{cell\text{-}type}$ and $p_{species\text{-}identity}$, respectively (Figure 1c, solid lines). Note that the independent distributions of F_i (cell type) and F_i (species identity) are not fully represented in the joint distribution of F_i (cell type) and F_i (species identity), because some genes in the "melanogasterized" gene set i had values of F_i (cell type) or F_i (species identity) – but not both – that were undefined due to variances equaling exactly to zero. This definition of $p_{cell\text{-}type}$ and $p_{species\text{-}identity}$ produced a new marker gene set i' with a cardinality $|i'| = 198$ (Table S2).

To validate the biological significance of the 198-gene set, which showed conserved expression among species using our method and is predicted to be functionally important, we performed a gene ontology (GO) analysis using the PANTHER database (43, 44)(Table S3). The eleven statistically significant major categories (FDR < 0.05) were highly concordant with known function during spermatogenesis, including cilium mobility, dynein-associated pathways, insulin signaling, centrosome cycle, protein maturation/proteolysis, and general reproduction. Notably, all of these major categories showed over-representation in comparison to expectation, demonstrating that the simple statistical procedure outlined above can effectively identify transcriptionally important genes that have remained functionally conserved in the dynamically and rapidly evolving environment of the *Drosophila* testis transcriptome.

Cell type clustering of testis scRNA-Seq data reveals increasing transcriptomic divergence throughout spermatogenesis

To validate the expression of the 198-gene set we identified, we visualized both UMAP and PCA projections of all analyzed cells with an overlay of previously identified cell type assignments (7)(Figure S3a). Importantly, cells from all species can be projected onto the same visualization using the data (subsetted by the 198-gene list) without extraneous processing beyond standard single-cell transcriptomics pipelines (Figure 1 d-f, Figure S3b). A visual examination of these projections shows a striking number of features concordant with known spermatogenesis function. While the sequence of spermatogenesis was not explicitly encoded into our statistical procedure or in the visualization process, the progression of cell types in the UMAP projection follows the known progression from germline stem cells to early and late spermatogonia, early and late spermatocytes, and finally to early and late spermatids. Despite

being a continuous developmental process where each stage does not produce spatially segregated clusters, the stages of the previous cell type assignments form relatively distinct boundaries, eliminating the need for more advanced pseudo-time projection techniques. In addition to the progression of germline differentiation, all somatic cell line types cluster separately from germ cell types. One further interesting feature of this projection is a gradual divergence in transcriptome expression associated with sperm maturation. While we are examining germline development rather than whole-organism development, this unexpected feature differs from more traditional "hourglass" models of whole-organism developmental processes (45–47), highlighting an intriguing aspect of expression variations among cells during spermatogenesis. Note that all of these features are readily observed in PCA projections of the same data set (Figure S3b).

Importantly, the application of this methodology allows for recapitulation of previously identified cell type clusters using unsupervised clustering on a unified, pan-species UMAP projection with the clustering resolution set to produce the same number of clusters reported in previous studies (7) (Leiden clustering, resolution = 3×10^{-4}, 9 clusters). While it is possible to be an artifact of species-specific sampling effects, these cell type assignments show that fewer cells are represented in the earlier germline stem cell, spermatogonia, and spermatocyte stages of *D. yakuba* when compared to *D. melanogaster.* We also observe a correlated accumulation of cells in later spermatid developmental stages. These cell type assignments also reveal the presence of an additional ancestral cell type present in *D. ananassae* that appears to be absent from *D. melanogaster* and *D. yakuba*. We also observe a low abundance of early and late spermatid stages in *D. ananassae* (Figure 1d-f, Figure S3b). Changes in cell type assignment and visualized patterns among species may reflect the expression patterns of the 198-gene list, technical biases during data generation, or fundamental processes driving species-specific differences in reproduction. To investigate these cell clusters further, we conducted experiments using RNA FISH, as described in the following sections.

High differential expression across cell types in D. melanogaster spermatogenesis does not necessarily indicate differential expression across cell types in closely related species

A cross-species analysis of the five most significantly differentially expressed genes for each cell type was performed. Interestingly, only one (*MtnA*) of thirteen previously reported markers (*aly*, *zfh1*, *Fas1*, *Hsp23*, *MtnA*, *His2Av*, *aub*, *bam*, *fzo*, *twe*, *soti*, *Dpy-30L2*, *p-cup*) (7)

remains as one of the top-5 differentially expressed genes by cell type (all spp., e.g. all GSC from all spp. considered together, Figure S4). These results suggest that while these genes may be highly differentially expressed across cell types in *D. melanogaster*, this highly differential expression is likely to have evolved in a species-biased manner. A lack of one-to-one orthology may also be excluded as a driving factor in these discrepancies, as only one of these thirteen markers (*zfh1*) was removed from the "melanogasterized" annotations in this analysis.

We then analyzed the full list of all significantly differentially expressed genes for each cell type, testing for both over- and under-expression (Dataset S2, Table S4). Among the twelve previously used markers present in the data set, *Fas1*, *fzo*, and *Dpy-30L2* were not found to be differentially expressed in any of these cell types (all spp.). In the somatic cells, *Hsp23* and *MtnA* remained highly differentially expressed with the correct cell type specificity. Similarly, early spermatogenesis markers did appear to be differentially expressed in GSC/Early spermatogonia cell types, but with highly attenuated signals. For example, of all 1326 significantly differentially expressed genes in GSC/Early spermatogonia, the most differentially expressed classic marker gene was *His2Av*, which was ranked 375 when sorted by log(fold change), while the next closest marker was *aub*, ranked 600. This attenuation of signal appears to continue through later stages of spermatogenesis, but reverses at the late spermatid stage. There, *p-cup* is ranked 18/639 and *soti* is ranked 24/639 respectively. Interestingly, the *ananassae*-specific cell type showed significant overexpression of both early and late spermatogenesis markers: *His2Av* (ranked 245/366) and *p-cup* (ranked 73/366), respectively. The attenuation of differential expression of previously utilized marker genes highlights how cell type-specific divergence contributed to the masking of proper cell type assignment between species.

Among the twelve genes in our data set, only *aly* appears in our 198-gene list, indicating that the remaining *D. melanogaster* marker genes' ability to predict cell type (above median *F* (cell type) values, c.f. Methods) likely occurs in a species-specific manner (above median *F* (species identity) values). Eight of the twelve marker genes did not appear in the 198-gene list as they showed high levels of predictive ability for species. Additionally, expression levels of four of the remaining twelve marker genes remained relatively similar across species. Of those, two genes (*aub*, *bam*) were excluded for having undefined *F* (cell type) values as a result of having little to no variance in certain cell types. This low/no variance occurs due to having uniformly few to no counts within different cell types. Interestingly, the remaining two genes (*Hsp23*,

MtnA) were not excluded for having evolutionarily diverged expression patterns, but they were excluded for having poor predictive ability for cell type (i.e., below-median *F* (cell type) values). This appears to contradict the appearance of *MtnA* in an analysis of the most highly differentially expressed genes across cell types (Table S4). The discrepancy arises from differences in statistical methodologies. *MtnA* expression is found only in a very small number of cells, primarily somatic cells (Figures S4). Differential expression analysis compares expression within one cell type (somatic cells for *MtnA*) vs expression across all other cell types ("one-vs-all"), revealing elevated *MtnA* expression in somatic cells despite its low detection in many other cell types. In contrast, ANOVA compares the between-cell type variance across all cell types, where *MtnA*'s limited detection results in a low F-statistic value due to minimal variance in most cell types. This highlights the value of performing cell type assignment on genes that show highly specific yet robust expression across cell types. This effect is particularly important in the context of scRNA-Seq technology, where the absence of detection may be as much the result of low read count as it is of low expression. Focusing on quantitative differences within widely expressed genes like those in our 198-gene list, rather than a reliance on non-detection events of relatively lowly expressed genes, may lead to more consistent cell type classification in future comparative scRNA-Seq studies.

RNA in-situ hybridization chain reaction validates cross-species cell type assignments

While we were able to transfer a reference species' labels across cell types both in our simulated data set and in our three-species data set, it was unclear whether the cell type assignments were grounded in biological reality. To determine the accuracy of the cell type assignments, we performed *in-situ* hybridization chain reactions (HCR RNA FISH) (48, 49) in whole-mount testis tissue from *D. melanogaster* and *D. ananassae*. Four genes were chosen: *Rbp4*, *B52*, *Pkd2*, and *soti* (Figure 2) to fulfill multiple criteria, such as highlighting predicted evolutionary divergence between species in common marker genes, demonstrating high cell-type specificity to aid interpretation of subsequent *in-situ* imaging, and to provide maximal information regarding our newly detected cell-type cluster. We note that during imaging, there was a large degree of crosstalk between the fluorophores representing *B52* and *soti*, so images were not amplified using the two fluorophores simultaneously nor were the same channels used for the same images during acquisition.

