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 simple 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, contrasting with traditional hourglass models of whole-organism development. With these cross-species cell type classifications, 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 the transcriptional bursting properties of neighboring genes during key stages of spermatogenesis.

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 (Levine *et al.* 2006; Chen, Zhang and Long 2010; Gubala *et al.* 2017), the study of the evolutionary mechanisms underlying the origination of *de novo* transcripts such as lncRNA in fruit flies has been lacking (Levine *et al.* 2006; Zhao *et al.* 2014; Cridland *et al.* 2020). 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* (Levine *et al.* 2006; Long *et al.* 2013; Zhao *et al.* 2014). Intriguingly, scRNA-Seq data has shown that *de novo* originated genes, unlike recently evolved duplicate genes, are biased toward expression in mid-spermatogenesis (Witt *et al.* 2019). 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 (Vibranovski 2014). While prior work has suggested that expression of genes on the X chromosome is globally down-regulated during spermatogenesis in *Drosophila* (Hense, Baines and Parsch 2007), it is not possible to determine anything beyond the general reduction of expression for X-linked genes (Meiklejohn *et al.* 2011; Vibranovski 2014). 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 (Witt *et al.* 2021). Alternatively, 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 (Kemkemer, Hense and Parsch 2011).

Transcription is an intrinsically random process which may be modeled stochastically (McAdams and Arkin 1997; Raj and van Oudenaarden 2008). Single molecule observations of transcriptional activity is intrinsically "bursty," consisting of short bouts of transcription that produce discrete quantities of mRNAs ("on" state) followed by periods of inactivity ("off" state) (Golding *et al.* 2005; Cai, Friedman and Xie 2006; Bothma *et al.* 2014). 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 (Golding *et al.* 2005; Raj *et al.* 2006; Suter *et al.* 2011). 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 (Walters *et al.* 1995; Fukaya, Lim and Levine 2016; Zoller, Little and Gregor 2018; Larsson *et al.* 2019). 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* (Forbes Beadle *et al.* 2023). 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 (Forbes Beadle *et al.* 2023), 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* (Raz *et al.* 2023; Witt *et al.* 2023) 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 genes expression changes and differences drastically vary in *Drosophila* (Nuzhdin *et al.* 2004; Khodursky *et al.* 2020), making it problematic to simply group cell types cross 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 (Tarashansky *et al.* 2021), 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' (Hao *et al.* 2021), Monocle 3's use of 'batchelor' (Haghverdi *et al.* 2018), and the independent Harmony software (Korsunsky *et al.* 2019) 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 (Snook, Markow and Karr 1994; Joly, Korol and Nevo 2004; Schärer, Da Lage and Joly 2008). Alternatively, a large degree of spermatogenesis (Figure 1a) has remained unaltered through similar evolutionary time scales (Yamashita, Fuller and Jones 2005; Lim, Tarayrah and Chen 2012), 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 (Alvi *et al.* 2013; Chang, Mejia Natividad and Malik 2023). 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 (Supplemental 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 (Lee, et al., under review), 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

To determine whether species-specific cell type transcriptomic divergence significantly hinders cell type classification using standard scRNA-Seq analysis software suites, we generated a simulated data

set consisting of nine unique cell types evolving in three different species (Methods, Supplemental Figure S2a). We then applied a simple methodology to calculate ANOVA F-statistic scores for cell type and species (Supplemental Figure S2b). The methodology was then able to capture the most informative genes (Supplemental Figure S2c). When standard batch correction techniques, such as Monocle3's implementation of batchelor via 'align_cds()' (Qiu *et al.* 2017; Haghverdi *et al.* 2018) , 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 (Supplemental Figure S2e). Interestingly, 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 (Supplemental 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.

Projection and clustering for the most-informative gene list showed distinct cell type clusters that still maintained species-specific differences (Supplemental Figure S2f, g). Cell type identities are easily recovered in this visualization, while species-specific evolution within these cell types is also easily visualized. This method identifies gene lists that are predictive of cell type but not species identity, which are important for clustering cell types across multiple species.

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

We further 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 (Suvorov *et al.* 2022)(Figure 1b). Previously published data and cell type labels from *D. melanogaster* were obtained from (Witt *et al.* 2019), 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

(Qiu *et al.* 2017), 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, Table S2). 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 (Supplemental 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, dotted 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). This is due to the fact that for certain genes in the "melanogasterized" gene set i, F_i (cell type) or F_i (species identity) – but not both – were undefined due to variances equaling exactly equal to zero. This definition of $p_{cell\text{-}type}$ and $p_{species\text{-}identity}$ produced a new gene set i' with cardinality $|i'| = 198$ (Table S3).

To validate the biological significance of our functionally conserved 198-gene set, we performed a gene ontology (GO) analysis using the PANTHER database (Mi *et al.* 2013; Thomas *et al.* 2022)(Table S4). 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, spermatogenesis (category), and general reproduction. Importantly, all of these major categories showed over-representation in comparison to expectation, demonstrating how the simple statistical procedure outlined above can 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

