Transcription Factors Drive Opposite Relationships between Gene Age and Tissue Specificity in Male and Female *Drosophila* Gonads

Evan Witt, Nicolas Svetec, Sigi Benjamin, and Li Zhao*

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

***Corresponding author**: E-mail: lzhao@rockefeller.edu.

Associate editor: John Parsch

Abstract

Evolutionarily young genes are usually preferentially expressed in the testis across species. Although it is known that older genes are generally more broadly expressed than younger genes, the properties that shaped this pattern are unknown. Older genes may gain expression across other tissues uniformly, or faster in certain tissues than others. Using *Drosophila* gene expression data, we confirmed previous findings that younger genes are disproportionately testis biased and older genes are disproportionately ovary biased. We found that the relationship between gene age and expression is stronger in the ovary than any other tissue and weakest in testis. We performed ATAC-seq on *Drosophila* testis and found that although genes of all ages are more likely to have open promoter chromatin in testis than in ovary, promoter chromatin alone does not explain the ovary bias of older genes. Instead, we found that upstream transcription factor (TF) expression is highly predictive of gene expression in ovary but not in testis. In the ovary, TF expression is more predictive of gene expression than open promoter chromatin, whereas testis gene expression is similarly influenced by both TF expression and open promoter chromatin. We propose that the testis is uniquely able to express younger genes controlled by relatively few TFs, whereas older genes with more TF partners are broadly expressed with peak expression most likely in the ovary. The testis allows widespread baseline expression that is relatively unresponsive to regulatory changes, whereas the ovary transcriptome is more responsive to *trans*-regulation and has a higher ceiling for gene expression.

***Key words:* testis, ovary, expression complexity, ATAC-seq, *Drosophila*, transcription factor expression.**

Introduction

For eons, genes have continuously arisen by a multitude of ways, from duplication and divergence to de novo origination from nongenic DNA (Ohno 1970; Long et al. 2003; Begun et al. 2006; Zhou et al. 2008; Tautz and Domazet-Lošo 2011; Zhao et al. 2014). Gene birth and death is a continuous and dynamic process in evolution, culled by natural selection or genetic drift (Kaessmann 2010; Palmieri et al. 2014). A large portion of young genes segregate within or recently reach fixation in populations, and most young genes are expressed specifically in the testis (Levine et al. 2006; Zhao et al. 2014), similar to duplicated genes (Long et al. 2013). The phrase "out of the testis" was originally used to describe young retroposed genes (Vinckenbosch et al. 2006), which gained expression by exploiting *cis*-regulatory machinery of nearby genes. Testis bias has since been observed in young X-linked duplicate genes, leading researchers to propose that young genes escape meiotic sex chromosome inactivation due to immature *cis*-regulatory machinery (Zhang, Vibranovski, Landback, et al. 2010). Testis expresses more genes in general than any other tissue (Soumillon et al. 2013), and studies from many taxa support that a large proportion of young genes then to show testis-biased or testis-specific expression and function (see review in Long et al. 2013).

The testis-biased expression of young genes has many possible explanations. Besides the obvious hypothesis that genes expressed in reproductive tissues may directly influence reproductive success and fitness (Zhang et al. 2004; Begun et al. 2006), many propose that the testis has a permissive chromatin environment facilitating the transcriptional birth of genetic novelties (Kaessmann 2010; Soumillon et al. 2013). Indeed, most genes are at least somewhat expressed in the testis (Soumillon et al. 2013; Witt et al. 2019). It has long been proposed that upregulation of universal transcriptional machinery facilitates such widespread transcription (Schmidt 1996). Such broad transcription may be a form of genomic surveillance, meant to detect and repair mutations via transcription-coupled repair or other mechanisms (Grive et al. 2019; Xia et al. 2020). It has also been proposed that permissive testis transcription is also due to reduced mRNA degradation of testis-specific genes (Mayr 2016). Young genes may also have low levels of "active" epigenetic markers across tissues, despite high expression in testis (Zhang and Zhou 2019). Results from Zhang and Zhou (2019) suggest that young genes have similar epigenetic profiles across tissues, yet show testis-biased expression, whereas older genes show consistently higher levels of "active" epigenetic marks. Their results indicate that the "out of the testis" pattern for the emergence of young genes may not be driven by specific epigenetic marks, but rather by a context-dependent *trans*-regulatory environment between tissues (Ding et al. 2010). Alternatively, recruitment of nearby testis-biased *cis*-regulatory elements by young genes may also be responsible for many testis-biased new genes (Majic and Payne 2020).

Although it is known that young genes are often testis specific, and that older genes are more broadly expressed than young genes (Zhou et al. 2008; Kondo et al. 2017), it is unknown how this relationship works. When genes age, do older genes lose expression in testis and retain relatively constant expression in other tissues? Or do older genes maintain relatively constant expression in testis, and gain expression in other tissues? If so, are all non-testis tissues equally conducive to old genes, or do the genomic characteristics of older genes produce higher expression in certain tissues? Once out of the testis, is any tissue the next hot target of tissue-biased expression when the genes expand their functions in other tissues?

One clue is that older duplicated genes are more likely to be retained if they are ovary biased (Assis 2019). This might imply the specific importance of older genes to ovary expression and function. To this effect, researchers have identified several modules of highly conserved, older genes with heightened importance in human ovarian function (Zhang et al. 2019). To see whether the *Drosophila* ovary drives the shift away from testis bias in older genes, we analyzed a database of RNA-seq data from FlyAtlas2 (Leader et al. 2018) to characterize tissue bias for genes of all ages in every tissue. We found that ovary has the largest relationship between gene age and expression, explaining why the oldest genes are often ovary biased. Conversely, testis shows a weaker relationship between gene age and gene expression than any other tissue.

To explain this trend, we examined the tissue-specific activity of the transcription factor (TF) regulators of every gene in the DroID database (Murali et al. 2011). We found that ovary-biased genes tend to have higher upstream TF expression than testis-biased genes of all age groups, yet young genes, with fewer TF partners, tend to be testis-expressed and old genes, with more TF partners, tend to be ovary biased. We found evidence that testis allows higher transcription than the ovary for genes with low TF expression. Conversely, genes with high TF expression have higher expression in the ovary than the testis. Additional upstream TF expression appears to confer diminishing returns on expression in testis but greatly benefits ovary expression, explaining why older genes with more TF partners tend to be ovary biased.

After establishing the different relationships between *trans*-regulation and gene expression in testis and ovary, we performed ATAC-seq to assess whether open promoter chromatin is equally predictive of expression in the two tissues. All age groups of genes are more likely to have open promoter chromatin in testis than ovary, indicating that open chromatin by itself is insufficient to explain age-related expression bias. In ovary, we found that high upstream TF expression is much more predictive of gene expression than the presence of open promoter chromatin; whereas in testis, high TF expression and open promoter chromatin are similarly predictive of gene expression. This indicates that gene expression in ovary is much more linked to *trans*-regulatory factors than testis expression. Taken together with our observation that young genes are less likely to be bound by annotated TFs than older genes, the opposite trends of gene age and tissue bias in testis and ovary make biological sense. We published a web app to allow users to interactively explore our tissue specificity data for any set of genes without coding experience necessary: https://zhao.labapps.rockefeller.edu/tissue-specificity/ (last accessed January 22, 2021).

Results

Testis and Ovary Show an Opposite Relationship between Gene Age and Tissue Bias

Using gene ages divided into Drosophilid (youngest), pre-Drosophilid (middle-aged), and pre-Bilateria (oldest), and tissue RNA-seq data from FlyAtlas2, we find results consistent with earlier work showing that younger genes are more tissue specific than older genes (fig. 1A). We plotted the proportion of genes from each age group with a maximum expression in testis, ovary, and male and female carcasses with the reproductive tracts removed. A plurality of young genes are testis biased, but the abundance of testis-biased genes declines for older genes (fig. 1B). Surprisingly, we found the opposite trend for ovary: Older genes are very likely to have maximum expression in ovary, but almost no younger genes are ovary biased. No other tissues displayed a relationship of this magnitude (supplementary fig. 1, Supplementary Material online), indicating that the two tissues that contribute most to gene age-related expression patterns are the male and female reproductive tissues. Whereas young genes are often testis biased and highly tissue specific, old genes are broadly expressed with peak expression in ovary.