Rpb4 was chosen as it was a previously employed cell-type marker for *D. melanogaster* GSC/Early spermatogonia cell types in prior publications (25). Our cross-species cell type classification predicted that this marker gene had a highly divergent expression between *D. melanogaster* and *D. ananassae* (Figure 2A). This prediction was validated, as *Rbp4* showed an unexpectedly wide expression pattern beginning with early spermatogonia and culminating in expression overlap with late spermatids as confirmed by *soti* expression and DAPI staining (Figure 2B-E). This is a surprising finding, as *Rpb4*, an RNA-binding protein, was found to be a regulator of mitosis. *Rbp4* interacts with *Fest* to regulate the translation of *Cyclin B* (50). Similarly, the genes *Lut* and *Syp* interact with *Fest*, regulating post-transcriptional control of *Cyclin B* (51). We then investigated the cell type-specific expression of the genes *Cyclin B*, *Fest*, and *Syp* in our data set – (*Lut* is not present as a one-to-one ortholog in our *melanogaster*-ized transcriptome)(Figure S5). The expression pattern of *Cyclin B* agrees with previously reported expression patterns in *D. melanogaster* (50), while also adopting a similarly specific pattern in *D. yakuba*. Like *Rpb4*, the expression pattern of *Cyclin B* appears to be non-specific in *D. ananassae*, showing broad expression across all of spermatogenesis. While the expression patterns of *Fest* and *Syp* were less specific than for *Rbp4* and *Cyclin B*, this specificity also appeared to decrease in *D. ananassae* as well (Figure S5). This result also indicates that changes in cell type biased patterns of genes may reflect functional changes at the pathway level.

B52 was chosen because it is predicted to express in early spermatogonia as well as in our newly identified *ananassae*-specific cluster. Similarly, *Pkd2* was also chosen as it was predicted to have maximal expression in our *ananassae*-specific cluster in addition to some residual expression in late spermatogonia and early spermatocytes in *D. melanogaster*. As predicted, *B52* showed high specificity in germline stem cells and spermatogonia, with a residual expression detected in our *ananassae*-specific cell type, which we were able to confirm is a type of spermatocyte. Additionally, *B52* expression appeared to be less specific in *D. ananassae* than in *D. melanogaster*. The specificity of the new *ananassae* spermatocyte cell type was also confirmed by examining the expression pattern of *Pkd2*. Surprisingly, a larger degree of spermatocyte-biased expression was detected for *Pkd2* in *D. melanogaster,* reflecting the increased sensitivity of HCR RNA FISH.

Finally, *soti* was chosen as it is also a previously employed cell-type marker for *D. melanogaster* late spermatid cell types in prior publications (7, 11, 25, 26). Our cross-species cell

type classification predicted that transcript levels in *D. melanogaster* would both be more highly expressed and less specific than in *D. ananassae*. This prediction was validated using our *in-situ* imaging. For example, *D. melanogaster soti* expression appears to be broader than in *D. ananassae*, forming a broader, "smearier" expression pattern when compared to the tighter, more individualized puncta seen in *D. ananassae*.

Transcriptional bursting in spermatogenesis is controlled via opposing mechanisms during meiotic transition

We then used our validated cell type assignments to study the evolution of transcriptional control by examining how transcriptional bursting properties of different genes change over time. General transcriptional activity first increases, then dramatically decreases the stage of spermatogenesis between the 16-spermatocytes stage and the second round of meiosis. These divisions subsequently result in 64 round spermatids in all three species (35). Importantly, the total pool of remaining transcripts is slowly degraded as they are translated in subsequent stages, supplemented only by a small number of genes that are transcribed post-meiotically. While the increased resolution of scRNA-Seq analysis (23, 52) and MS2-MS2 coat protein system imaging (21, 22) has brought significant advances in understanding the fundamental molecular mechanisms underlying transcription, shedding light on the chemical kinetics that drives stochastic gene expression, global control of this pre- to post-meiotic transition during *Drosophila* spermatogenesis has remained under studied.

We begin by calculating the burst size and frequency for all annotated genes within our annotated cell types in *D. melanogaster* using txburst (23). We then repeated these calculations for tissue derived from *D. yakuba*, and *D. ananassae* using FlyBase reference annotations (Figure 3). After applying default txburst quality control criteria, the number of genes that were analyzed per cell type varied by cell type and species (Table S5). Interestingly, we find that in the early, pre-meiotic stages of spermatogenesis (i.e., germline stem cells, early spermatogonia, late spermatogonia, and early spermatocytes), transcriptional bursting is primarily controlled via alterations in burst size. While burst frequency also increases during these stages, overall burst size appears to reach a maximum during the late spermatogonia stage before declining during spermatocyte stages (Figure 3a, b). Interestingly, in *D. melanogaster*, burst size drops sharply between the early and late spermatocyte stages, while in *D. yakuba* and *D. ananassae*, burst size

decreases between the spermatogonia and spermatocyte stages but remains relatively stable from early to late spermatocyte stages (Figure 3a-d). This result suggests that phenotypic divergence during spermatogenesis in *Drosophila* may be evolving via global, genome-level alterations of transcriptional control via pol II initiation ('k_{ini}'), rather than through alterations in a smaller subset of genes.

The control of bursting frequency appears to become the more dominant mechanism during post-meiotic stages of spermatogenesis (i.e. late spermatocyte, and spermatid stages, Figure 3b). A correlated increase in bursting frequency in *D. melanogaster* is observed in the late spermatocyte stage simultaneously with a decrease in burst size. A similar correlated change is observed between late spermatogonia and early spermatocyte in *D. yakuba*, where a decrease in burst size occurs simultaneously with an increase in bursting frequency (Figure 3c, d). In *D. ananassae*, an increase in burst frequency is observed in the early spermatocyte stage followed by an even larger increase in the late spermatocyte stage (Figure 3e, f). While the increase in burst frequency during the transition between late spermatogonia and early spermatocyte stages is correlated with a decrease in burst size in the same stages, the second increase of burst frequency during the late spermatocyte stage does not show a correlated change in burst size. This increase in burst frequency is potentially an outlier, as not only did a very low number of annotated genes pass quality control measures (202 genes, Figure S6C), but also the variance in the observed distribution is also quite high. These observations could also reflect a deeper evolutionary divergence in cell type, resulting from a weaker cell type assignment as revealed by its sparse representation in UMAP and PCA projections (Figure 1e). If the inferred distribution is closer in value to distributions found for the early spermatocytes and *D. ananassae* spermatocyte cell types, it would appear that *D. melanogaster* may have undergone a species-specific alteration where the size-to-frequency control transition has been delayed from the early spermatocyte stage to the late spermatocyte stage.

Overall, these results are consistent with a model of transcriptional control where an abundance of pol II is present during earlier stages of spermatogenesis and becomes depleted during the meiotic transition. During the period of high overall pol II availability and activity, general levels of transcription remain high, and bursting kinetics are dominated by increased burst size. Subsequent to this depletion, transcription then decreases significantly, shifting global transcriptional control into a more burst frequency-dominated mode of activity. It is likely that

this modulation of burst frequency is achieved by changes in transcription factor stoichiometry, as is commonly observed with enhancer sequences (20, 22, 53).

Bursting kinetics of X-linked vs autosomal genes reveals that partial dosage compensation likely occurs in somatic and germline tissue

To help elucidate the timing of dosage compensation during spermatogenesis, we examined the transcriptional bursting kinetics for both X-linked (N=2198) and autosomal (N=11749) protein-coding genes using 'txburst' (23). Interestingly, our data set shows a highly consistent pattern where transcriptional burst size and frequency are significantly lower for X-linked genes than for autosomal genes across cell types and species (Mann-Whitney U test, Benjamini-Hochberg correction, FDR < 0.05) (Figure 4). Perhaps most surprisingly, while significant differences between X-linked and autosomal bursting kinetics remain consistent throughout spermatogenesis, a significant difference was also detected for somatic tissue in *D. melanogaster*. While this pattern of lower activity for X-linked genes than autosomal genes continues even when examining cell types that lack the statistical power to produce a significant difference, the transcriptional bursting frequency during the early spermatocyte stage in *D. yakuba* is significantly higher for X-linked genes than in autosomal genes. Additionally, while nearly every cell type in *D. melanogaster* and *D. yakuba* produces a significant difference between X- and autosomal-linked genes' transcriptional kinetics, the observed differences in *D. ananassae* are typically not significant after multiple hypothesis correction. This is likely because the number of genes that pass quality control in txburst is much lower for *D. ananassae* than for *D. melanogaster* and *D. yakuba*, resulting from the relatively poor reference genome quality for *D. ananassae* vs *D. melanogaster* and *D. yakuba* (Table S1).