As a positive control 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 (Witt *et al.* 2019)(Supplementary Figure S3a). Importantly, all cells from all species are able to be projected onto the same projection using the subsetted data without extraneous processing beyond standard single-cell transcriptomics pipelines (Figure 1 d-f, Supplementary 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 cell types related to gametogenesis. One further interesting feature of this projection is a gradual divergence in transcriptome expression correlating with sperm maturation. This unexpected feature stands in contrast to more traditional "hourglass" models of whole-organism developmental processes (Duboule 1994; Kalinka *et al.* 2010; Raff 2012). Note that all of these features are easily observed in PCA projections of the same data set (Supplemental 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 (Witt *et al.* 2019) (Leiden clustering, resolution = 3×10^{-4}, 9 clusters). These cell type assignments show that gametogenic progression in *D. yakuba* testis is potentially accelerated in comparison to *D. melanogaster*. While it is possible this is an artifact of species-specific sampling effects, we find that fewer cells are represented in the earlier germline stem cell, spermatogonia, and spermatocyte stages of *D. yakuba* when compared to *D. melangaster.* 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 is absent from *D. melanogaster* and *D. yakuba*. Interestingly, correlated alterations in early and late spermatocyte stages and the near absence of early and late spermatid stages in *D. ananassae* spermatogenesis suggest that this ancestral cell type represents a different spermatid cell type that is distinct from early and late spermatid cell types in *D. melanogaster* and *D. yakuba*. The evolution of distinct spermatid cell types, in combination with the late-stage transcriptional divergence observed across all three species, hints at more general molecular mechanisms underlying species divergence (Figure 1d-f, Supplementary Figure 3b). These results demonstrate how transcriptomic control of cell type-specific evolution via genes controlling sperm mobility, protein maturation and degeneration, and insulin signaling in mature spermatids (Table S4) may be a fundamental process driving species-specific differences in reproductive processes.

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

An 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*) (Witt *et al.* 2019) remains as one of the

top-5 differentially expressed genes by cell type (Supplementary 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-specific 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. The remaining twelve genes from the (Witt *et al.* 2019) marker gene list were present in all downstream analyses.

Only one of the remaining twelve genes (*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 value)). Eight of the twelve marker genes did not appear in the 198-gene list as they showed high levels of predictive ability for species. Alternatively, 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. However, a crucial difference in statistical methodologies between differential expression analysis and an analysis of variance (ANOVA) underlies these results.

The key distinction appears because *MtnA* expression is found only in a very small number of cells (Supplementary Figures S4). On one hand, differential expression analysis compares expression within one cell type (somatic cells in the case of *MtnA*) vs expression across all other cell types ("one-vs-all"). In this case, the inferred low expression of *MtnA* resulting from a lack of detection in many cell types contrasts with the elevated expression of *MtnA* in somatic cells, indicating a high degree differential expression of *MtnA*. Alternatively, the ANOVA compares the between-cell type variance across all cell types. In this case, the lack of detection of *MtnA* in many cells results in a low F-statistic value, as many cell types have low to no variance. This discrepancy between statistical methodologies highlights the value of performing cell type assignment on genes that show highly specific expression while still maintaining 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. We believe it is possible that the utilization of 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, will lead to more consistent cell type classification in future comparative scRNA-Seq studies.

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

General transcriptional activity first increases, then dramatically decreases during the stage of spermatogenesis when 16 spermatocytes undergo two rounds of meiosis. These divisions subsequently result in 64 round spermatids in all three species (Schärer, Da Lage and Joly 2008). 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 (Larsson *et al.* 2019; Hunt *et al.* 2024) and MS2-MS2 coat protein system imaging (Fukaya, Lim and Levine 2016; Zoller, Little and Gregor 2018) 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-explored.

We begin by calculating the burst size and frequency for all annotated genes within our annotated cell types (c.f. *Cell type clustering of testis scRNA-Seq data reveals increasing transcriptomic divergence throughout spermatogenesis*) in *D. melanogaster* using txburst (Larsson *et al.* 2019). We then repeated these calculations for tissue derived from *D. yakuba*, and *D. ananassae* using FlyBase reference annotations (Figure 2). 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 2a, b). Interestingly, while burst size in *D. melanogaster* declines precipitously between early and late spermatocyte stages, burst size declines between spermatogonia and spermatocyte stages then remains relatively constant between early and late spermatocyte stages in *D. yakuba* and *D. ananassae* (Figure 2a-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.

Interestingly, 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 2a, b). 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 2c, d). Alternatively, an increase in burst frequency is observed in the early spermatocyte stage in *D. ananassae*, followed by an even larger increase in the late spermatocyte stage (Figure 2e, 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, Supplementary Figure S5C), 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* spermatid 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 (Walters *et al.* 1995; Zoller, Little and Gregor 2018; Alamos *et al.* 2023).

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' (Larsson *et al.* 2019). 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 3). Perhaps most surprisingly, while significant differences between X-linked and autosomal expression remain consistent throughout spermatogenesis, a significant difference was also detected for somatic tissue in *D. melanogaster* and *D. yakuba*. 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 due to the fact that 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 S3).

The ratio between X-linked and autosomal genes passing quality-control for each stage in each species remains relatively constant for stages in all species (Larsson *et al.* 2019) (Figure 3, Supplementary Figure S5, Table S5), which is reflective of the number of genes that are expressed in each stage. A lack of bias in the relative ratio of X to autosomal genes expressed in each stage suggests that there is no particular stage of spermatogenesis during which the entire X chromosome is entirely inactivated. This lack of bias continues 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* (Supplementary Figure S5). This marked decrease in burst size and frequency is consistent with well-characterized cessation of transcription during the meiotic transition of *Drosophila* spermatogenesis. The lack of bias between X-linked and autosomal genes persists even despite global transcriptional changes that shift control of transcriptional kinetics from a primarily burst sized dominated mode to a burst frequency dominated mode. As this lack of bias persists even when comparing somatic cell types to germline cell types (Figure 3), 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 (Mahadevaraju *et al.* 2021). 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 identification of new tissue-specific transcripts arising *de novo* from previously non-transcribing regions of the genome is in 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 (Weisman, Murray and Eddy 2022) as in the case of FlyBase (Gramates *et al.* 2022), 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 (Witt *et al.* 2019)) onto the *D. simulans*, *D. sechellia*, *D. erecta*, *D. yakuba* and *D. ananassae* genomes from FlyBase (Gramates *et al.* 2022)(c.f. 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 (Table S6), 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 (Lee, et al., under review). Unlike in the case of alteration of existing transcription-factor binding sites, as in the case of enhancer sequences which 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 (Fukaya, Lim and Levine 2016) or supercoiling-mediated transcriptional coupling (Johnstone and Galloway 2022). As a result, a key prediction of this

model is that the origination of a *de novo* transcript would alter 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 4). 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 4, Supplementary Figure S6). 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.