Although a plurality of old genes are ovary biased, this is not due to an increased likelihood of expression for old genes in ovary. Young genes are most commonly expressed with FPKM > 2 in testis (65%) and least commonly expressed in ovary (13%), whereas testis, ovary and somatic tissues express a similar proportion of old genes (all between 73% and 85%; fig. 1C, supplementary fig. 2, Supplementary Material online). Therefore, the age-related decline in testis bias is not due to an absence of old gene expression in testis. The proportion of genes expressed between age groups varies the least in testis, and the most in ovary, indicating that ovary may have a disproportionately large relationship between gene age and expression. We confirmed that young duplicate genes were not confounding these results by repeating the analysis from figure 1 with *melanogaster*-specific genes removed (supplementary fig. 3, Supplementary Material online). We also confirmed these results with an alternate set of gene age assignments (supplementary fig. 4, Supplementary Material online).

Testis Shows a Weak, and Ovary Shows a Strong Relationship between Gene Age and Expression

We wanted to further unpack how gene expression correlates with gene age across tissues to understand our observed

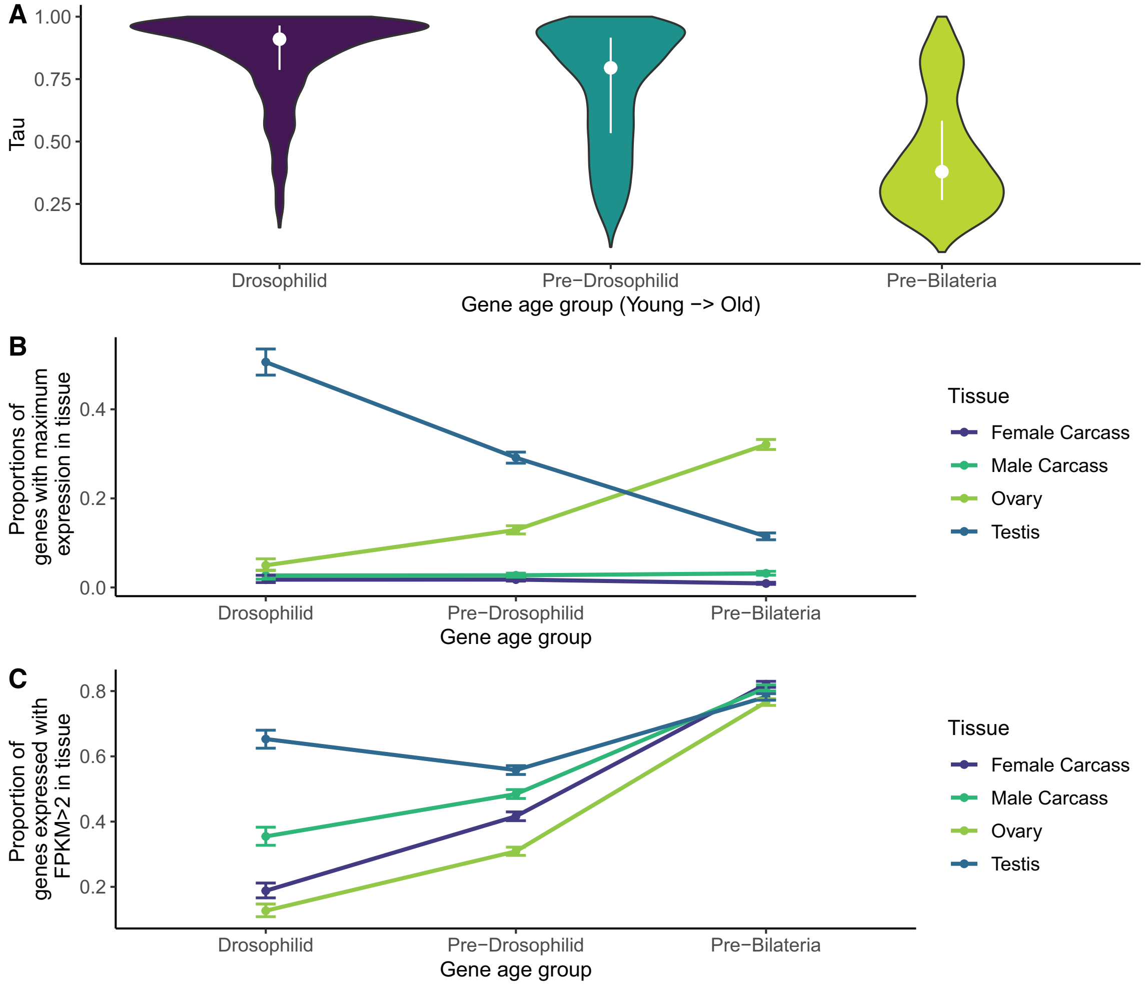

Fig. 1. Young genes are testis specific, old genes are broadly expressed and often ovary biased. (*A*) Average Tau values among genes of each age group, dots represent medians and vertical lines are interquartile ranges. Young genes are more tissue specific (higher Tau) than older genes. (*B*) For four tissues, the proportion of genes of each age group with maximum expression in that tissue. Younger genes usually have the highest expression in testis, but the proportion of testis-biased genes declines with gene age. Ovary-biased young genes are rare, but old genes are more often biased toward ovary than any other tissue. Error bars are 95% confidence intervals for proportion test. (*C*) For four tissues, proportion of genes of each age group with FPKM >2 in that tissue. In testis, ovary, and carcass, old genes are more likely to be expressed than young genes, but this disparity is smallest in testis and largest in ovary. By this measure, old genes are no longer biased in testis. Ovary bias of older genes is not explained by the relative proportion of genes expressed between tissues.

patterns of testis bias and ovary bias. For each tissue, we plotted gene expression (Log 2(FPKM + 1)) from FlyAtlas2 conditioning by gene age. In every tissue, expression of old genes was higher for pre-Bilateria genes than for Drosophilid genes as measured with a pairwise Wilcoxon test (fig. 2A). In every tissue except testis, Drosophilid genes were less expressed on average than pre-Drosophilid genes. In testis, these two groups were statistically similar. This may be because in testis, unlike other tissues, a similar proportion of genes are expressed for each age group (fig. 1C). A qualitative comparison shows that expression of the three age groups is least different in testis (median FPKM 8.53 [Drosophilid], 3.19 [pre-Drosophilid], 8.24 [pre-Bilateria]), and most dramatically different in ovary (median FPKM 0.071, 0.25, 21.10, respectively) (fig. 2A).

To quantitatively compare tissue-specific gene expression as a function of age group, we performed a one-way analysis of variance (ANOVA) on each tissue and age group from figure 2A. The ANOVA *F*-statistic is the ratio of between-group variation to intragroup variation. For similar groups, the *F*-statistic is close to 1. The ANOVA *F*-statistic is highest in ovary, meaning that the age groups are more variable in this tissue than any other. In testis, the *F*-statistic is lower, meaning that gene expression varies less between age groups. Young genes have a relatively similar expression in testis across all age groups, in contrast to other tissues where gene expression is highly stratified across age groups, with young genes the least and old genes the most expressed.