The ratio between X-linked and autosomal genes passing quality-control remains relatively constant across all stages and species (23) (Figure 4, Figure S6, Table S5), reflecting the number of genes that are expressed in each stage. The consistent ratio of X to autosomal genes expressed suggests that there is no particular stage of spermatogenesis during which the entire X chromosome is entirely inactivated. This lack of bias persists even despite a strong stage-specific decrease in the total number of genes passing quality-control (a proxy for total genes expressed) after either the early spermatocyte stage in *D. melanogaster* or the late spermatogonia stage in *D. yakuba* and *D. ananassae* (Figure S6). As this lack of bias persists even when comparing

somatic cell types to germline cell types (Figure 4), it is likely that the process of X-downregulation occurs prior to the establishment of somatic vs. germline identity during the differentiation of testis tissue (54). This pattern could potentially be further tested by comparing the transcriptional bursting characteristics of X-linked to autosomal genes in earlier stages of testis organogenesis, e.g., in larval testis tissue.

Application of independent lift-over methodologies reveals evolutionarily young testis-specific transcription through positive identification of transcriptional inactivity using bulk RNA-Seq samples

One significant technical difficulty in the identification of new tissue-specific transcripts arising *de novo* from previously non-transcribing regions of the genome is in the positive identification of non-detection of these transcripts in outgroup species. More specifically, the absence of homologous annotations in outgroup species, even when annotation methodologies are kept constant across species (55) as in the case of FlyBase (56), may not necessarily be used as positive evidence of absence. Thus, to positively identify a set of transcripts gaining transcription in testis tissues within the *D. melanogaster*-*D. yakuba*-*D. ananassae* divergences, we applied two orthogonal methodologies to lift over an annotation file that included *D. melanogaster* annotations from both FlyBase and a list of previously identified *de novo* genes (retrieved from (7)) onto the *D. simulans*, *D. sechellia*, *D. erecta*, *D. yakuba* and *D. ananassae* genomes from FlyBase (56)(Methods and Materials). We thus identified a set of 170 genes (156 annotated in FlyBase) whose transcription activated in the *Drosophila* testis in a species- or clade-specific manner (Dataset S3), including 58 *melanogaster*-specific, 55 shared between *D. melanogaster*, *D. simulans*, and *D. sechellia*, and 57 shared between *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. erecta* and *D. yakuba*. Of these transcripts, 32 are annotated in FlyBase as protein-coding genes, 102 are annotated in FlyBase as lncRNAs, 14 were previously identified *de novo* variants or genes, and 22 are annotated in FlyBase as specific types of ncRNAs like tRNAs, siRNAs, etc.

We note that this methodology does not distinguish between transcripts that may or may not be transcribed in non-testis tissue in ancestral states due to a lack of such data. Similarly, it is also possible that some of these transcripts identified through this methodology may be a result of loss events affecting *D. ananassae*, particularly in the 57 genes shared between *melanogaster*

and *yakuba*, however this would be impossible to distinguish as without further ancestral data sets. This would thus explain the presence of certain evolutionarily conserved genes in this 170-gene list of testis-activated transcripts including *let-7* and *Jhe*. However, rather than excluding these genes based on results from manual searches with arbitrary criteria, we have included them in this data set for completion, particularly as our evidence demonstrates that these genes have likely activated in *D. melanogaster* testis tissue in a species- or clade-specific manner. Despite the 170-gene list containing a number of older, conserved genes, the vast majority of the list likely consists of transcripts that have activated *de novo* from previously inactive sequence as only 31 and 28 of these *D. melanogaster* genes have orthologs identified by FlyBase in *D. yakuba* and *D. ananassae* respectively.

Origination of new transcripts alters regulation of neighboring genes by increasing burst size during meiotic transition

The cultivator model of *de novo* gene origination hypothesizes that the most evolvable method of altering the expression of a pre-existing genes is the appearance of a new proximal promoter element (40). Unlike changes to existing transcription-factor binding sites, as enhancer sequences typically act on longer range distances (~1.5+ kb), the cultivator model posits that the appearance of a promoter would act on relatively short ranges (~100-500 bp), likely through enhancer-mediated promoter interference (21) or supercoiling-mediated transcriptional coupling (57). One key prediction of this model would be that ncRNAs are found to be closer to their protein-coding neighbors than expected by chance – this is indeed the case (40). Another key prediction of this model is that the origination of a *de novo* transcript would alter the regulation of its immediate neighboring genes, but not more distant neighbors as might be expected by enhancers. To test this prediction, we examined the transcriptional bursting characteristics of protein-coding genes that are neighboring newly activated *de novo* transcripts using our scRNA-Seq data set.

We begin by identifying four neighboring protein-coding genes for all 170 transcripts identified through bulk RNA-Seq analysis, two upstream and two downstream ("L2," "L1," "R1," and "R2" respectively, Figure 5). Note that the "L2" and "L1" genes are defined to be to the 5' of the *de novo* transcript, regardless of the sense or anti-sense orientation of the transcript. Using these neighboring genes' orthologous genes in *D. yakuba* and *D. ananassae*, we compared

the burst size and frequency for each class of neighbors to genome-wide distributions. Unlike the 170 evolutionarily young *de novo* transcripts identified through our comparative analysis of bulk RNA-Seq data, the neighboring protein-coding genes are not expected to be biased in an evolutionary age.

Consistent with the cultivator model, the origination of these *de novo* transcripts is associated with a significant increase in transcriptional burst size in their 1st upstream neighboring genes, but not further neighboring genes. These differences are also found concurrently with the meiotic transition, suggesting that these genes may be functionally important in later stages of spermatogenesis. More specifically, significant differences are found either in the early spermatocyte (*D. melanogaster*, $p=5.17 \times 10^{-4}$ and *D. yakuba*, $p=4.56 \times 10^{-4}$, Mann-Whitney U-test) and late spermatocyte (*D. yakuba*, $p=8.25 \times 10^{-4}$) stages. These results remain significant after correction for multiple hypothesis testing using a Benjamini-Hochberg procedure (4 neighbors x 8 clusters x 3 species = 96 hypotheses, FDR < 0.05; Figure 5, Figure S7). Importantly, the cultivator model also predicts that there is no significant difference in transcriptional bursting kinetics prior to the origination of these *de novo* transcripts as these neighboring genes pre-date the appearance of the new *de novo* transcripts in the outgroup species, *D. ananassae*. Consistent with this hypothesis, no significant differences in burst size are detected in any neighboring genes. Furthermore, no significant differences are detected in burst frequency in any neighbors at any stage in any species. Given an increase in burst size but no alteration of burst frequency, these results suggest that a promoter-based effect, rather than an enhancer-based mechanism, drives the *cis*-regulatory effect of new promoter birth.

To further validate this finding, we compared the genomic distribution of mean read count per cell with the same distributions calculated for L2, L1, R1, and R2 neighboring gene groups (Figure S8, Table S6). While many significant differences in mean read count per cell were found, most of these significant differences persisted between in-group (*D. melanogaster* and *D. yakuba*) and out-group (*D. ananassae*) comparisons, or had opposing directions between species (e.g. significantly increased in *D. melanogaster* and decreased in *D. yakuba*). Only three neighboring group/cell type combinations showed significant differences within in-group species and while also being directionally consistent: L1 neighbors in early spermatocytes, R1 neighbors in early spermatocytes, and R1 neighbors in late spermatids.

In alignment with our previous finding of increased burst size, we find significantly elevated expression of L1 neighboring genes in early spermatocytes when compared to genomic expression (Figure S8). Interestingly, we also find significantly decreased expression of R1 neighboring genes in both early spermatocytes and early spermatids when compared to genomic expression. The combination of increased L1 neighboring gene expression as well as decreased R1 neighboring gene expression is particularly striking as it is consistent with supercoiling-mediated coupling (57). The appearance of a new promoter element in the tandem upstream/downstream conformations as shown in Figure 5A would be expected to generate under-winding just upstream of and over-winding just downstream of the new promoter. This would result in both increased transcription in the upstream gene (in this case, the L1 neighbor), while decreasing transcription of the downstream gene (in this case, the R1 neighbor).

Discussion

Feature selection through ANOVA allows for the identification of conserved cell type-specific genes

Our work demonstrates how the identification of the evolutionarily conserved 198-gene list allowed for cell type classifications from *D. melanogaster* to be extended to a panel of closely related *Drosophila* species. While it was utilized in this study to overcome cell type-specific evolution, the methodology used here should be generalizable to a large range of problems where conditions and transcriptome expression vary in a correlated manner. For example, the transcriptional activity of specific cell types may exhibit cyclical fluctuations throughout the day, potentially leading to cell type-specific differences that are confounded with circadian rhythms. Alternatively, strong cell type-specific responses to differing treatment regimes could similarly obscure cell type assignments. The application of a similar ANOVA-based methodology could be utilized to identify genes that differ strongly between cell types but remain robust to changes in time-of-day or treatment category.

Interestingly, the 198-gene list that we identified did not show large overlap with more traditionally utilized cell type markers in *D. melanogaster* while still being significantly enriched for previously identified spermatogenesis-related function. One potential interpretation of this observation is that while the markers in the 198-gene list do have known cell type-specific function, it is possible that they do not perform as well as other, more classic markers within *D.*

melanogaster. Alternatively, it is also possible that the list of classic marker genes, while they served very well over multiple generations of studies, has been assembled one at a time like building blocks and reflects more of the history of the discoveries made in *D. melanogaster* rather than the global transcriptional progression through spermatogenesis in the *Drosophila melanogaster* species group. Regardless of the reason, our work has shown that the utilization of these marker genes in *D. melanogaster* does not necessarily imply that they are the most optimal markers when examining multiple species' genomes. Given the increasing popularity of SAMap (29), it appears likely that an increasing number of researchers will be faced with similar technical difficulties. We believe that the utilization of our ANOVA-based methodology will provide a simple solution to these difficulties that will aid future scRNA-Seq cell type classifications.