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 reflect 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 (Tarashansky *et al.* 2021), 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.

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 (Larschan *et al.* 2011). Alternatively, a more canonical pathway involving the lncRNAs *roX1*, *roX2* and the male-specific lethal complex (MSL) is possible (Franke and Baker 1999; Conrad and Akhtar 2012). 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 (Valsecchi *et al.* 2021). 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 (Forbes Beadle *et al.* 2023). However, it has also been suggested that elements of the MSL complex, such as *maleless* do not associate with the X chromosome during spermatogenesis (Rastelli and Kuroda 1998).

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 (Witt *et al.* 2021), 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 (Anderson, Henikoff and Ahmad 2023). While these results are consistent with some manner of incomplete dosage compensation (Witt *et al.* 2021; Wei, Chatla and Bachtrog 2022), 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 (Forbes Beadle *et al.* 2023). 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

Statistical procedure

These were represented by 9,999 simulated genes within 2700 simulated cells. These cells were distributed across each simulated species such that each species contained 900 cells each with 100 cells of each of the 9 cell types. Cell type information was distributed evenly across genes (1111 genes per cell type) by sampling between two different distributions of count values, represented as data matrix $\boldsymbol{C}$, where C_{ij} and $i = 1, 2, \ldots, 9{,}999$ represents each gene and $j = 1, 2, \ldots, 900$ represents each cell. Additionally, cell type identities are encoded in the vector $c_j \in$ {cell type 1, cell type 2, … cell type 9}. Similarly, species information was also independently distributed across these genes (3333 genes per species), represented as data matrix S_{ij}. Species identity is also encoded in the vector $s_j \in$ {species 1, species 2, species 3}. These two matrices were then combined by multiplication with a random value β_i (drawn from a beta distribution between 0 and 1 with shape parameters $\alpha = \beta = 7$) and $1-\beta_i$ respectively, such that the final data matrix $\boldsymbol{D}$ where $D_{ij} = \beta_i C_{ij} + (1-\beta_i)S_{ij}$ .This value β_i thus represents the relative information content regarding cell type vs species passed on to the final simulated data matrix, resulting in the data structure seen in Supplementary Figure S2a (c.f. Methods and Materials). Values of β_i closer to 0 thus represent genes with more information content for cell type identity than species while β_i values closer to 1 represent genes with more information content regarding species than cell type identity.

Importantly, UMAP and PCA projection of matrix $\boldsymbol{D}$ shows 27 distinct clusters that separate according to both cell type and species, with clusters grouping more closely by species than by cell type (Supplementary Figure S2d).

While many genes contained information regarding both cell type identity as well as species identity, recall that the random value β_i dictated the degree to cell type and species information was composited into the final expression value for gene i. We hypothesized that we may recover cell type-specific clusters while still maintaining species-specific information by clustering on the highest dimensions using a UMAP or PCA transformation by using a subset of genes with high between cell type variance and low between species variance (i.e. genes with low β_i values). However, rather than performing statistical inference to estimate β_i for each gene, we calculated F-statistics resulting from one-sided ANOVA tests for each gene i against cell type label vector $c_{j'}$ and species vector s_j. Importantly, as cell type identity in the testis scRNA-Seq data exists only for a subset of samples $j' \in j$ where $s_{j'} \in$ {species 1} – that is to say that cell type information only exists for one species (e.g. *D. melanogaster*) – the calculated F-statistics for cell type is over cell type label vector $c_{j'}$ and not c_j. The resulting distributions for the F-statistics F_i (cell type) and F_i (species identity) respectively were then examined to generate a candidate gene list for further clustering analyses (Supplementary Figure S2d). Specifically, to identify genes with high predictive power for cell type identity and low predictive power for species identity (i.e. cell type-specific genes that have been evolutionarily conserved), we selected a list of 200 genes $i' \in i$ such that $F_{i'}$ (cell type) $\geq$ Q(F_i (cell type), $p_{cell\text{-}type}$) and $F_{i'}$ (species identity) $\leq$ Q(F_i (species identity), $p_{species\text{-}identity}$) where is Q($F$, p) is the quantile function for distribution F and percentile p. In this case, $p_{cell\text{-}type} = p_{species\text{-}identity}$ were chosen as the highest possible percentile to produce i' such that cardinality $|i'| = 200$.