For each tissue, we also calculated the summed pairwise mean differences between every group. This measure is the absolute value of the difference between the mean of each age group within a tissue, summed for each pair of groups (Drosophilid vs. pre-Drosophilid, pre-Drosophilid vs. Bilateria, Drosophilid vs. Bilateria). By this measure,

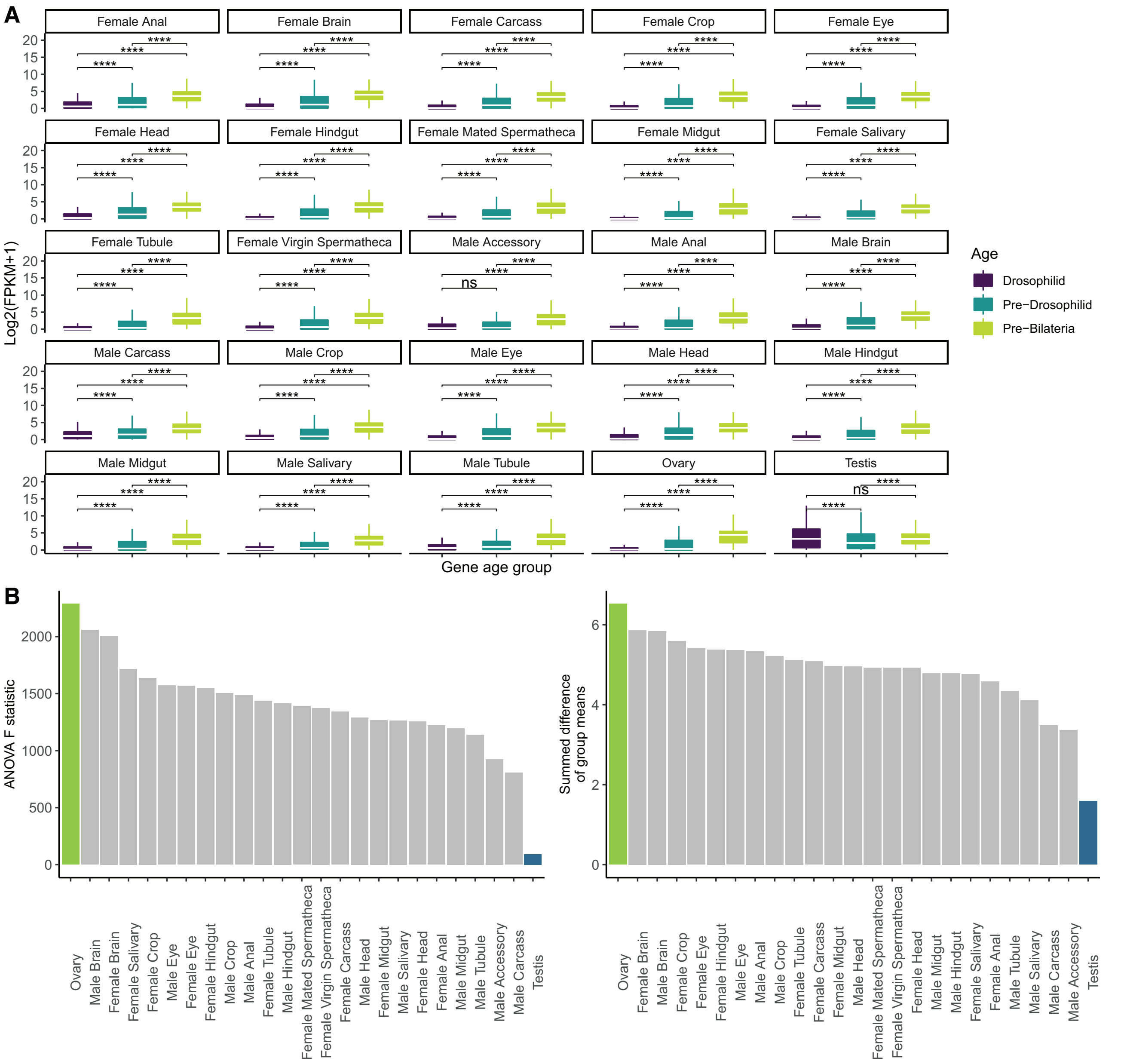

Fig. 2. Tissue-specific trends in the gene age/expression relationship. (*A*) The log-scaled expression of every gene in that tissue versus gene age for every adult tissue (see Materials and Methods). In almost every tissue scaled expression is very low for young genes, and high for older genes. The testis is an outlier, with a statistically similar expression between Drosophilid and Pre-Drosophilid genes. Asterisks represent *P* values are adjusted with Bonferroni's correction (****$P < 0.00005$). Raw and Bonferroni-adjusted *P* values are in supplementary table 1, Supplementary Material online. (*B*) Rankings of ANOVA statistics for all tissues. We performed an ANOVA on each of the panels from part A, comparing, for each tissue, the ratio of intergroup variation to between-group variation (*F*-statistic). By this measure, ovary has the largest relationship between gene age and expression (because old genes are often ovary biased), and testis has the smallest (because old and young genes are similarly expressed in the testis). We also took the mean difference between groups and summed their absolute values for each tissue. Testis has the smallest mean expression difference between age groups, and ovary has the largest. This conclusion held when we repeated the analysis using an alternate set of gene age assignments (supplementary fig. 6, Supplementary Material online).

mean testis expression is the least different between gene age groups and ovary expression varies the most of any tissue (fig. 2*B*). The results in figure 2*B* hold if *melanogaster*-specific genes are removed (supplementary fig. 5, Supplementary Material online), or with an alternate method of gene age assignments (supplementary fig. 6, Supplementary Material online).

Testis Expression Requires Lower TF Activity than Ovary Expression

We hypothesized that TFs may play a role in the discrepancy between age/expression relationships between the testis and ovary. We designed a proxy measure of TF network activity for every gene in every tissue. For every gene with bound TFs listed by DroID (Murali et al. 2011), we defined the summed scaled expression of the upstream TFs of a gene in a tissue as "TF expression." Higher TF expression in a tissue indicates that gene's TF partners are more transcriptionally active in that tissue. This metric is based on data from ChIP-seq and ChIP-chip experiments for individual TFs and only considers whether a TF binds to a given gene's promoter. Although such a method does not reveal whether a TF–gene relationship is one of activation or repression, it is unbiased with regard to gene age because the whole-genome binding profile of a TF is agnostic to the degree of study a particular gene has received (as young genes are often less studied than older genes with mammalian homologs).

The purpose of our TF expression metric is not to infer gene expression (for which RNA-seq is much better suited), but rather to assess the relative dynamics between gene expression and *trans*-regulation across tissues. For this purpose, the metric performs consistently well across tissues even though some TFs are repressive in nature. For more details about TF expression, see Materials and Methods.

We compared the TF expression of young, middle-aged, and older genes between the testis and ovary. We thought that as young genes are more specifically expressed in testis, young genes would have higher upstream TF expression in testis than in ovary. We found that no age group of genes shows higher TF expression in testis than ovary (fig. 3*A*). The testis specificity of young genes must be due to factors other than increased TF expression in testis. Exploring further, we found that young genes have fewer identified TF-gene interactions than middle-aged genes, which in turn have fewer TF binding partners than old genes (fig. 3*B*). We confirmed these results using an alternate list of gene ages in supplementary figure 7, Supplementary Material online.

We then sought to correlate expression with TF activity between testis and ovary and found that genes with low TF expression are much more active in testis than in ovary. Conversely, genes with high TF expression are often more active in the ovary than in testis (fig. 3*C*). It appears that testis expression requires fewer TFs than ovary expression, explaining why young genes, with fewer TFs, would have testis-biased expression. Having many TF partners, a property of older genes, appears to boost expression in ovary more than in testis. To confirm that this property was not sex specific we compared TF expression and gene expression in the male and female brain, two sexually dimorphic tissues, and observed no major differences (fig. 3*D*). Additionally, we made this comparison across all tissues in FlyAtlas2 (supplementary fig. 8, Supplementary Material online) and found that gene expression is least correlated to TF expression in testis (Pearson's $r = 0.22$) and most responsive in ovary (Pearson's $r = 0.67$).