Biological differences in cell type classification

One lingering question in regard to our cell type classifications is whether our newly identified *ananassae* spermatocyte cluster represents different biology. A possibility is that this difference in classification arises from asymmetric divergence times between *D. melanogaster*, *D. yakuba*, and *D. ananassae*. For example, our modeling strategy is agnostic to divergence time or phylogenetic structure. Similarly, there is still some residual species-specific signal that should be present in our subsetted gene list. Consequently, it is also possible that these differences in cell type classification are artefactual. On the other hand, it is also possible that either primary or secondary spermatocytes have multiple subtypes at various stages and types of differentiation, thereby creating a distinguished cell cluster at the expression level.

When we further examined the *ananassae* spermatocyte cluster, however, we did find biological evidence that there is meaningful transcriptional divergence in these cells. For example, while these cells are poorly represented in *D. melanogaster* and *D. yakuba*, we found unusual patterns in our differential expression analysis in our full data set. Recall that we found significant differential expression of *His2Av*, an early spermatogenesis marker in *D. melanogaster*, as well as *p-cup*, a late spermatogenesis marker in *D. melanogaster*. Notably, neither of these genes is found in our 198-gene list. Such transcriptional divergence during the spermatocyte stage suggests that the rapid evolution of the testis may be dominated by changes

during this stage, reminiscent of earlier observations that *de novo* genes tend to express most highly in spermatocytes (7).

We also found that *Rpb4* has a surprisingly broad expression pattern in *D. ananassae*. This provokes the interesting question of whether this broad expression pattern is reflective of the ancestral state or evolved in an ancestor of *D. ananassae* subsequent to *D. melanogaster* speciation. It is likely that investigation into the upstream transcriptional differences that drive the alteration of *Rpb4* expression would lead to more general insight into the evolution of transcriptional programming in the testis. Similarly, investigation into the downstream effects of broad *Rpb4* expression would likely be fruitful. *Rpb4*, in conjunction with *fest* and *Syp*, has been shown to regulate *Cyclin B* during meiosis in *D. melanogaster* (50, 51). However, *Rbp4* and *Cyclin B* have much higher cell type specificity in *D. melanogaster* than *D. ananassae*, but in slightly different tissue types. How did these expression patterns evolve? Is the evolutionary appearance of this specificity coupled between the two genes? Alternatively, did this loss of specificity occur only in the *D. ananassae* lineage? These questions require further investigation in the future.

We also note that our cross-species categorization detected an absence of spermatids in our scRNA-Seq data. However, our *in-situ* results showed highly specific expression of *soti*, while our brightfield microscopy revealed cells that can clearly be identified as late spermatids based on morphology alone. Interestingly, the gross morphology of these *D. ananassae* spermatids appeared to diverge from *D. melanogaster* as well, as they appeared to have a "stringier" and "fuzzier" appearance in in addition to downstream morphological differences observed in the accessory gland, where tissue appeared to be larger and whiter. As the total amount of transcripts becomes increasingly depleted as spermatids fully mature, the transcriptional programming controlling these phenotypic differences likely occurs in earlier stages, e.g. *anannassae* spermatocytes. Consequently, the downstream biological effects of this altered transcriptional programming will likely be opaque to transcription-based techniques. This highlights the need for different methods of inquiry, such as the application of newer proteomic technology, or the continued application of more traditional genetic tools in outgroup species beyond *D. melanogaster*.

Transcriptional bursting kinetics offers insight into male sex chromosome regulation

Many unanswered questions remain in regard to the mechanisms that drive decreased X-chromosome expression during spermatogenesis, leading to a large variety of proposed models. For example, differential polymerase elongation rates have been suggested as a mechanism underlying dosage compensation in males (58). Alternatively, a more canonical pathway involving the lncRNAs *roX1*, *roX2*, and the male-specific lethal complex (MSL) is possible (59, 60). One attractive aspect of the canonical MSL pathway is that recent results have shown that the intrinsically disordered C-terminal region of *MSL2* and the *roX* lncRNAs induce X chromosome compartmentalization that functions similarly to phase separation (61). This has led to the suggestion that this X chromosome compartment could be enriched for pol II, resulting in increased initiation of transcription and an increase in transcriptional burst size (24). However, it has also been suggested that elements of the MSL complex, such as *maleless*, do not associate with the X chromosome during spermatogenesis (62).

Our results demonstrate a clear transition from a burst size-dominated mode of regulation to a burst frequency-dominated mode during the meiotic transition, suggesting that this transition may be driven by a global decrease in polymerase availability. However, in alignment with a prior analysis of previous results (11), our analysis of transcriptional bursting kinetics demonstrates that significantly decreased burst size and burst frequency for X-linked genes precedes not only the meiotic transition, but likely also the establishment of germline stem cells. This decreased burst size is consistent with a study demonstrating decreased pol II binding on X chromosomes derived from both isolated spermatocytes and dissociated testis in comparison to wing imaginal discs (63). While these results are consistent with some manner of incomplete dosage compensation (11, 64), our consistent observation of both decreased burst size and burst frequency suggests that a depletion of pol II localized to the X chromosome alone is not a complete explanation for the observed X-linked decrease in expression. More specifically, a local depletion of pol II would decrease transcriptional burst size without altering burst frequency as observed in the developed embryo (24). Instead, the as-yet unknown mechanism for decreased X chromosome expression should instead affect both the binding rate as well as the initiation rate of pol II to X chromosome-linked promoter sequences.

Methods and Materials

Statistics and simulation

Information regarding statistical modeling, sampling procedure, classification framework, feature selection, and quality assessment are provided in the Supplementary Information.

scRNA-Seq library preparation and sequencing

Testis from *D. melanogaster*, *D. yakuba*, and *D. ananassae* were dissected in ice-cold PBS before being transferred to ice. Dissociation was performed via desheathing using a lysis buffer consisting of 2% collagenase (100mg/ml stock) in 1X TrypLE buffer. Samples were centrifuged, then mildly vortexed, and then incubated at room temperature for 10 minutes. This procedure was repeated three times (30 minutes total), then passed through 35 µm cell strainers. The resulting samples were centrifuged for 7 minutes at 1200 rpm at 4°C. The resulting pellet was washed with 200 µL of cold HBSS and centrifuged again for 7 minutes at 1200 rpm at 4°C. The resulting pellet was suspended in 35 µL of cold HBSS. 5 µL of the resulting single cell suspension was stained with 5 µL of trypan blue for cell counting and the remaining 30 µL of suspension was used for library preparation and sequenced using the 10X platform by the Rockefeller University Genomics Resource Center.

scRNA-Seq pipeline

Raw data was processed using cellranger (v7.1.0) and aligned to FlyBase genomes (dmel-r6.44, dyak-r1.05, and dana-r1.06). Resulting count matrices were loaded into Monocle3, where only "melanogasterized" genes with one-to-one orthologs reported by FlyBase were retained. Ribosomal genes, tRNA genes, and mitochondrial genes were also excluded from further analysis. These data sets were preprocessed and aligned using Monocle3. Differential expression analyses were performed in Seurat (v 4.1.3) using cell type assignments performed in Monocle3. Further information may be found in the Supplementary Information.

In situ hybridization chain reaction (HCR RNA FISH)

Dissection of testis tissue followed (65), hybridization chain reaction protocols from (49) were adapted for *Drosophila* testis and oligo pools/probes were generated using HCRProbeMakerCL v2.0 (https://github.com/rwnull/HCRProbeMakerCL). Specific probe

sequences paired with amplifier information are provided in Table S7. Further information may be found in the Supplementary Information.

Bulk RNA-Seq pipeline and identification of de novo transcripts

We used both HALPER (66), a HAL-based (67) methodology that relies on ProgressiveCactus (68) multi-genome alignments, and LiftOff (69), a minimap2-based (70) methodology which does not require multi-genome alignments and thus does not rely on synteny.

We then obtained a panel of publicly available bulk testis RNA-Seq data sets from these species (SRR1024003, SRR1024002, SRR1617567, SRR486109, SRR5839446, SRR16213299, SRR16213300, SRR6667440, SRR6667439, SRR931549) and calculated TPM values for both annotations generated through HALPER and LiftOff. This approach thus allows us to identify the homologous regions in outgroup species for transcripts, allowing us to positively identify non-detection events in outgroup species.

Inference of transcriptional bursting properties

Raw cell counts were exported into .csv files for each cell type. Each cell count file was then input into 'txburstML.py' from (23). Resulting .pkl files were then converted and then imported into R. Genes failing quality-control from 'txburstML.py' were discarded, and orthology calls were made using FlyBase orthology information ("dmel_orthologs_in_drosophila_species_fb_2022_01"). Inferred burst size and burst frequency values were further analyzed in R.