To validate whether gene set i' contained more cell type information than species identity information, we examined the distributions of $\beta_{i'}$ and β_i (Supplementary Figure S4c), thereby confirming that our outlined statistical procedure could, in fact, identify functionally conserved genes that retain cell type-specific information. Importantly, as we did not utilize β_i in order identify gene set i', this procedure is non-parametric and inference-free, allowing it to be applied widely across different data sets. We subsequently performed UMAP transformation over subsetted data $\boldsymbol{D'}=D_{i'j}$ (Supplementary Figure S2e-g)

scRNA-Seq library preparation and sequencing

Testis from *D. melanogaster*, *D. yakuba*, and *D. ananassae* were dissected in ice-cold PBS before transfer 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 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. Count matrices were combined and cell type labels from a previous study (Witt *et al.* 2019) were imported, and species labels were applied. ANOVA F-statistics were calculated for each gene as described in *Statistical procedure* and the resulting 198-gene list was retained. Raw counts were then extracted for the 198-gene list in Monocle3 then subsequently combined, preprocessed, and aligned using default settings. Dimensional reduction was performed and leiden clustering was performed using default settings with $3x10^{-4}$ resolution. This resulted in 9 clusters, where two clusters overlapping with early spermatocytes were combined and two other clusters overlapping with early spermatid were also combined. Differential expression analyses were performed in Seurat (v 4.1.3) using cell type assignments performed in Monocle3.

Bulk RNA-Seq pipeline and identification of de novo transcripts

We used both HALPER (Zhang *et al.* 2020), a HAL-based (Hickey *et al.* 2013) methodology that relies on ProgressiveCactus (Armstrong *et al.* 2020) multi-genome alignments, and LiftOff (Shumate and Salzberg 2021), a minimap2-based (Li 2018) 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 (Larsson *et al.* 2019). Resulting .pkl files were then converted 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. We thank the Zhao lab members for helpful discussion. 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. 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

Alamos S, Reimer A, Westrum C *et al.* Minimal synthetic enhancers reveal control of the probability of transcriptional engagement and its timing by a morphogen gradient. *Cell Syst* 2023;**14**:220-236.e3.

Alvi ZA, Chu T-C, Schawaroch V *et al.* Protamine-like proteins in 12 sequenced species of Drosophila. *Protein Pept Lett* 2013;**20**:17–35.

Anderson JT, Henikoff S, Ahmad K. Chromosome-specific maturation of the epigenome in the Drosophila male germline. Struhl K (ed.). *eLife* 2023;**12**:RP89373.

Armstrong J, Hickey G, Diekhans M *et al.* Progressive Cactus is a multiple-genome aligner for the thousand-genome era. *Nature* 2020;**587**:246–51.

Bothma JP, Garcia HG, Esposito E *et al.* Dynamic regulation of eve stripe 2 expression reveals transcriptional bursts in living Drosophila embryos. *Proceedings of the National Academy of Sciences* 2014;**111**:10598–603.

Cai L, Friedman N, Xie XS. Stochastic protein expression in individual cells at the single molecule level. *Nature* 2006;**440**:358–62.

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

Chen S, Zhang YE, Long M. New genes in Drosophila quickly become essential. *Science* 2010;**330**:1682–5.

Conrad T, Akhtar A. Dosage compensation in Drosophila melanogaster: epigenetic fine-tuning of chromosome-wide transcription. *Nat Rev Genet* 2012;**13**:123–34.

Cridland JM, Majane AC, Sheehy HK *et al.* Polymorphism and Divergence of Novel Gene Expression Patterns in Drosophila melanogaster. *Genetics* 2020;**216**:79–93.

Duboule D. Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Dev Suppl* 1994:135–42.

Forbes Beadle L, Zhou H, Rattray M *et al.* Modulation of transcription burst amplitude underpins dosage compensation in the Drosophila embryo. *Cell Rep* 2023;**42**:112382.

Franke A, Baker BS. The rox1 and rox2 RNAs Are Essential Components of the Compensasome, which Mediates Dosage Compensation in Drosophila. *Molecular Cell* 1999;**4**:117–22.

Fukaya T, Lim B, Levine M. Enhancer Control of Transcriptional Bursting. *Cell* 2016;**166**:358–68.

Golding I, Paulsson J, Zawilski SM *et al.* Real-Time Kinetics of Gene Activity in Individual Bacteria. *Cell* 2005;**123**:1025–36.

Gramates LS, Agapite J, Attrill H *et al.* FlyBase: a guided tour of highlighted features. *Genetics* 2022;**220**:iyac035.

Gubala AM, Schmitz JF, Kearns MJ *et al.* The Goddard and Saturn Genes Are Essential for Drosophila Male Fertility and May Have Arisen De Novo. *Mol Biol Evol* 2017;**34**:1066–82.

Haghverdi L, Lun ATL, Morgan MD *et al.* Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* 2018;**36**:421–7.

Hao Y, Hao S, Andersen-Nissen E *et al.* Integrated analysis of multimodal single-cell data. *Cell* 2021;**184**:3573-3587.e29.

Hense W, Baines JF, Parsch J. X Chromosome Inactivation during Drosophila Spermatogenesis. *PLoS Biol* 2007;**5**:e273.

Hickey G, Paten B, Earl D *et al.* HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* 2013;**29**:1341–2.

Hunt G, Vaid R, Pirogov S *et al.* Tissue-specific RNA Polymerase II promoter-proximal pause release and burst kinetics in a Drosophila embryonic patterning network. *Genome Biology* 2024;**25**:2.

Johnstone CP, Galloway KE. Supercoiling-mediated feedback rapidly couples and tunes transcription. *Cell Rep* 2022;**41**:111492.

Joly D, Korol A, Nevo E. Sperm size evolution in Drosophila: inter- and intraspecific analysis. *Genetica* 2004;**120**:233–44.

Kalinka AT, Varga KM, Gerrard DT *et al.* Gene expression divergence recapitulates the developmental hourglass model. *Nature* 2010;**468**:811–4.

Kemkemer C, Hense W, Parsch J. Fine-Scale Analysis of X Chromosome Inactivation in the Male Germ Line of Drosophila melanogaster. *Molecular Biology and Evolution* 2011;**28**:1561–3.