Testis Promoter Chromatin Is Broadly Open across All Gene Ages

To see whether promoter chromatin environment explains TF expression differences in testis and ovary, we performed ATAC-seq on *Drosophila* testis and obtained ATAC-seq data sets for *Drosophila* ovarian somatic cells (Iwasaki et al. 2016) and S2 cells (Vaid et al. 2020). We annotated peaks in the promoters of genes from each age group and compared the proportion of genes with detectable high-quality peaks in each tissue (fig. 4*A*). In every tissue, young genes were the least likely to have detectable chromatin accessibility in their promoters, and old genes were the most likely to have detectable peaks. Every age group of genes was more likely to have peaks in testis, and least likely to have peaks in ovary, indicating that chromatin at the promoter is more broadly open in the testis. In addition, a majority of genes from each age group exhibited more frequent detectable open promoter chromatin in testis. In ovary, by contrast, pre-Bilateria genes are the only age group of which a majority of genes (68%) have detectable ATAC-seq peaks. Every other age group of genes is less likely to contain detectable promoter ATAC-seq peaks, especially young genes, of which only 26% have open chromatin in ovary, compared with 56% of young genes in testis. Our observation that every gene-age group is more likely to have testis peaks than ovary peaks indicates that open chromatin does not underlie the ovary bias of older genes.

The presence of an ATAC-seq peak generally corresponds to increased gene expression in analogous tissues (fig. 4*B–D*). Similarly, genes with an ATAC-seq peak in a tissue have heightened activity of their partner TFs compared with genes with no peak in a tissue (fig. 4*E* and *F*). This indicates that TF expression and promoter chromatin state are useful proxies of a gene's network activity (Sigalova et al. 2020).

The low proportion of young genes with ovary ATAC-seq peaks does not entirely explain the paucity of young ovary-biased genes. In the ovary, we found that 13% of young genes are expressed whereas 26% of them have open promoter chromatin. We, therefore, sought next to separate the relative influences of TF expression and promoter chromatin for testis and ovary expression.

High Upstream TF Expression Boosts Gene Expression in Ovary More than in Testis

We quantified expression for genes with and without detectable ATAC-seq peaks, conditioning on whether they had high or low TF expression in the tissue (fig. 5). Many genes in testis have a surprisingly high expression (median FPKM 1.04) without nearby detectable ATAC-seq peaks or high TF expression, indicating that baseline transcription is higher in testis than in ovary (median FPKM 0.13). Without the aid of many TF partners or open promoter chromatin detectable by ATAC-seq, plenty of genes have surprisingly high expression in testis but not ovary. In both testis and ovary, the presence of detectable ATAC-seq peaks or high TF expression (greater than the tissue median) is associated with an expression boost. In testis, however, these fold differences in median expression are smaller than in ovary (table 1). Furthermore, in ovary, high TF expression boosts expression 169.23-fold in genes without a detectable ATAC-seq peak. For genes with a detectable ATAC-seq peak in ovary, high TF expression is associated with a further 15% boost in expression. Ovary expression is 24.18-fold higher for genes with high TF expression but no detectable ATAC-seq peaks compared with genes with open chromatin but low TF expression, indicating that high TF

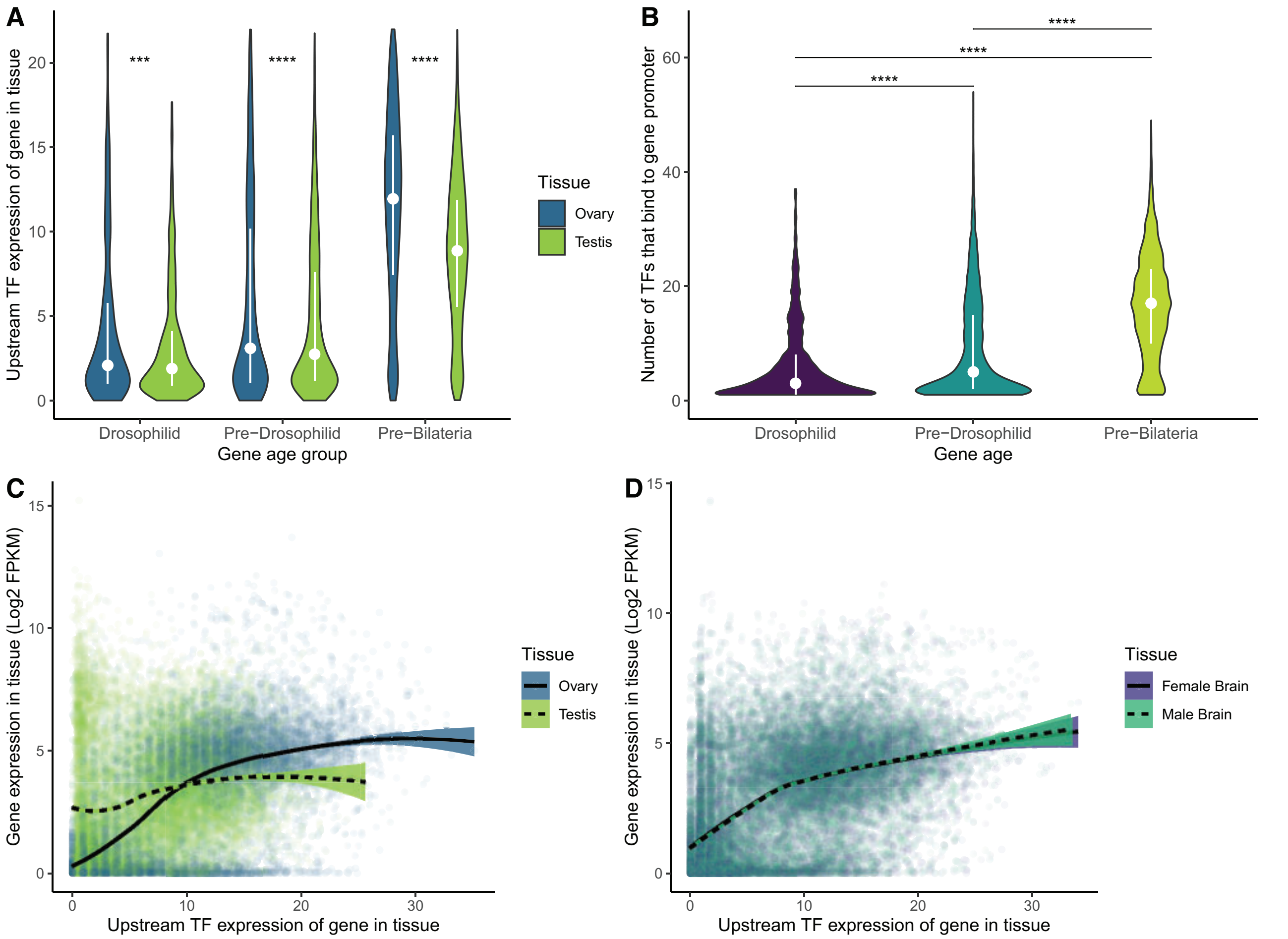

Fig. 3. Transcription factor expression explains gene age/expression trends in testis and ovary. (*A*) Upstream TF expression in testis and ovary for genes with different ages, using DroID's curated database of TF-binding profiles from the modENCODE project. For every gene with a confirmed TF–promoter interaction, we calculated TF expression in testis and ovary by scaling the expression for each TF from 0 to 1, and summing the scaled expression of every TF that binds to the promoter of a given gene. Older genes have much higher TF expression than younger genes in both tissues, and no age group of genes shows elevated TF expression in testis compared with ovary. White dots are medians and lines are interquartile ranges. Asterisks represent *P* values adjusted with Bonferroni's correction (***$P < 0.0005$, ****$P < 0.00005$). (*B*) In DroID data, the promoters of older genes have been shown able to be bound by more TFs than younger genes. (*C*) Log-scaled gene expression versus upstream TF expression in gonads. Genes require less upstream TF expression for expression in testis than in ovary. Genes have fairly high testis expression even without much TF expression in testis, but genes with low ovary TF expression are relatively lowly expressed in ovary. This indicates that genes require less TF expression for testis expression than ovary expression. In ovary, higher TF expression corresponds to higher expression, moreso than testis, where adding TF expression makes relatively little difference in testis gene expression. (*D*) Log-scaled gene expression versus upstream TF expression in brains. These sexually dimorphic tissues show no difference in their relationships between TF expression and gene expression. In these tissues, low TF expression yields low expression, and high TF expression yields high expression, much like the ovary and much unlike the testis. Lines are smoothed loess regressions with 95% confidence intervals. Other tissues are shown in supplementary figure 3, Supplementary Material online.

expression is more predictive of expression than chromatin environment in ovary.