Data and code availability

ScRNA-Seq reads were upload to NCBI under the project number PRJNA995212. Scripts for analyses may be found at https://github.com/LiZhaoLab/scripts_ul_2024.

Acknowledgements

We thank Helen Duan and Connie Zhao from The Rockefeller University Genomics Resource Center for the help with scRNA-Seq library preparation, RRID:SCR_020986. Confocal microscopy was performed in the Rockefeller University's Bio-Imaging Resource Center, RRID:SCR_017791. We thank the Zhao lab members for helpful discussion. UL would like to

thank Huang Suqi for helpful discussions regarding statistics. The figures were partly constructed using BioRender.

Author contribution

U.L., N.S., and L.Z. conceived the study, designed the experiments, and formulated the analyses. C.L., C.B.L, and N.S. generated the data. U.L. conducted all computational analyses. U.L. and L.Z. wrote the paper.

Funding

This work was supported by National Institutes of Health (NIH) MIRA R35GM133780, the Robertson Foundation, and an Allen Distinguished Investigator Award from Paul G. Allen Family Foundation to L.Z, and National Science Foundation postdoctoral fellowship 2410289 to U.L. The content of this study is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

Reference cited

1. M. T. Levine, C. D. Jones, A. D. Kern, H. A. Lindfors, D. J. Begun, Novel genes derived from noncoding DNA in Drosophila melanogaster are frequently X-linked and exhibit testis-biased expression. *Proceedings of the National Academy of Sciences* **103**, 9935–9939 (2006).

2. S. Chen, Y. E. Zhang, M. Long, New genes in Drosophila quickly become essential. *Science* **330**, 1682–1685 (2010).

3. A. M. Gubala, *et al.*, The Goddard and Saturn Genes Are Essential for Drosophila Male Fertility and May Have Arisen De Novo. *Mol Biol Evol* **34**, 1066–1082 (2017).

4. L. Zhao, P. Saelao, C. D. Jones, D. J. Begun, Origin and Spread of de Novo Genes in Drosophila melanogaster Populations. *Science* **343**, 769–772 (2014).

5. J. M. Cridland, A. C. Majane, H. K. Sheehy, D. J. Begun, Polymorphism and Divergence of Novel Gene Expression Patterns in Drosophila melanogaster. *Genetics* **216**, 79–93 (2020).

6. M. Long, N. W. VanKuren, S. Chen, M. D. Vibranovski, New gene evolution: little did we know. *Annual Review of Genetics* **47**, 307–333 (2013).

7. E. Witt, S. Benjamin, N. Svetec, L. Zhao, Testis single-cell RNA-seq reveals the dynamics of de novo gene transcription and germline mutational bias in Drosophila. *eLife* **8**, e47138 (2019).

8. M. D. Vibranovski, Meiotic Sex Chromosome Inactivation in Drosophila. *J Genomics* **2**, 104–117 (2014).

9. W. Hense, J. F. Baines, J. Parsch, X Chromosome Inactivation during Drosophila Spermatogenesis. *PLoS Biol* **5**, e273 (2007).

10. C. D. Meiklejohn, E. L. Landeen, J. M. Cook, S. B. Kingan, D. C. Presgraves, Sex Chromosome-Specific Regulation in the Drosophila Male Germline But Little Evidence for Chromosomal Dosage Compensation or Meiotic Inactivation. *PLOS Biology* **9**, e1001126 (2011).

11. E. Witt, N. Svetec, S. Benjamin, L. Zhao, Transcription Factors Drive Opposite Relationships between Gene Age and Tissue Specificity in Male and Female Drosophila Gonads. *Molecular Biology and Evolution* **38**, 2104–2115 (2021).

12. C. Kemkemer, W. Hense, J. Parsch, Fine-Scale Analysis of X Chromosome Inactivation in the Male Germ Line of Drosophila melanogaster. *Molecular Biology and Evolution* **28**, 1561–1563 (2011).

13. H. H. McAdams, A. Arkin, Stochastic mechanisms in gene expression. *PNAS* **94**, 814–819 (1997).

14. A. Raj, A. van Oudenaarden, Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. *Cell* **135**, 216–226 (2008).

15. I. Golding, J. Paulsson, S. M. Zawilski, E. C. Cox, Real-Time Kinetics of Gene Activity in Individual Bacteria. *Cell* **123**, 1025–1036 (2005).

16. L. Cai, N. Friedman, X. S. Xie, Stochastic protein expression in individual cells at the single molecule level. *Nature* **440**, 358–362 (2006).

17. J. P. Bothma, *et al.*, Dynamic regulation of eve stripe 2 expression reveals transcriptional bursts in living Drosophila embryos. *Proceedings of the National Academy of Sciences* **111**, 10598–10603 (2014).

18. A. Raj, C. S. Peskin, D. Tranchina, D. Y. Vargas, S. Tyagi, Stochastic mRNA Synthesis in Mammalian Cells. *PLOS Biology* **4**, e309 (2006).

19. D. M. Suter, *et al.*, Mammalian Genes Are Transcribed with Widely Different Bursting Kinetics. *Science* **332**, 472–474 (2011).

20. M. C. Walters, *et al.*, Enhancers increase the probability but not the level of gene expression. *Proc Natl Acad Sci U S A* **92**, 7125–7129 (1995).

21. T. Fukaya, B. Lim, M. Levine, Enhancer Control of Transcriptional Bursting. *Cell* **166**, 358–368 (2016).

22. B. Zoller, S. C. Little, T. Gregor, Diverse Spatial Expression Patterns Emerge from Unified Kinetics of Transcriptional Bursting. *Cell* **175**, 835-847.e25 (2018).

23. A. J. M. Larsson, *et al.*, Genomic encoding of transcriptional burst kinetics. *Nature* **565**, 251–254 (2019).

24. L. Forbes Beadle, H. Zhou, M. Rattray, H. L. Ashe, Modulation of transcription burst amplitude underpins dosage compensation in the Drosophila embryo. *Cell Rep* **42**, 112382 (2023).

25. A. A. Raz, *et al.*, Emergent dynamics of adult stem cell lineages from single nucleus and single cell RNA-Seq of Drosophila testes. *eLife* **12**, e82201 (2023).

26. E. Witt, C. B. Langer, N. Svetec, L. Zhao, Transcriptional and mutational signatures of the Drosophila ageing germline. *Nat Ecol Evol* **7**, 440–449 (2023).

27. S. V. Nuzhdin, M. L. Wayne, K. L. Harmon, L. M. McIntyre, Common Pattern of Evolution of Gene Expression Level and Protein Sequence in Drosophila. *Molecular Biology and Evolution* **21**, 1308–1317 (2004).

28. S. Khodursky, N. Svetec, S. M. Durkin, L. Zhao, The evolution of sex-biased gene expression in the Drosophila brain. *Genome Res.* (2020). https://doi.org/10.1101/gr.259069.119.

29. A. J. Tarashansky, *et al.*, Mapping single-cell atlases throughout Metazoa unravels cell type evolution. *eLife* **10**, e66747 (2021).

30. Y. Hao, *et al.*, Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).

31. L. Haghverdi, A. T. L. Lun, M. D. Morgan, J. C. Marioni, Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* **36**, 421–427 (2018).

32. I. Korsunsky, *et al.*, Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* **16**, 1289–1296 (2019).

33. R. R. Snook, T. A. Markow, T. L. Karr, Functional nonequivalence of sperm in Drosophila pseudoobscura. *Proc Natl Acad Sci U S A* **91**, 11222–11226 (1994).

34. D. Joly, A. Korol, E. Nevo, Sperm size evolution in Drosophila: inter- and intraspecific analysis. *Genetica* **120**, 233–244 (2004).

35. L. Schärer, J.-L. Da Lage, D. Joly, Evolution of testicular architecture in the Drosophilidae: A role for sperm length. *BMC Evolutionary Biology* **8**, 143 (2008).

36. Y. M. Yamashita, M. T. Fuller, D. L. Jones, Signaling in stem cell niches: lessons from the Drosophila germline. *Journal of Cell Science* **118**, 665–672 (2005).

37. C. Lim, L. Tarayrah, X. Chen, Transcriptional regulation during Drosophila spermatogenesis. *Spermatogenesis* **2**, 158–166 (2012).

38. Z. A. Alvi, T.-C. Chu, V. Schawaroch, A. V. Klaus, Protamine-like proteins in 12 sequenced species of Drosophila. *Protein Pept Lett* **20**, 17–35 (2013).

39. C.-H. Chang, I. Mejia Natividad, H. S. Malik, Expansion and loss of sperm nuclear basic protein genes in Drosophila correspond with genetic conflicts between sex chromosomes. *Elife* **12**, e85249 (2023).

40. U. Lee, S. Mozeika, L. Zhao, A Synergistic, Cultivator Model of De Novo Gene Origination. *Genome Biology and Evolution* evae103 (2024). https://doi.org/10.1093/gbe/evae103.