Khodursky S, Svetec N, Durkin SM *et al.* The evolution of sex-biased gene expression in the Drosophila brain. *Genome Res* 2020, DOI: 10.1101/gr.259069.119.

Korsunsky I, Millard N, Fan J *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods* 2019;**16**:1289–96.

Larschan E, Bishop EP, Kharchenko PV *et al.* X chromosome dosage compensation via enhanced transcriptional elongation in Drosophila. *Nature* 2011;**471**:115–8.

Larsson AJM, Johnsson P, Hagemann-Jensen M *et al.* Genomic encoding of transcriptional burst kinetics. *Nature* 2019;**565**:251–4.

Levine MT, Jones CD, Kern AD *et al.* 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* 2006;**103**:9935–9.

Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 2018;**34**:3094–100.

Lim C, Tarayrah L, Chen X. Transcriptional regulation during Drosophila spermatogenesis. *Spermatogenesis* 2012;**2**:158–66.

Long M, VanKuren NW, Chen S *et al.* New gene evolution: little did we know. *Annual Review of Genetics* 2013;**47**:307–33.

Mahadevaraju S, Fear JM, Akeju M *et al.* Dynamic sex chromosome expression in Drosophila male germ cells. *Nat Commun* 2021;**12**:892.

McAdams HH, Arkin A. Stochastic mechanisms in gene expression. *PNAS* 1997;**94**:814–9.

Meiklejohn CD, Landeen EL, Cook JM *et al.* Sex Chromosome-Specific Regulation in the Drosophila Male Germline But Little Evidence for Chromosomal Dosage Compensation or Meiotic Inactivation. *PLOS Biology* 2011;**9**:e1001126.

Mi H, Muruganujan A, Casagrande JT *et al.* Large-scale gene function analysis with the PANTHER classification system. *Nat Protoc* 2013;**8**:1551–66.

Nuzhdin SV, Wayne ML, Harmon KL *et al.* Common Pattern of Evolution of Gene Expression Level and Protein Sequence in Drosophila. *Molecular Biology and Evolution* 2004;**21**:1308–17.

Qiu X, Mao Q, Tang Y *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* 2017;**14**:979–82.

Raff RA. *The Shape of Life: Genes, Development, and the Evolution of Animal Form*. University of Chicago Press, 2012.

Raj A, van Oudenaarden A. Nature, Nurture, or Chance: Stochastic Gene Expression and Its Consequences. *Cell* 2008;**135**:216–26.

Raj A, Peskin CS, Tranchina D *et al.* Stochastic mRNA Synthesis in Mammalian Cells. *PLOS Biology* 2006;**4**:e309.

Rastelli L, Kuroda MI. An analysis of *maleless* and histone H4 acetylation in *Drosophila melanogaster* spermatogenesis. *Mechanisms of Development* 1998;**71**:107–17.

Raz AA, Vida GS, Stern SR *et al.* Emergent dynamics of adult stem cell lineages from single nucleus and single cell RNA-Seq of Drosophila testes. Buszczak M, Banerjee U, Buszczak M, et al. (eds.). *eLife* 2023;**12**:e82201.

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

Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. *Bioinformatics* 2021;**37**:1639–43.

Snook RR, Markow TA, Karr TL. Functional nonequivalence of sperm in Drosophila pseudoobscura. *Proc Natl Acad Sci U S A* 1994;**91**:11222–6.

Suter DM, Molina N, Gatfield D *et al.* Mammalian Genes Are Transcribed with Widely Different Bursting Kinetics. *Science* 2011;**332**:472–4.

Suvorov A, Kim BY, Wang J *et al.* Widespread introgression across a phylogeny of 155 *Drosophila* genomes. *Current Biology* 2022;**32**:111-123.e5.

Tarashansky AJ, Musser JM, Khariton M *et al.* Mapping single-cell atlases throughout Metazoa unravels cell type evolution. Shalek AK, Barkai N (eds.). *eLife* 2021;**10**:e66747.

Thomas PD, Ebert D, Muruganujan A *et al.* PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Science* 2022;**31**:8–22.

Valsecchi CIK, Basilicata MF, Georgiev P *et al.* RNA nucleation by MSL2 induces selective X chromosome compartmentalization. *Nature* 2021;**589**:137–42.

Vibranovski MD. Meiotic Sex Chromosome Inactivation in Drosophila. *J Genomics* 2014;**2**:104–17.

Walters MC, Fiering S, Eidemiller J *et al.* Enhancers increase the probability but not the level of gene expression. *Proc Natl Acad Sci U S A* 1995;**92**:7125–9.

Wei KH-C, Chatla K, Bachtrog D. Single cell RNA-seq in Drosophila testis reveals evolutionary trajectory of sex chromosome regulation. 2022:2022.12.07.519494.

Weisman CM, Murray AW, Eddy SR. Mixing genome annotation methods in a comparative analysis inflates the apparent number of lineage-specific genes. *Current Biology* 2022;**32**:2632-2639.e2.

Witt E, Benjamin S, Svetec N *et al.* Testis single-cell RNA-seq reveals the dynamics of de novo gene transcription and germline mutational bias in Drosophila. Landry CR, Wittkopp PJ, White-Cooper H (eds.). *eLife* 2019;**8**:e47138.

Witt E, Langer CB, Svetec N *et al.* Transcriptional and mutational signatures of the Drosophila ageing germline. *Nat Ecol Evol* 2023;**7**:440–9.

Witt E, Svetec N, Benjamin S *et al.* Transcription Factors Drive Opposite Relationships between Gene Age and Tissue Specificity in Male and Female Drosophila Gonads. *Molecular Biology and Evolution* 2021;**38**:2104–15.