In testis, both the presence of open chromatin and high TF expression are associated with an expression boost, but every pairwise comparison shows a smaller magnitude difference than in ovary, indicating that *trans*-regulation influences expression in ovary more than in testis. Although the presence of ATAC-seq peaks correlates with gene expression, TF expression is generally both necessary and sufficient for gene expression in ovary. Most genes in testis have a low but genuine expression ($\text{FPKM} > 1$) without nearby detectable ATAC-seq peaks or high TF expression, indicating that leaky transcription may be commonplace in testis. The same category of genes in ovary has a median FPKM of 0.13, negligible by comparison.

Discussion

Our results shape the contours of a model where gene age correlates with tissue-specific determinants of gene expression patterns. Genes are typically born under simpler regulatory machinery (*cis*-regulation with fewer TF binding sites; Zhao et al. 2014), sufficient to drive expression in the testis but not other tissues. As a gene ages, it will likely recruit more *trans*-acting TF partners, strengthen existing *cis*-acting TF binding sites (Tuğrul et al. 2015), or gain novel binding sites

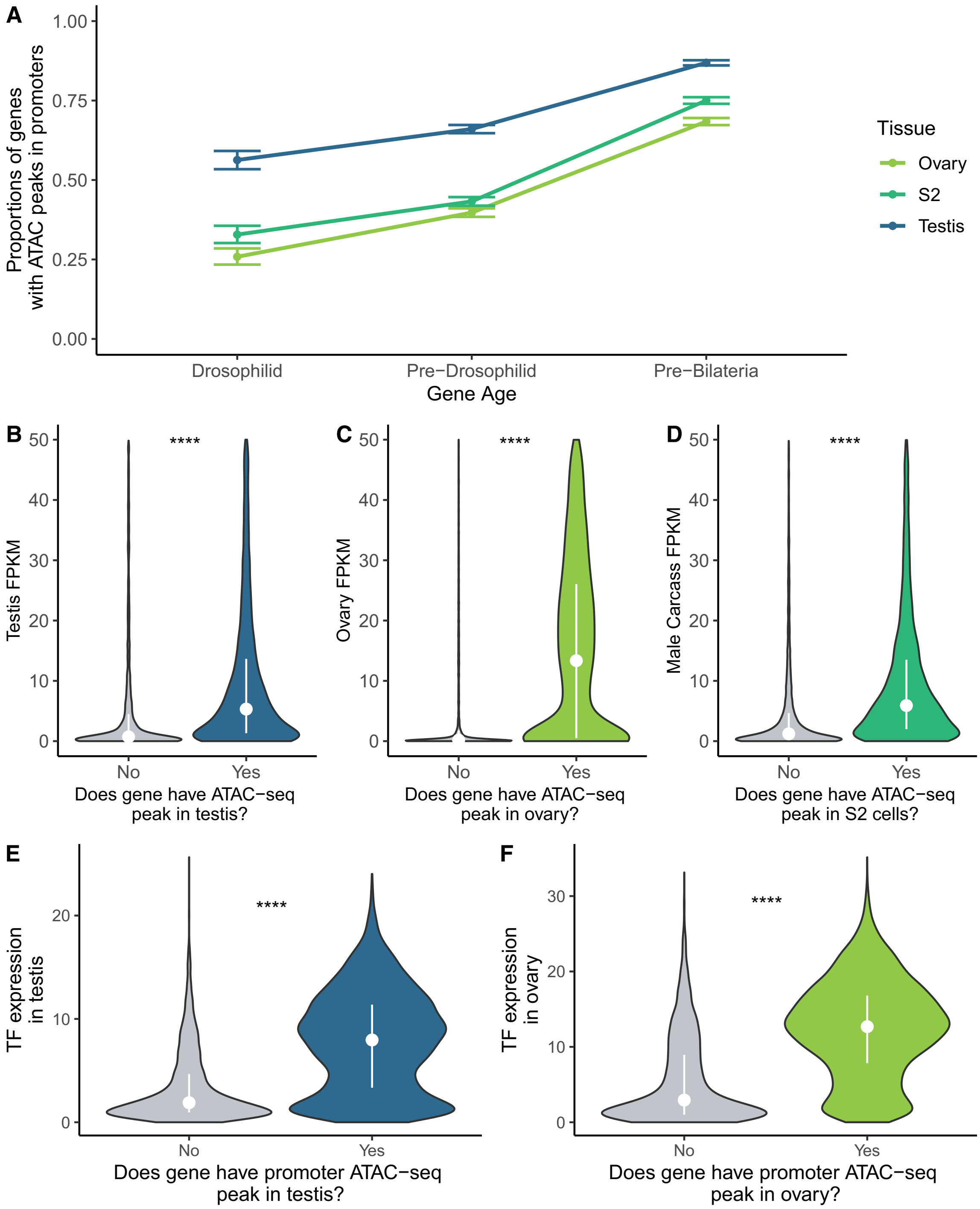

Fig. 4. ATAC-seq peaks show an age-related trend in multiple tissues. (*A*) The relative proportions of genes with a detectable ATAC-seq peak in their promoters, for three gene age groups and three data sets. For each data set, young genes were the least likely to have open chromatin in their promoters. Testis is unique among these data sets because a majority of genes of each age group have open promoter chromatin. (*B*) FPKM for genes with and without a detected promoter ATAC-seq peak in testis. Genes with open promoter chromatin in testis have generally higher expression in FlyAtlas2 data. Dots are medians, and the white line is the interquartile range. (*C*) Genes with open promoter chromatin in ovary have higher FlyAtlas2 expression, and the FPKM difference between genes with and without peaks is much larger than the other two tissues. (*D*) Genes with promoter peaks in S2 cells generally have higher expression in male carcass, the most analogous FlyAtlas2 tissue to this cell line. (*E*) TF expression for genes with and without detectable ATAC-seq peaks. Genes with ATAC-seq promoter peaks tend to have higher TF expression in testis, (*F*) as well as ovary. **** represents adjusted *P* values <0.00005.

(Trizzino et al. 2017; Levran et al. 2020). In ovary, the presence of ATAC-seq peaks alone does not correlate with increased expression without the help of *trans*-acting members of a gene's network. The accumulation of TF partners boosts expression in other tissues more than testis, lowering the probability that a middle-aged gene will be testis biased. Of course, many TF partners are repressive, meaning that their expression would be anticorrelated with that of their target gene. Despite this, older genes with larger TF networks are expressed across a greater variety of tissues and with consistently higher expression levels than younger genes. Old genes likely continue to recruit more TF sites and relationships, and complex regulatory machinery such as enhancers or insulators. These features only marginally increase expression in the permissive transcriptional environment of the testis, but will substantially increase expression in other tissues, especially ovary. This trend will lead to common ovary bias of older genes. The age-related complexity of a gene's TF network may, therefore, drive functional recruitment of young genes to the testis and old genes to the ovary.