41. X. Qiu, *et al.*, Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* **14**, 979–982 (2017).

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

43. H. Mi, A. Muruganujan, J. T. Casagrande, P. D. Thomas, Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc* **8**, 1551–1566 (2013).

44. P. D. Thomas, *et al.*, PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Science* **31**, 8–22 (2022).

45. D. Duboule, Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Dev Suppl* 135–142 (1994).

46. A. T. Kalinka, *et al.*, Gene expression divergence recapitulates the developmental hourglass model. *Nature* **468**, 811–814 (2010).

47. R. A. Raff, *The Shape of Life: Genes, Development, and the Evolution of Animal Form* (University of Chicago Press, 2012).

48. H. M. T. Choi, *et al.*, Third-generation in situ hybridization chain reaction: multiplexed, quantitative, sensitive, versatile, robust. *Development* **145**, dev165753 (2018).

49. H. S. Bruce, *et al.*, Hybridization Chain Reaction (HCR) In Situ Protocol. (2021).

50. C. C. Baker, B. S. Gim, M. T. Fuller, Cell type-specific translational repression of Cyclin B during meiosis in males. *Development (Cambridge, England)* **142**, 3394 (2015).

51. C. C. Baker, *et al.*, Cell-type-specific interacting proteins collaborate to regulate the timing of Cyclin B protein expression in male meiotic prophase. *Development (Cambridge, England)* **150**, dev201709 (2023).

52. G. Hunt, *et al.*, Tissue-specific RNA Polymerase II promoter-proximal pause release and burst kinetics in a Drosophila embryonic patterning network. *Genome Biology* **25**, 2 (2024).

53. S. Alamos, *et al.*, Minimal synthetic enhancers reveal control of the probability of transcriptional engagement and its timing by a morphogen gradient. *Cell Syst* **14**, 220-236.e3 (2023).

54. S. Mahadevaraju, *et al.*, Dynamic sex chromosome expression in Drosophila male germ cells. *Nat Commun* **12**, 892 (2021).

55. C. M. Weisman, A. W. Murray, S. R. Eddy, Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes. *Current Biology* **32**, 2632-2639.e2 (2022).

56. L. S. Gramates, *et al.*, FlyBase: a guided tour of highlighted features. *Genetics* **220**, iyac035 (2022).

57. C. P. Johnstone, K. E. Galloway, Supercoiling-mediated feedback rapidly couples and tunes transcription. *Cell Rep* **41**, 111492 (2022).

58. E. Larschan, *et al.*, X chromosome dosage compensation via enhanced transcriptional elongation in Drosophila. *Nature* **471**, 115–118 (2011).

59. A. Franke, B. S. Baker, The rox1 and rox2 RNAs Are Essential Components of the Compensasome, which Mediates Dosage Compensation in Drosophila. *Molecular Cell* **4**, 117–122 (1999).

60. T. Conrad, A. Akhtar, Dosage compensation in Drosophila melanogaster: epigenetic fine-tuning of chromosome-wide transcription. *Nat Rev Genet* **13**, 123–134 (2012).

61. C. I. K. Valsecchi, *et al.*, RNA nucleation by MSL2 induces selective X chromosome compartmentalization. *Nature* **589**, 137–142 (2021).

62. L. Rastelli, M. I. Kuroda, An analysis of *maleless* and histone H4 acetylation in *Drosophila melanogaster* spermatogenesis. *Mechanisms of Development* **71**, 107–117 (1998).

63. J. T. Anderson, S. Henikoff, K. Ahmad, Chromosome-specific maturation of the epigenome in the Drosophila male germline. *eLife* **12**, RP89373 (2023).

64. K. H.-C. Wei, K. Chatla, D. Bachtrog, Single cell RNA-seq in Drosophila testis reveals evolutionary trajectory of sex chromosome regulation. [Preprint] (2022). Available at: https://www.biorxiv.org/content/10.1101/2022.12.07.519494v1 [Accessed 25 April 2024].

65. L. Shao, *et al.*, Eukaryotic translation initiation factor eIF4E-5 is required for spermiogenesis in Drosophila melanogaster. *Development* **150**, dev200477 (2023).

66. X. Zhang, I. M. Kaplow, M. Wirthlin, T. Y. Park, A. R. Pfenning, HALPER facilitates the identification of regulatory element orthologs across species. *Bioinformatics* **36**, 4339–4340 (2020).

67. G. Hickey, B. Paten, D. Earl, D. Zerbino, D. Haussler, HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29**, 1341–1342 (2013).

68. J. Armstrong, *et al.*, Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* **587**, 246–251 (2020).

69. A. Shumate, S. L. Salzberg, Liftoff: accurate mapping of gene annotations. *Bioinformatics* **37**, 1639–1643 (2021).

70. H. Li, Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

Figure Legends

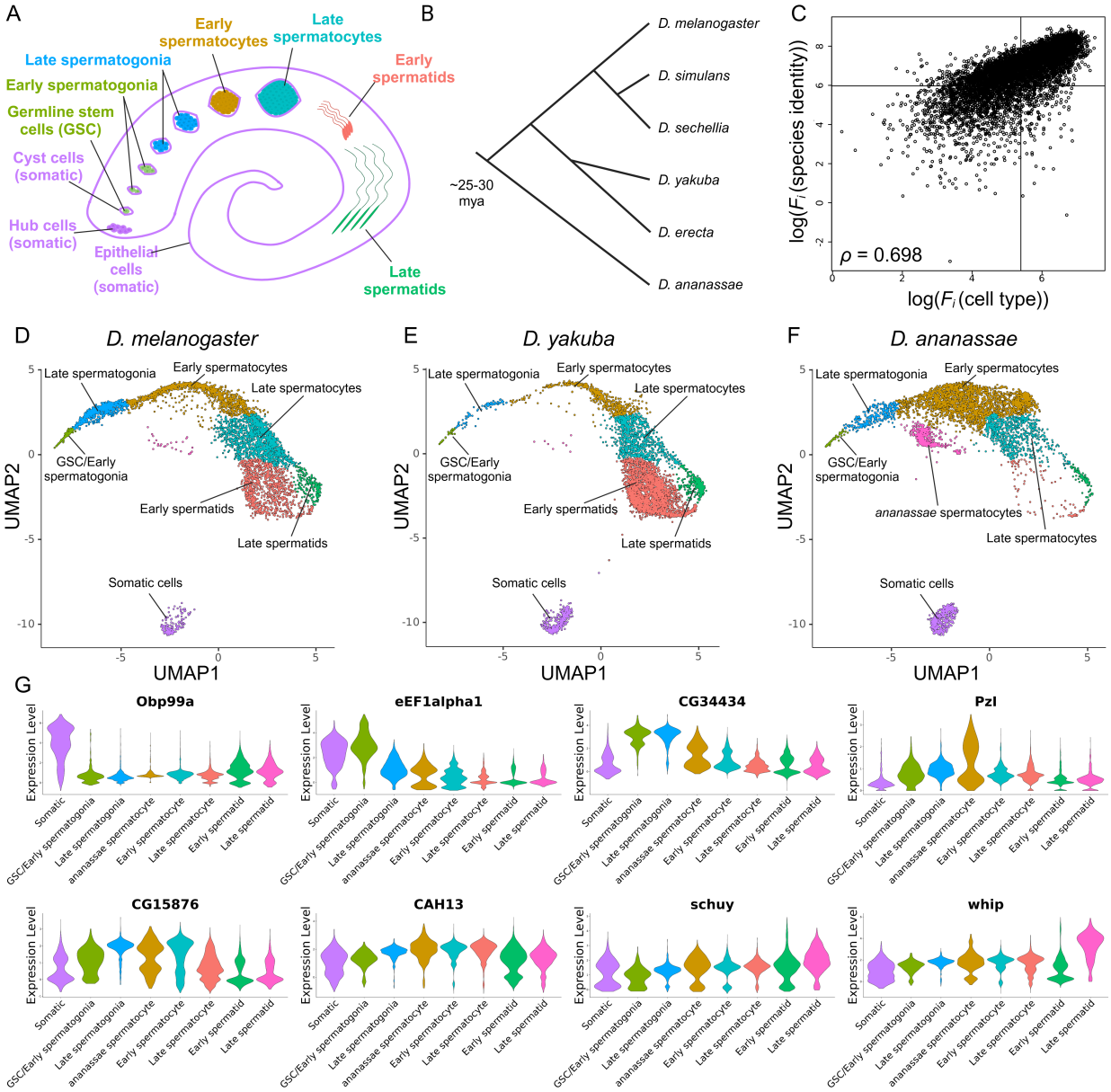

Figure 1: Identification of 198-gene list allows for cell type classification across species in a rapidly diverging process. **(A)** Cartoon figure of spermatogenesis progression, beginning with germline stem cells and finishing in late spermatids. **(B)** Phylogeny of species used in this study with estimated divergence times. **(C)** The joint distribution of ANOVA log(F-statistic) values derived from scRNA-Seq data using previously identified *D. melanogaster* cell types and cross-species labels from *D. melanogaster*, *D. yakuba*, and *D. ananassae* tissue shows a high level of correlation resulting in a large degree of species-specific cell type divergence. Drawn lines correspond to median *F*-statistic values for each (independent) marginal distribution. **(D-F)** UMAP projections resulting from the 198-gene list reveal a progression of intermediate cell types spanning germline stem cells to late spermatids from *D. melanogaster*, *D. yakuba*, and *D. ananassae*. **(G)** Violin plots from the most differentially expressed gene within each cell type.

