Title

Population biology of accessory gland-expressed de novo genes in Drosophila melanogaster

Permalink

https://escholarship.org/uc/item/6bm706vk

Journal

Genetics, 220(1)

ISSN

0016-6731

Authors

Cridland, Julie M
Majane, Alex C
Zhao, Li
et al.

Publication Date

2022-01-04

DOI

10.1093/genetics/iyab207

Peer reviewed

OXFORD GENETICS

GENETICS, 2022, 220(1), iyab207
https://doi.org/10.1093/genetics/iyab207
Advance Access Publication Date: 17 November 2021
Investigation

Population biology of accessory gland-expressed *de novo* genes in *Drosophila melanogaster*

Julie M. Cridland,[1,*] Alex C. Majane,[1] Li Zhao,[2] and David J. Begun[1]

[1]Department of Evolution and Ecology, University of California, Davis, Davis, CA 95616, USA and
[2]Laboratory of Evolutionary Genetics and Genomics, The Rockefeller University, New York, NY 10065, USA

*Corresponding author: Department of Evolution and Ecology, University of California—Davis, 2320 Storer Hall, Davis, CA 95616, USA.
Email: jmcridland@ucdavis.edu

Abstract

Early work on *de novo* gene discovery in *Drosophila* was consistent with the idea that many such genes have male-biased patterns of expression, including a large number expressed in the testis. However, there has been little formal analysis of variation in the abundance and properties of *de novo* genes expressed in different tissues. Here, we investigate the population biology of recently evolved *de novo* genes expressed in the *Drosophila melanogaster* accessory gland, a somatic male tissue that plays an important role in male and female fertility and the post mating response of females, using the same collection of inbred lines used previously to identify testis-expressed *de novo* genes, thus allowing for direct cross tissue comparisons of these genes in two tissues of male reproduction. Using RNA-seq data, we identify candidate *de novo* genes located in annotated intergenic and intronic sequence and determine the properties of these genes including chromosomal location, expression, abundance, and coding capacity. Generally, we find major differences between the tissues in terms of gene abundance and expression, though other properties such as transcript length and chromosomal distribution are more similar. We also explore differences between regulatory mechanisms of *de novo* genes in the two tissues and how such differences may interact with selection to produce differences in *D. melanogaster de novo* genes expressed in the two tissues.

Keywords: *Drosophila*; evolution; RNA-seq; de novo genes; accessory gland; testis

Introduction

The especially rapid divergence of male-limited behavioral and morphological phenotypes in many animal lineages, presumed to be a consequence of various forms of sexual selection, is mirrored in the genome, most conspicuously in the portion functioning specifically in male-specific reproductive tissues. This rapid divergence of genes exhibiting male-biased or male-specific expression applies to several evolutionary phenomena observed in *Drosophila*, including protein sequence evolution (*e.g.*, Coulthart and Singh 1988; Swanson *et al.* 2001; Wagstaff and Begun 2005; Haerty *et al.* 2007), the evolution of canonical gene duplications (*e.g.*, Wagstaff and Begun, 2005; Mikhaylova *et al.* 2008; Belote and Zhong 2009; Sorourian *et al.* 2014), gene expression divergence (*e.g.*, Meiklejohn *et al.* 2003; Zhang *et al.* 2007), and the origination of genetic novelties, such as retrogene duplications (Long and Langley 1993; Betrán *et al.* 2002) and *de novo* genes (*e.g.*, Begun *et al.* 2006; Levine *et al.* 2006), the last of which is the focus of this report.

We define *de novo* genes here as DNA sequences producing derived transcripts, coding or noncoding, that are independent of mature ancestral transcripts and located in ancestrally intergenic or intronic DNA. Such genes have been found in several taxa, including *Drosophila* (Begun *et al.* 2006, 2007; Levine *et al.* 2006; Zhou *et al.* 2008; Zhao *et al.* 2014), rodents (Heinen *et al.* 2009; Murphy and McLysaght 2012; Neme and Tautz 2013; Casola 2018), primates (Knowles and McLysaght 2009), plants (Zhang *et al.* 2019), and fungi (Cai *et al.* 2008; Li *et al.* 2010; Carvunis *et al.* 2012; Vakirlis *et al.* 2018). Nevertheless, the abundance (Casola 2018), persistence times (Palmieri *et al.* 2014), and functions of *de novo* genes (*e.g.*, Cai *et al.* 2008; Heinen *et al.* 2009; Li *et al.* 2010) remain unclear. While the identification of *de novo* genes could be viewed operationally as an annotation problem, it is challenging for several reasons, some of which may derive from the properties of *de novo* genes themselves (*e.g.*, low expression levels; Zhao *et al.* 2014) and others of which derive from the fact that identification of *de novo* genes relies upon marshaling evidence in support of gene absence in orthologous DNA of nonfocal species (reviewed in Van Oss and Carvunis 2019).