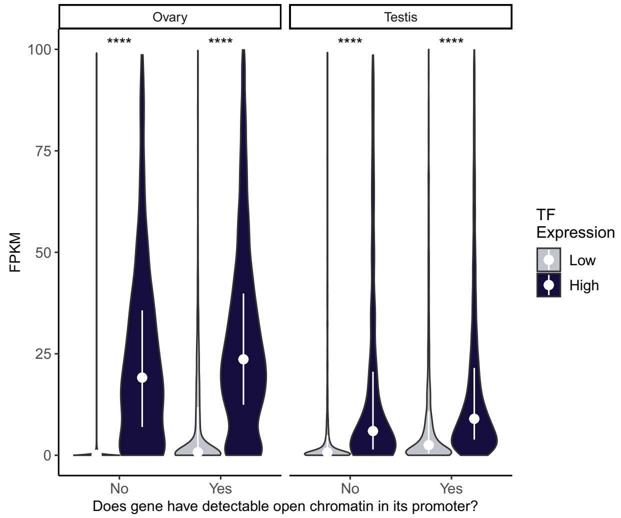

Fig. 5. High TF expression disproportionately predicts gene expression in ovary. For testis and ovary, FPKM for genes with and without detectable chromatin peaks, grouped by "high" or "low" upstream TF expression. Genes are classified as high or low activity in a tissue if they are above or below the median TF expression for genes in the tissue. In testis, both high TF expression and open promoter chromatin confer a similar, modest expression benefit. In ovary, genes with low TF expression are generally very lowly expressed regardless of the presence of a promoter peak. This indicates that TF expression influences ovary expression more than chromatin environment. In ovary, high TF expression is necessary and sufficient for gene expression. White dots are the median values for each group, used to calculate fold changes in table 1. Vertical lines are interquartile ranges. Asterisks represent P values are adjusted with Bonferroni's correction (****$P < 0.00005$)

DroID does not show whether a TF–gene relationship is one of an activator or repressor. For each gene, we use the same set of TF interactions across every tissue, so the activator/repressor balance should not bias the expression/upstream TF expression relationship between tissues. Although our TF expression measure does not consider whether a TF is an activator or repressor, it is still quite predictive of expression. Indeed, the fact that this measure shows a fairly robust Pearson's r with gene expression across tissues might indicate that most of these relationships are activation, consistent with Zhang and Zhou's finding that genes accrue activating TFs and activating epigenetic marks concurrently as they age (Zhang and Zhou 2019). Our TF expression metric relies on the assumption that most TFs are not an activator of a gene in one tissue and a repressor of the same gene in another tissue. It does not require the assumption that all TF–gene interactions are activation. Although TF expression would be a poor method to predict gene expression, TF expression is a useful method to compare the relationship between *trans*-regulation and gene expression between tissues.

It is also true that younger genes are often less studied compared with older genes. Fortunately, the DroID TF–gene interaction database shows ChIP-chip profiles of known TFs, giving us a whole-genome holistic comparison of confirmed TF–gene interactions without regard to the age of the target gene. This means that if a TF binds to the promoter of a younger gene, we will still be able to experimentally confirm this interaction even if the gene's function is unknown. This high-throughput approach means that although we have a comprehensive list of confirmed TF–gene interactions, the activation/repression relationship and network modularity of many of these interactions are not yet known barring future lower-throughput experiments.

It has been proposed that the testis is uniquely positioned to drive the evolution of new genes due to an open

Table 1. High TF Expression Confers a Disproportionate Fold Difference in Gene Expression in Ovary.

Category	Ovary Median FPKM	Ovary Fold Difference	Testis Median FPKM	Testis Fold Difference
Low TF expression, no ATAC peak	0.13	7.00	1.04	3.37
Low TF expression, ATAC peak	0.91		3.50	
Low TF expression, no ATAC peak	0.13	169.23	1.04	6.92
High TF expression, no ATAC peak	22.00		7.20	
Low TF expression, no ATAC peak	0.13	200.77	1.02	9.90
High TF expression, ATAC peak	26.10		10.10	
Low TF expression, ATAC peak	0.91	24.18	3.50	2.06
High TF expression, no ATAC peak	22.00		7.20	
Low TF expression, ATAC peak	0.91	28.68	3.50	2.89
High TF expression, ATAC peak	26.10		10.10	
High TF expression, no ATAC peak	22.00	1.19	7.20	1.40
High TF expression, ATAC peak	26.10		10.10	

Note.—Corresponding to the median values shown in figure 5, these are the pairwise fold differences in median FPKM for genes with and without promoter peaks, and genes with upstream TF expression above or below the median for a tissue. In ovary, genes with no detectable peak have 169.23-fold higher expression if their TF expression is higher than the median TF expression for genes in ovary. In testis, fold differences in median expression are much smaller between groups.

chromatin environment (Kaessmann 2010; Assis 2019). Our findings indicate that this general pattern of open chromatin may be reflected on local levels, where we find a substantial proportion of genes of all age groups with ATAC-seq peaks in their promoters. Our findings indicate, however, that TF expression is more predictive of ovary gene expression than the presence of an ATAC-seq peak. This indicates that *trans*-regulation is especially important for ovarian gene expression, more so than for testis expression.

Even though TF expression is higher in ovary than in testis for every age group of genes, this activity does not result in ovary bias for young and middle-aged genes. A fitting analogy is that testis gene expression is like a bicycle in low gear: easy to initiate movement, but total speed is limited despite the rider's best efforts. Ovary gene expression is more like a bike in high gear: hard to initiate, but given a favorable environment (like biking downhill) the rider can reach greater speeds as a function of their energy input. This may explain why in the long term the ovary becomes a top niche for older genes.

As a good number of the genes in this study originated before multicellular organisms (and therefore animal tissues such as testis and ovary), it is intriguing that such genes are affected by the relationship between gene age and tissue specificity. Our results do not mean that the fate of all genes is to evolve in testis and gain expression in the ovary. Our results are a snapshot of the relationship between gene age and expression pattern as it occurs now, not a reconstruction of a guaranteed path for the evolution of a given gene's expression.

It is instead clear that properties related to gene age differentially influence a gene's potential roles in various tissues. Young genes have relatively few TF binding sites, a state not conducive to expression in most tissues except the testes. Older genes accumulate more TF binding sites (Tuğrul et al. 2015) and gain expression in nontestis tissues. Eventually, adding TF binding sites yields diminishing returns as a gene approaches expression saturation in a tissue. In ovary, however, added TF activity boosts expression more than in other tissues, making ovary-biased expression more likely for older genes. In testis, by comparison, adding TF binding sites appears to have a marginal effect on expression.

Future work could focus on the TF aspect of this model. Given that old genes have more TFs than young genes, we would aim to simulate the evolution of a gene's expression trajectory by adding a variable number of TF sites to the promoter of a reporter construct and analyzing the tissue-specific expression patterns of the construct. This could tell us about the probable evolutionary "fate" of a stereotypical gene's expression: to originate with testis bias, gain expression in every other tissue, but end with the highest expression in ovary. Why ovary currently becomes the top niche remains enigmatic and warrants future studies.

Of course, gene expression evolution takes place over millions or billions of years. Newly originated genes, if they reach fixation in the population, will likely acquire TF sites over time. In another billion years, the regulatory characteristics that today confer testis bias or ovary bias may confer bias toward other tissues or even tissues that have not yet emerged.

Materials and Methods

Processing of FlyAtlas2 RNA-Seq Data

Fastq files of adult FlyAtlas2 tissues were obtained from EBI under accession number PRJEB22205 and reads were trimmed with Trimmomatic, set to remove the Illumina universal adapter. Reads were aligned with Hisat2 (Kim et al. 2016), default parameters to the Flybase dmel-r6.15 genome assembly (Thurmond et al. 2019). Reads with mapping quality less than 10 were removed. FPKM values were calculated with Stringtie (Kim et al. 2016) using default parameters. For each gene, FPKMs were averaged across replicates of a tissue.