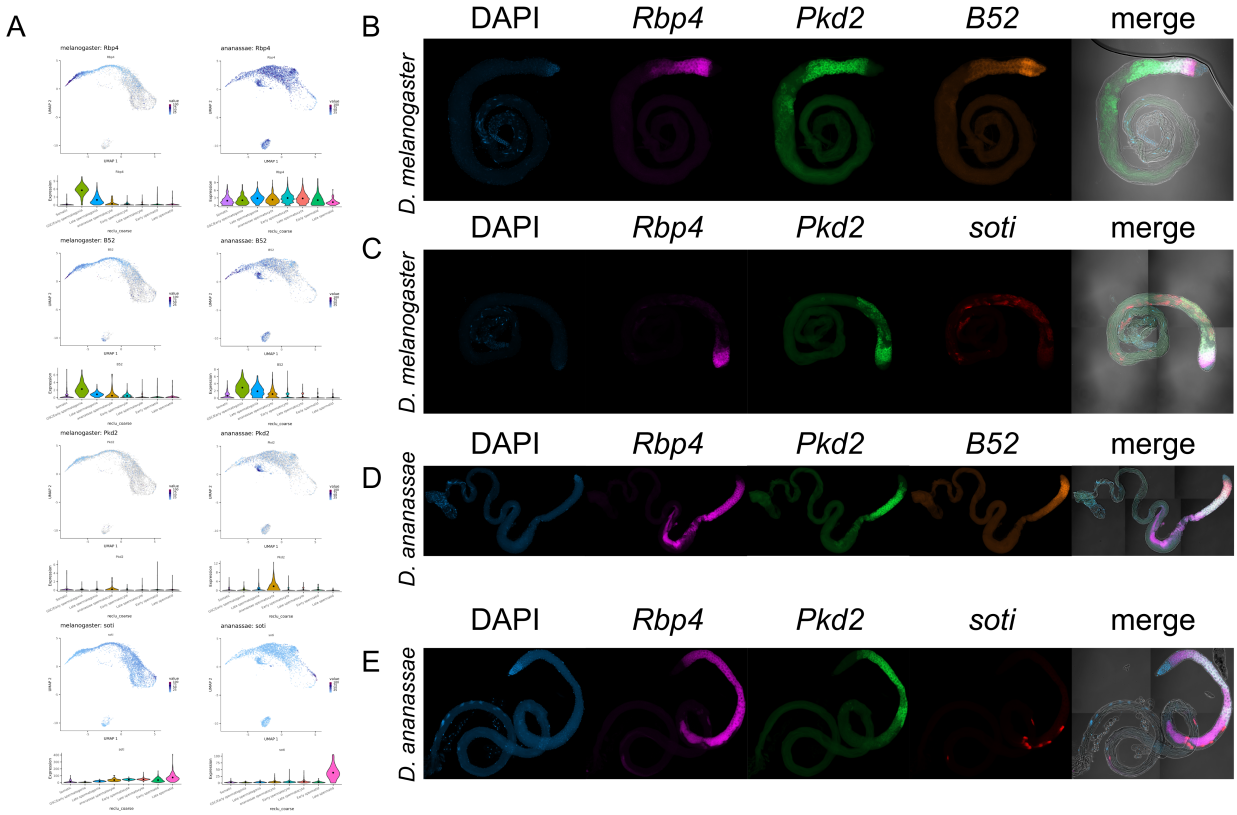

Figure 2: Validation of expression pattern divergence and cell type classification using HCR RNA FISH. **(A)** scRNA-Seq UMAP and predicted expression patterns in *D. melanogaster* and *D. ananassae* for four genes: *Rbp4*, *B52*, *Pkd2*, and *soti*. **(B-E)** Imaging of RNA FISH performed in *D. melanogaster* **(B, C)** and *D. ananassae* **(D, E)** using these same genes. Nuclear staining with DAPI (**B-E**, left column) and overlay of these channels over brightfield microscopy are also presented (**B-E**, right column). *Rbp4* is a marker gene commonly used in *D. melanogaster* to indicate germline stem cells/spermatogonia. In contrast, scRNA-Seq cell type classification using the conserved 198-gene list predicts that *Rbp4* has a broad expression pattern in *D. ananassae*. *B52* is a gene predicted to have a restricted expression pattern in *D. melanogaster* germline stem cells/spermatogonia. This expression pattern is predicted to be slightly broader in *D. ananassae*, with its pattern extending into the newly identified *D. ananassae* spermatocyte cluster. *Pkd2* is predicted to have an expression pattern that is primarily restricted to the newly identified *D. ananassae* spermatocyte cluster in both *D. melanogaster* and *D. ananassae*. *soti* is a marker gene commonly used in *D. melanogaster* to indicate spermatid cell types. scRNA-Seq cell type classification using the conserved 198-gene list predicts that *soti* has both higher overall expression levels and lower cell type specificity in *D. melanogaster* than in *D. ananassae*.

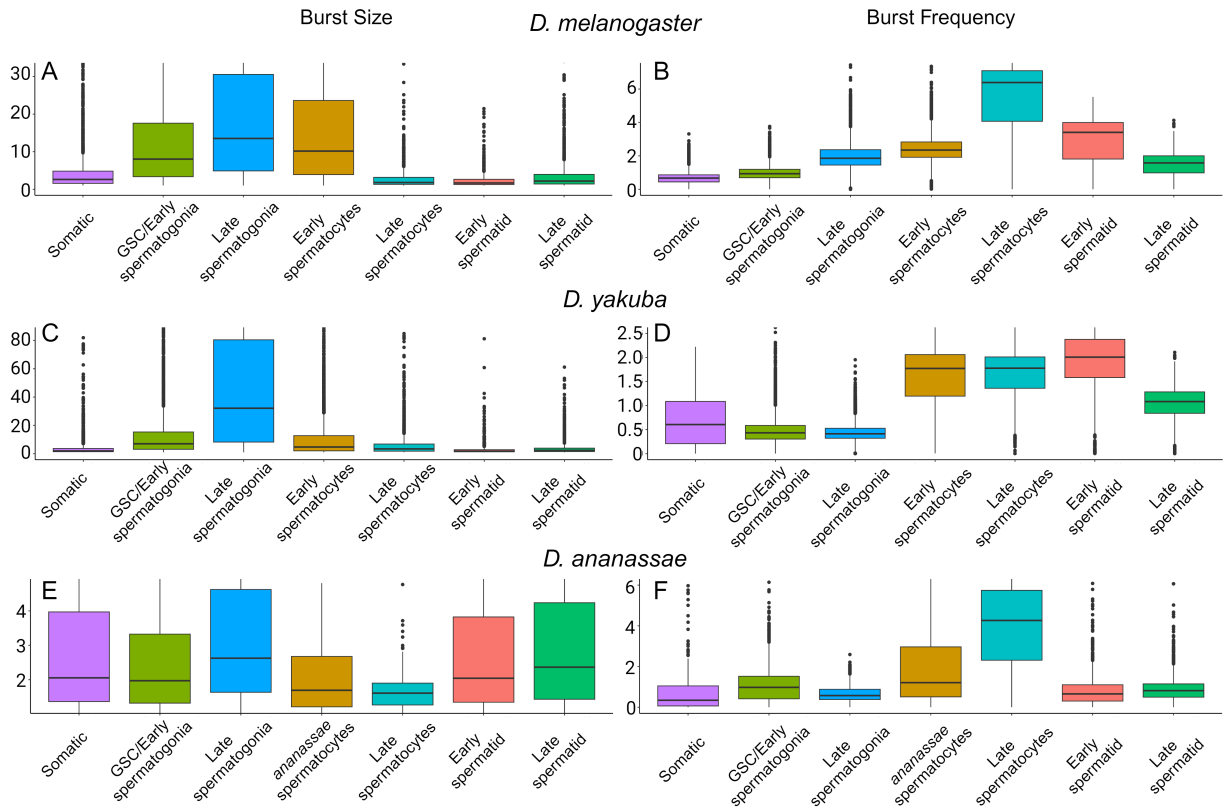

Figure 3: Pre- and post-meiotic cell types in spermatogenesis are controlled by opposing mechanisms. The burst size and burst frequency distributions for **(A, B)** *D. melanogaster*, **(C, D)** *D. yakuba*, and **(E, F)** *D. ananassae* reveal that transcription in the earlier states of spermatogenesis are primarily controlled via burst size, while transcription in later stages of spermatogenesis is primarily controlled via burst frequency. This transition between control mechanisms correlates with the transition between mitosis and meiosis during spermatogenesis.