Yamashita YM, Fuller MT, Jones DL. Signaling in stem cell niches: lessons from the Drosophila germline. *Journal of Cell Science* 2005;**118**:665–72.

Zhang X, Kaplow IM, Wirthlin M *et al.* HALPER facilitates the identification of regulatory element orthologs across species. *Bioinformatics* 2020;**36**:4339–40.

Zhao L, Saelao P, Jones CD *et al.* Origin and Spread of de Novo Genes in Drosophila melanogaster Populations. *Science* 2014;**343**:769–72.

Zoller B, Little SC, Gregor T. Diverse Spatial Expression Patterns Emerge from Unified Kinetics of Transcriptional Bursting. *Cell* 2018;**175**:835-847.e25.

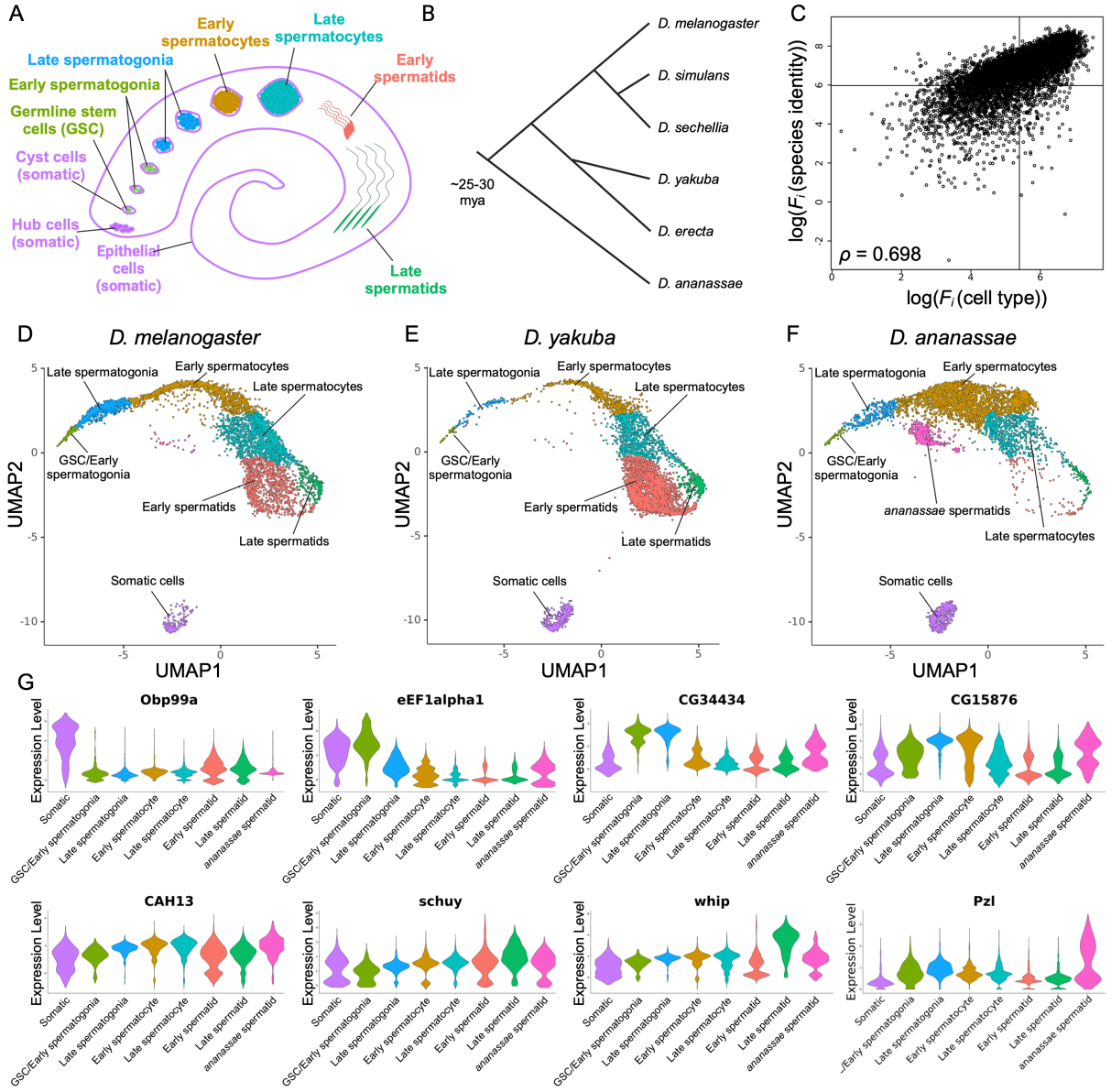

Figure 1: Identification of 198-gene list allows for cell type classification across species in a rapidly diverging process. **(A)** Cartoon figure of spermatogenesis progression, beginning with germline stem cells and finishing in late spermatids. **(B)** Phylogeny of species used in this study with estimated divergence times. **(C)** The joint distribution of ANOVA log(F-statistic) values derived from single cell RNA-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.