Determination of Consensus Gene Ages

To allow for better statistical power and relatively uniform group sizes between gene age groups, we binned genes into three groups: genes that emerged after the pan-Drosophilid divergence (Drosophilid), genes that emerged sometime before the pan-Drosophilid divergence but before the divergence of Bilateria (pre-Drosophilid), and genes that emerged before Bilateria (pre-Bilateria). To define Drosophilid genes, we used genes assigned to branches 1–5 in the gene age data set from Zhang, Vibranovski, Krinsky, et al. (2010). Ages of older genes were assigned using gene ages from Kondo et al. (2017). Genes without ages defined in either data set were not included for figures that segment genes by age but were included for analyses of TF expression and open chromatin that did not consider gene age. For supplementary figures, Supplementary Material online, we reproduced the main figures defining genes from all three age groups only according to the ages assigned by Kondo et al. (2017) and observed no differences that would change our main findings.

Calculation of Tissue Specificity

We used the tau method (Kryuchkova-Mostacci and Robinson-Rechavi 2017) to calculate tissue specificity based on a gene's FPKM across adult tissues, with replicates averaged (Kryuchkova-Mostacci and Robinson-Rechavi 2017). A tau close to 1 indicates a tissue-specific gene, with a tau of 1 indicating that a gene is only expressed in one tissue. A tau close to zero indicates that a gene is equally expressed in every tissue.

Calculation of Scaled Gene Expression

FPKM is not normalized between genes, so we scaled gene expression to compare genes with different thresholds of activity. For a tissue *i*, scaled expression of a gene *j* is log-transformed FPKM in tissue *i* divided by gene *j*'s max logFPKM in any tissue. A scaled expression of 1 is gene *j*'s maximum expression in any tissue, and a scaled expression of 0 means expression is not detected. A scaled expression of 0.5 means that the logFPKM of a gene in a particular tissue is half the maximum observed logFPKM in any tissue.

Calculation of TF Expression for Genes/Tissues

For each gene, we wanted a measure for the activity of its upstream regulators in every tissue. We used the DroID database (Murali et al. 2011), which, for over 700 TFs, lists all genes whose promoters are bound by each TF as annotated with ChIP-chip and ChIP-seq by the modENCODE project (Roy et al. 2010). For this analysis, we only used genes with at least one TF annotated by DroID.

For a gene in a tissue, the TF expression score is the summed scaled expression of all annotated TF partners of that gene in that tissue. For example: if a gene's TF partners have scaled expression values of 1, 1, and 0.5 in a tissue, and 0, 0, 0.5 in another tissue, the activity score for that gene would be 2.5 in the first tissue and 0.5 in the second, reflecting higher network activity in the first tissue. As the TF expression values are scaled first, this measure allows for holistic comparisons of TF expression patterns between genes and tissues. The correlation between open promoter chromatin and TF expression in multiple tissues assures us that this metric measures biologically meaningful activity.

ATAC-Seq of *Drosophila* Testis

We performed ATAC-seq experiment and analysis using 2-day-old testis of *Drosophila melanogaster* RAL517 stain. For each sample, 25 newly emerged males were collected and transferred to three new vials (performed in triplicate). Forty-eight hours later, we dissected testes in cold PBS. Tissues were lysed in 200 μl of ATAC-seq lysis buffer (10 mM Tris-HCl, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% IGEPAL CA-630) and manually homogenized with a plastic pestle, followed by a 1-min incubation on ice, this process was repeated three times. The samples were pelleted at 4 °C (100 g for 10 min) to recover the nuclei. The buffer was removed and the nuclear pellet was resuspended in 200 μl of lysis buffer. The nuclei preparation was filtered through a 30-μm Nitex nylon mesh (Genesee Scientific #57-105); the filter was further washed with another 200 μl of lysis buffer to ensure optimal nuclear recovery. The purified nuclei were isolated by centrifugation at 1,000 × g for 10 min at 4 °C. Following buffer removal, the nuclei were processed for the tagmentation reaction by adding: 12.5 μl Nextera Tagment DNA Buffer, 11.25 μl ddH2O and 1.25 μl Tn5 Transposase (Illumina Kit # FC-121-1030). The reaction was carried out in a thermal cycler for 30 min at 37 °C with an additional mixing step 15 min into the reaction. The fragments were then purified using the Qiagen MinElute PCR Purification Kit (#28004) according to instructions. Libraries were constructed using the same primers as Buenrostro et al. (2015) and following a similar workflow: The purified DNA was first amplified for five cycles by PCR using the NEB Ultra II PCR mix (M0544). Then, an aliquot of the PCR reaction was analyzed by qPCR to determine the remaining optimal number of PCR cycles. Libraries were finally purified using SPRI beads with a two-step size selection protocol with bead-to-sample ratios of 0.55× and 1.00× for the first and second steps, respectively. An aliquot of the purified library was used for quality control, and tested on an Agilent D1000 Tapestation platform, where concentration and peak periodicity were assessed. The samples were additionally tested for quality using Qubit and sequenced on a 75-bp paired-end Hiseq X platform.

Processing of ATAC-Seq Data from Testis, Ovary, and S2 Cells

We generated three replicates of testis ATAC-seq data from *D. melanogaster*. Two replicates of OSC data were used: SRR3503078 and SRR3503086 (Iwasaki et al. 2016). Two replicates of S2 cells were used: SRR5985082 and SRR5985083 (Ibrahim et al. 2018). Reads were aligned with bowtie2 (Langmead and Salzberg 2012), default parameters against the flybase dmel_r6.24 reference genome (Thurmond et al. 2019). BAM files for each tissue were then merged with samtools merge (Li et al. 2009). Macs2 (Zhang et al. 2008) was used to call peaks for each tissue with the –nomodel parameter. The narrowpeak files were then loaded into R for further processing with Chippeakanno (Zhu et al. 2010) (details in supplementary Rmd on Github). Only peaks with a *q* value <0.05 were used. Chippeakanno was run to find peaks overlapping the region 2,000 bp upstream–100 bp downstream of every gene's Transcription Start Site (TSS).

Data Reproducibility

The data needed to reproduce this work can be found in this link https://github.com/LiZhaoLab/TissueSpecificity (last accessed January 22, 2021). It includes calculated FPKM for every gene and tissue in FlyAtlas2, files used to calculate consensus gene ages from Kondo et al. and Zhang et al., narrowpeak files we calculated for each of the three ATAC-seq data sets, a csv file with calculated TF expression (connectivity.csv) for each gene and tissue, and a file from DroID showing every experimentally annotated TF–gene interaction (tf_gene.txt). These files are all referenced by the Rmd script on our Github page. The free web app which allows users to interactively explore our tissue-specificity data for any set of genes without coding experience necessary is: https://zhao.labapps.rockefeller.edu/tissue-specificity/ (last accessed January 22, 2021).

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank Stuart Macdonald from The University of Kansas and Margo Herre from Leslie Vosshall's lab at The Rockefeller University for the help and suggestions on ATAC-seq, and Luis Gracia at the Rockefeller University for the help with the web app. We thank the members of the Zhao lab for helpful discussions and critically reading an earlier version of the manuscript. The work was supported by NIH MIRA R35GM133780. L.Z. was supported by the Robertson Foundation, a Monique Weill-Caulier Career Scientist Award, an Alfred P. Sloan Research Fellowship (FG-2018-10627), a Rita Allen Foundation Scholar Program, and a Vallee Scholar Program (VS-2020-35).

Author Contributions

E.W. and L.Z. conceived the study. N.S. and S.B. generated the ATAC-seq data. E.W. performed all the analysis in this manuscript. E.W. and L.Z. wrote the manuscript with the input from all authors.

Data Availability

Scripts and processed data needed to reproduce figures are deposited in https://github.com/LiZhaoLab/TissueSpecificity. Testis ATAC-seq of *Drosophila melanogaster* Ral517 is deposited at NCBI under biosample accession number SAMN16259271.