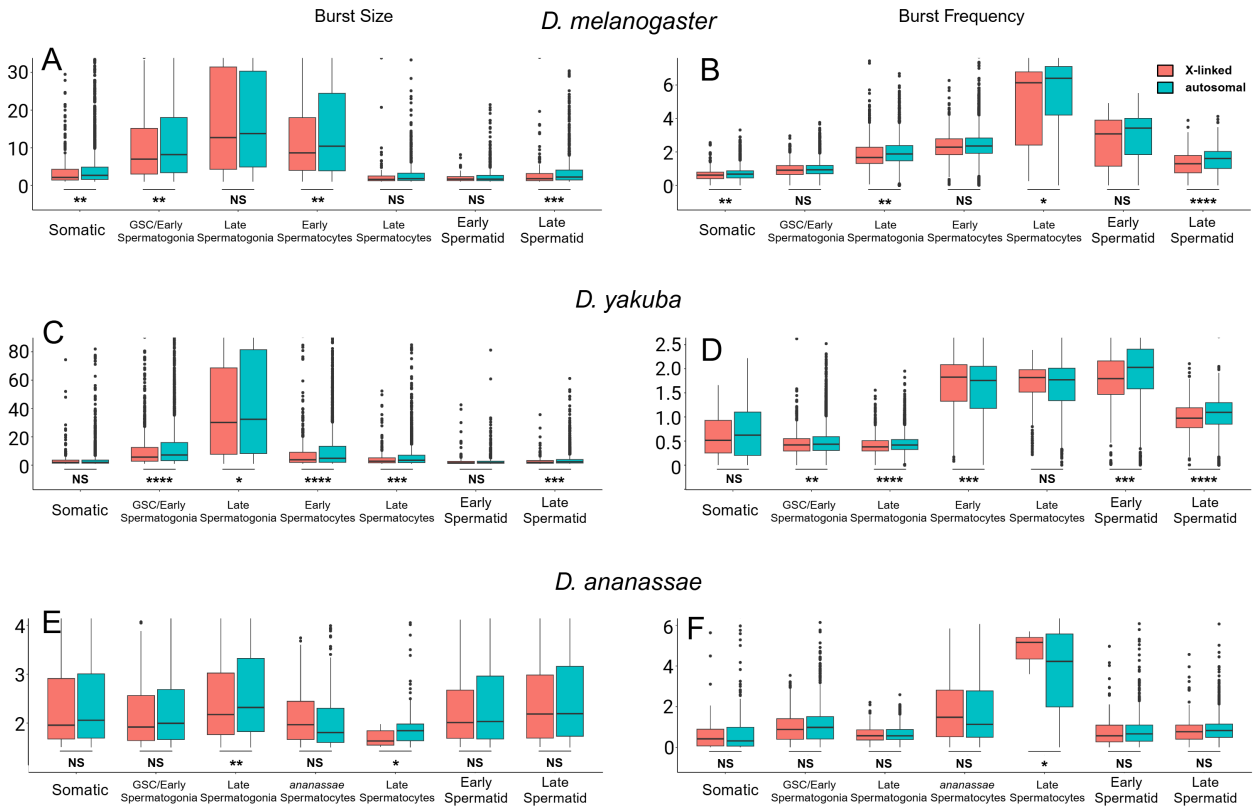

Figure 4: X-linked genes show lower bursting kinetics than autosomal genes prior to spermatogenesis. The burst size and burst frequency distributions for **(A, B)** *D. melanogaster*, **(C, D)** *D. yakuba*, and **(E, F)** *D. ananassae* reveal a conserved pattern where X-linked genes show significantly smaller average burst sizes and lower bursting frequencies than autosomal genes across cell types (* = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$. Note these are p-values prior to application of the Benjamini-Hochberg, FDR < 0.05). There is no stage of spermatogenesis at which X-linked genes begin to show lower burst kinetics than autosomal genes. The significant repression of X-linked genes in somatic cells suggests that the onset of decreased meiotic X-chromosome expression occurs in a developmental stage preceding testis maturation.

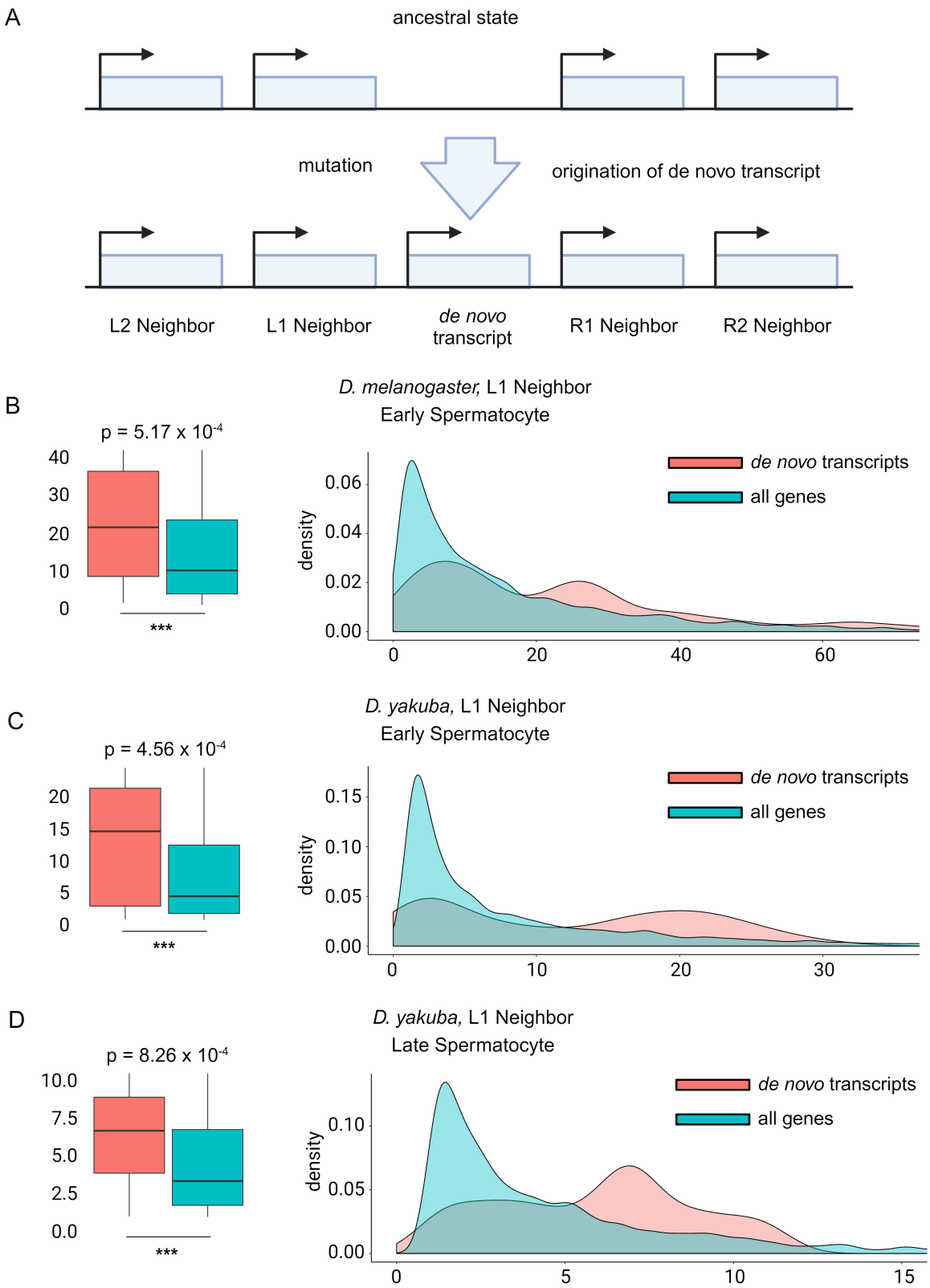

Figure 5: Significantly higher burst sizes of L1 neighboring genes of newly evolved *de novo* transcripts support a cultivator model of *de novo* gene origination. (A) The initial steps of

the cultivator model of *de novo* gene origination posit that the origination of a new promoter element near a pre-existing gene may serve to regulate nearby genes in *cis*. This may be tested by comparing the bursting kinetics of genes immediately neighboring evolutionarily new *de novo* transcripts to further neighboring genes. Note that in our study, *de novo* transcripts include those that may have arisen specifically in *D. melanogaster*, as well as those shared across the *D. melanogaster-D.simulans* or the *D. melangaster-D.yakuba* clades. **(B-D)** L1 neighboring genes just upstream of new *de novo* transcripts show significantly higher burst sizes than the distribution of all genes during the early and late spermatocyte stages, coinciding with the onset of meiosis. These differences remain significant after the application of a Benjamini-Hochberg correction (FDR < 0.05). No other significant differences are found when comparing the distributions of further neighboring genes or within spermatocytes preceding the origination of the *de novo* transcript (Figure S7).