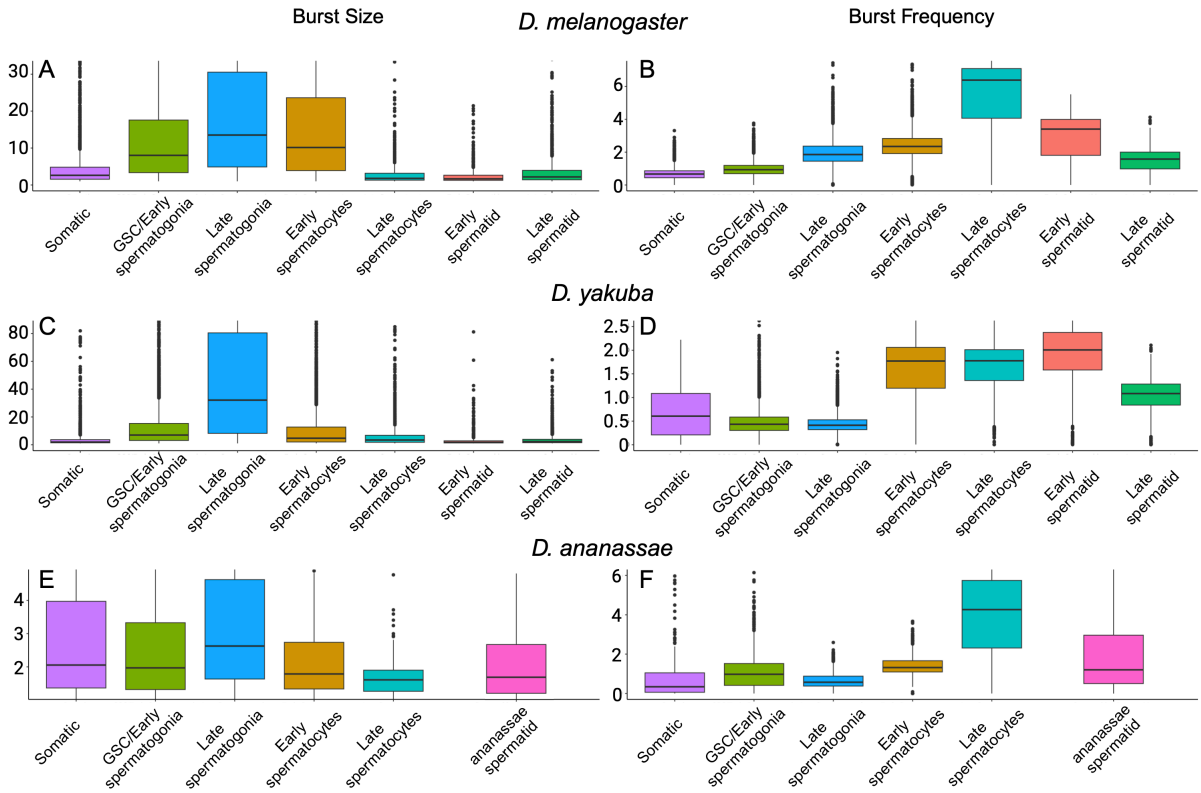

Figure 2: Pre- and post-meiotic cell types in spermatogenesis are controlled by opposing mechanisms. Burst size and burst frequency distributions for **(A, B)** *D. melanogaster*, **(C, D)** *D. yakuba*, and **(E, F)** *D. ananassae* reveal that transcription in the earlier states of spermatogenesis 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.