References

Assis R. 2019. Out of the testis, into the ovary: biased outcomes of gene duplication and deletion in *Drosophila*. *Evolution (NY)* 73(9):1850–1862.

Begun DJ, Lindfors HA, Thompson ME, Holloway AK. 2006. Recently evolved genes identified from *Drosophila yakuba* and *D. erecta* accessory gland expressed sequence tags. *Genetics* 172(3):1675–1681.

Buenrostro JD, Wu B, Chang HY, Greenleaf WJ. 2015. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr Protoc Mol Biol*. 109:21.29.1–21.29.9.

Ding Y, Zhao L, Yang S, Jiang Y, Chen Y, Zhao R, Zhang Y, Zhang G, Dong Y, Yu H, et al. 2010. A young *Drosophila* duplicate gene plays essential roles in spermatogenesis by regulating several Y-linked male fertility genes. *PLoS Genet*. 6(12):e1001255.

Grive KJ, Hu Y, Shu E, Grimson A, Elemento O, Grenier JK, Cohen PE. 2019. Dynamic transcriptome profiles within spermatogonial and spermatocyte populations during postnatal testis maturation revealed by single-cell sequencing. *PLOS Genet*. 15(3):e1007810.

Ibrahim MM, Karabacak A, Glahs A, Kolundzic E, Hirsekorn A, Carda A, Tursun B, Zinzen RP, Lacadie SA, Ohler U. 2018. Determinants of promoter and enhancer transcription directionality in metazoans. *Nat Commun*. 9(1):4472.

Iwasaki YW, Murano K, Ishizu H, Shibuya A, Iyoda Y, Siomi MC, Siomi H, Saito K. 2016. Piwi modulates chromatin accessibility by regulating multiple factors including histone H1 to repress transposons. *Mol Cell*. 63(3):408–419.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res*. 20(10):1313–1326.

Kim D, Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. 2016. Transcript-level expression analysis of RNA- seq experiments with HISAT, StringTie and Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat Protoc*. 11(9):1650–1667.

Kondo S, Vedanayagam J, Mohammed J, Eizadshenass S, Kan L, Pang N, Aradhya R, Siepel A, Steinhauer J, Lai EC. 2017. New genes often acquire male-specific functions but rarely become essential in *Drosophila*. *Genes Dev*. 31(18):1841–1846.

Kryuchkova-Mostacci N, Robinson-Rechavi M. 2017. A benchmark of gene expression tissue-specificity metrics. *Brief Bioinform*. 18(2):205–214.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods*. 9(4):357–359.

Leader DP, Krause SA, Pandit A, Davies SA, Dow JAT. 2018. FlyAtlas 2: a new version of the *Drosophila melanogaster* expression atlas with RNA-Seq, miRNA-Seq and sex-specific data. *Nucleic Acids Res*. 46(D1):D809–D815.

Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc Natl Acad Sci U S A*. 103(26):9935–9939.

Levran O, Even-Tov E, Zhao L. 2020. A hominid-specific shift in cerebellar expression, upstream retrotransposons, and a potential cis-regulatory mechanism: bioinformatics analyses of the mu-opioid receptor gene. *Heredity (Edinb.)* 124(2):325–335.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16):2078–2079.

Long M, Betrán E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. *Nat Rev Genet*. 4(11):865–875.

Long M, VanKuren NW, Chen S, Vibranovski MD. 2013. New gene evolution: little did we know. *Annu Rev Genet*. 47(1):307–333.

Majic P, Payne JL. 2020. Enhancers facilitate the birth of de novo genes and gene integration into regulatory networks. *Mol Biol Evol*. 37(4):1165–1178.

Mayr C. 2016. Evolution and biological roles of alternative 3′UTRs. *Trends Cell Biol*. 26(3):227–237.

Murali T, Pacifico S, Yu J, Guest S, Roberts GG, Finley RL. 2011. DroID 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for *Drosophila*. *Nucleic Acids Res*. 39(Suppl 1):D736–D743.

Ohno S. 1970. Evolution by gene duplication. Berlin/Heidelberg: Springer Berlin Heidelberg.

Palmieri N, Kosiol C, Schlötterer C. 2014. The life cycle of *Drosophila* orphan genes. *Elife*. 3:e01311.

Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF, et al.; The modENCODE Consortium. 2010. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* 330(6012): 1787–1797.

Schmidt EE. 1996. Transcriptional promiscuity in testes. *Curr Biol*. 6(7):768–769.

Sigalova OM, Shaeiri A, Forneris M, Furlong EEM, Zaugg JB. 2020. Predictive features of gene expression variation reveal mechanistic link with differential expression. *Mol Syst Biol*. 16(8):e9539.

Soumillon M, Necsulea A, Weier M, Brawand D, Zhang X, Gu H, Barthès P, Kokkinaki M, Nef S, Gnirke A, et al. 2013. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep*. 3(6):2179–2190.

Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. *Nat Rev Genet*. 12(10):692–702.

Thurmond J, Goodman JL, Strelets VB, Attrill H, Gramates LS, Marygold SJ, Matthews BB, Millburn G, Antonazzo G, Trovisco V, et al.; The FlyBase Consortium. 2019. FlyBase 2.0: the next generation. *Nucleic Acids Res*. 47(D1):D759–D765.

Trizzino M, Park Y, Holsbach-Beltrame M, Aracena K, Mika K, Caliskan M, Perry GH, Lynch VJ, Brown CD. 2017. Transposable elements are the primary source of novelty in primate gene regulation. *Genome Res*. 27(10):1623–1633.

Tuğrul M, Paixão T, Barton NH, Tkačik G. 2015. Dynamics of transcription factor binding site evolution. *PLoS Genet*. 11(11):e1005639.

Vaid R, Wen J, Mannervik M. 2020. Release of promoter-proximal paused Pol II in response to histone deacetylase inhibition. *Nucleic Acids Res*. 48(9):4877–4890.

Vinckenbosch N, Dupanloup I, Kaessmann H. 2006. Evolutionary fate of retroposed gene copies in the human genome. *Proc Natl Acad Sci U S A*. 103(9):3220–3225.

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

Xia B, Yan Y, Baron M, Wagner F, Barkley D, Chiodin M, Kim SY, Keefe DL, Alukal JP, Boeke JD, et al. 2020. Widespread transcriptional scanning in the testis modulates gene evolution rates. *Cell* 180(2):248–262.e21.

Zhang J-Y, Zhou Q. 2019. On the regulatory evolution of new genes throughout their life history. *Mol Biol Evol*. 36(1):15–27.

Zhang L, Tan Y, Fan S, Zhang X, Zhang Z. 2019. Phylostratigraphic analysis of gene co-expression network reveals the evolution of functional modules for ovarian cancer. *Sci Rep*. 9(1):2623.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nussbaum C, Myers RM, Brown M, Li W, et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol*. 9(9):R137.

Zhang YE, Vibranovski MD, Krinsky BH, Long M. 2010. Age-dependent chromosomal distribution of male-biased genes in *Drosophila*. *Genome Res*. 20(11):1526–1533.

Zhang YE, Vibranovski MD, Landback P, Marais GAB, Long M. 2010. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. *PLoS Biol*. 8(10):e1000494.

Zhang Z, Hambuch TM, Parsch J. 2004. Molecular evolution of sex-biased genes in *Drosophila*. *Mol Biol Evol*. 21(11): 2130–2139.

Zhao L, Saelao P, Jones CD, Begun DJ. 2014. Origin and spread of de novo genes in *Drosophila melanogaster* populations. *Science* 343(6172):769–772.

Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, Zhan Z, Li X, Ding Y, Yang S, Wang W. 2008. On the origin of new genes in *Drosophila*. *Genome Res*. 18(9):1446–1455.

Zhu LJ, Gazin C, Lawson ND, Pagès H, Lin SM, Lapointe DS, Green MR. 2010. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics* 11(1):237.