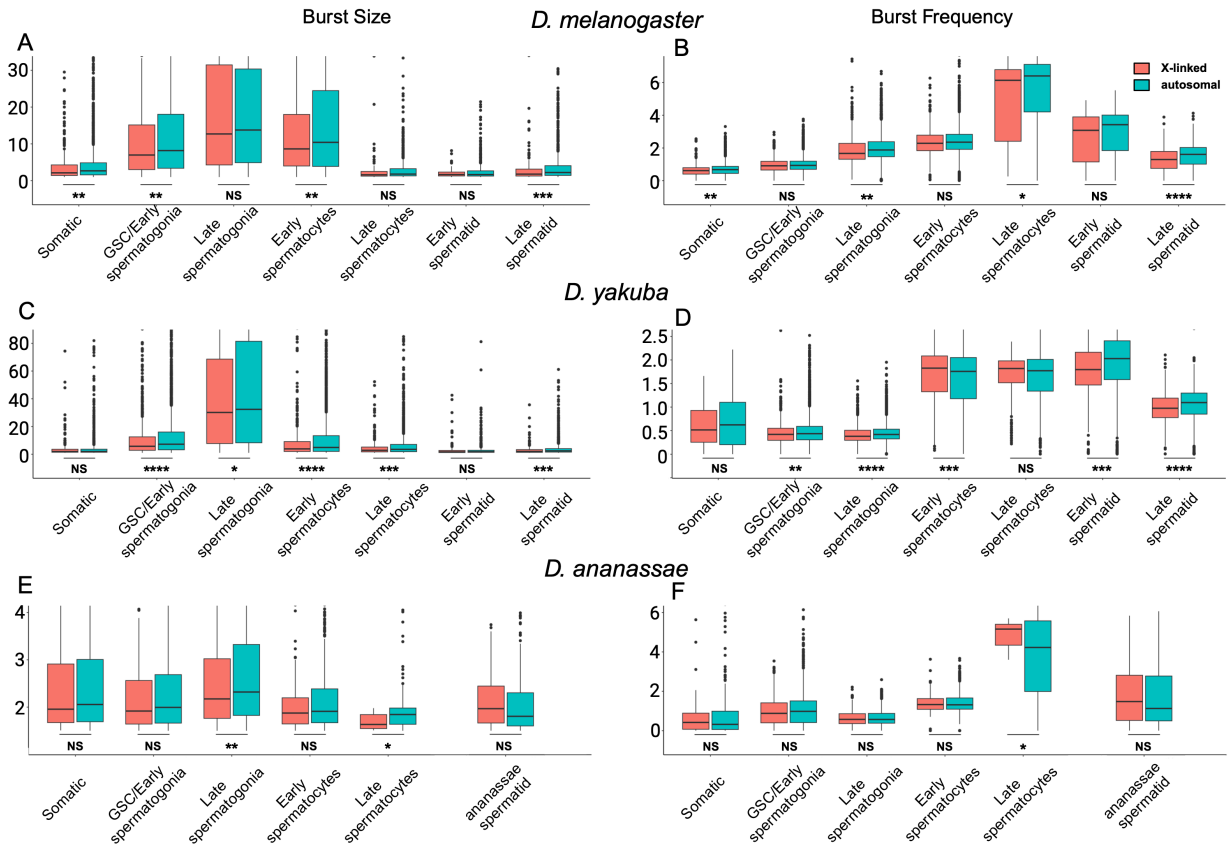

Figure 3: X-linked genes show lower bursting kinetics than autosomal genes prior to spermatogenesis. Burst size and burst frequency distributions for **(A, B)** *D. melanogaster*, **(C, D)** *D. yakuba*, and **(E, F)** *D. ananassae* reveal a conserved pattern where X-linked genes show significantly smaller average burst sizes and lower bursting frequencies than autosomal genes across cell types (* = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, **** = $p < 0.0001$). 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.

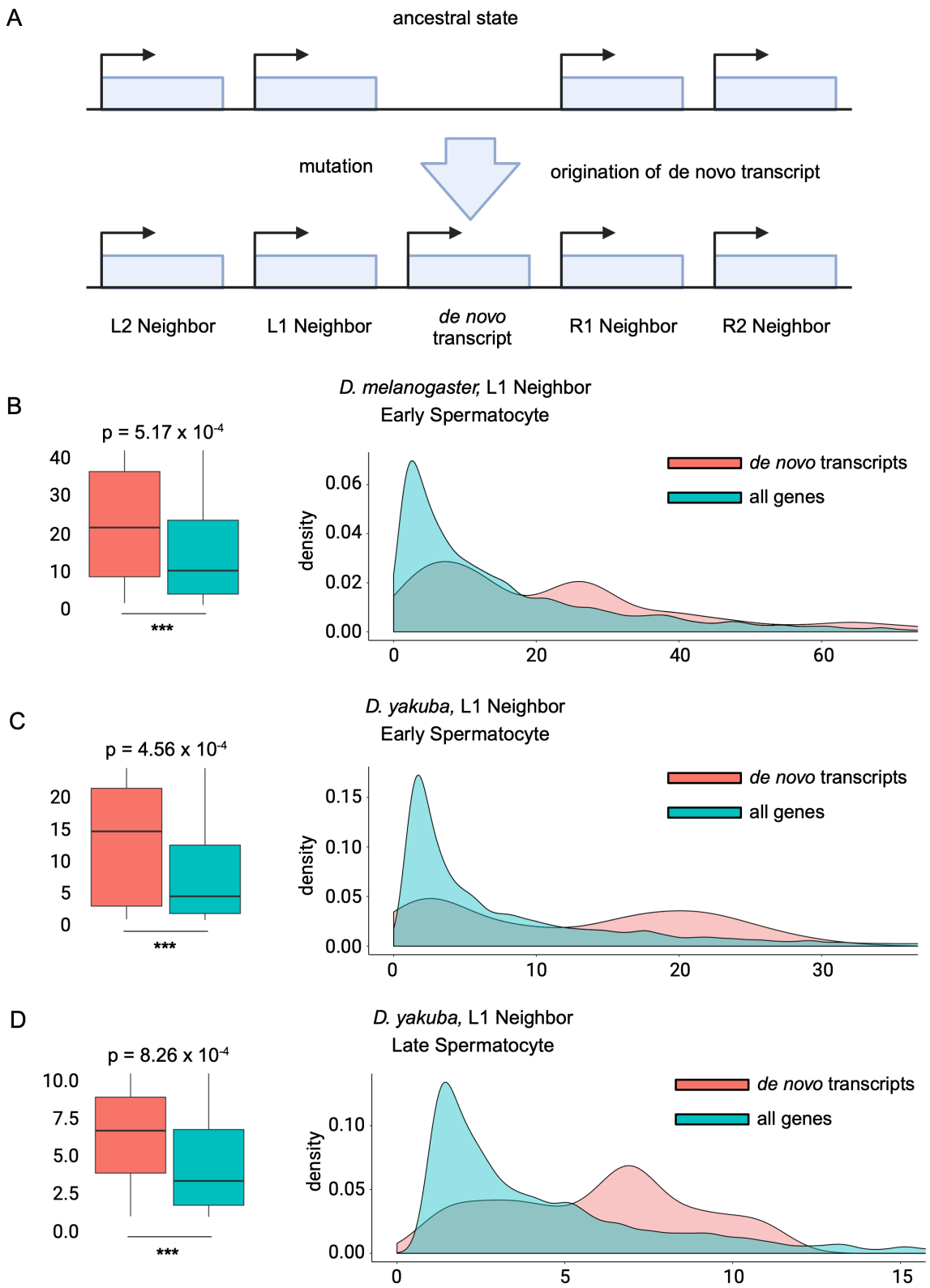
A
ancestral state
mutation
origination of de novo transcript
L2 Neighbor
L1 Neighbor
de novo transcript
R1 Neighbor
R2 Neighbor
B
p = 5.17 x 10-4
***
D. melanogaster, L1 Neighbor
Early Spermatocyte
de novo transcripts
all genes
density
C
p = 4.56 x 10-4
***
D. yakuba, L1 Neighbor
Early Spermatocyte
de novo transcripts
all genes
density
D
p = 8.26 x 10-4
***
D. yakuba, L1 Neighbor
Late Spermatocyte
de novo transcripts
all genes
density

Figure 4: 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 posits 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. **(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 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 (Supplementary Figure S6).