The first experimental investigation of *de novo* gene evolution (Begun *et al.* 2006) took place in the context of the accessory gland (AG), which produces among other molecules, secreted proteins that are transferred to the female along with sperm during mating. These molecules are required for fertility, mediate a number of female postmating physiological responses, and may also influence female sperm storage and sperm competition (reviewed in Wilson *et al.* 2017; Wigby *et al.* 2020). In that work, early *Drosophila* genome assemblies from the *melanogaster* subgroup (Begun *et al.* 2007; Clark *et al.* 2007) and AG-derived expressed sequence tags from cDNA libraries were used to reveal evidence of

Received: September 14, 2021. **Accepted:** November 08, 2021

several small, AG-expressed genes in *Drosophila yakuba* or *Drosophila erecta* that appeared to be unexpressed in related species, but for which orthologous, syntenic sequence could be identified (Begun *et al.* 2006). Those genes were hypothesized to have recently originated *de novo* from ancestral intergenic DNA in *D. yakuba*, *D. erecta*, or their common ancestor.

An early annotation-based phylogenetic investigation of *de novo* gene evolution in the *melanogaster* subgroup (Levine *et al.* 2006) found that *de novo* genes often exhibit testis-biased or testis-specific expression, thereby providing the first clue that *D. melanogaster de novo* genes may be biased toward male reproductive functions. A similar pattern was observed in the obscura group of *Drosophila* (Palmieri *et al.* 2014). To investigate the youngest class of *de novo* genes, Zhao *et al.* (2014) carried out a detailed population level investigation of testis transcriptomes in a sample of six *D. melanogaster* inbred genotypes, which detected 106 putative *de novo* genes that had fixed since the split from common ancestor with *Drosophila simulans*, and 142 that were segregating in *D. melanogaster* (Zhao *et al.* 2014). Many of the polymorphic testis-expressed *de novo* genes occurred at intermediate or high frequency, and population genetic evidence suggested that these genes had been influenced by directional selection. Nevertheless, because that study investigated only the testis it could not speak to the question of whether the abundance and population biology properties of *de novo* genes in the testis are typical or atypical. Thus, while the *Drosophila* data point overall to a role for *de novo* genes in the evolution of genetic novelty in both testis and AG transcriptomes, comparison of the abundance and properties of *de novo* genes expressed in these two organs awaits more thorough investigation of the *D. melanogaster* AG.

Several functional or evolutionary attributes of AG function could facilitate the origin and spread of AG-expressed *de novo* genes. First, AG-specific proteins tend to be small (*e.g.*, Findlay *et al.* 2009). If *de novo* genes are protein-coding and often originate from ancestrally noncoding DNA carrying latent open reading frames (ORFs), shorter *de novo* genes would be more common than longer ones simply because *ceteris paribus*, shorter latent ORFs are more abundant than longer latent ORFs in noncoding eukaryotic DNA. Second, while the majority of secreted seminal fluid proteins require a signal peptide, the protein sequence constraints for signal peptides are fairly lax (Nielsen *et al.* 1997), which might also lead to high origination rates of novel proteins competent for secretion (Begun *et al.* 2006). Third, while many protein functional domains are widely shared among seminal fluid proteins, several *Drosophila* seminal fluid proteins have no known functional domains (Findlay *et al.* 2008). Genes coding for such proteins may be more likely than many functional gene classes to have atypical structures or functions, and thus, more likely to arise *de novo*. Finally, regardless of whether most *de novo* evolved genes are coding *vs* noncoding (see below), strong selection favoring novelty in male–male or male–female interactions could facilitate the spread of *de novo* AG-expressed genes. Thus, while the fixation rate of *de novo* genes will always depend on the cellular processes underlying the expression of noncoding or nongenic DNA (Begun *et al.* 2006), the breadth of properties of novel proteins or RNAs capable of functioning in a particular tissue or cell type, and the strength of selection acting on evolutionary novelties in a given tissue or cell type, the investigation of how these factors may interact to influence *de novo* gene origination and fixation is still in its infancy.

One possibility is that the phenomena promoting the spread of *Drosophila de novo* genes are relatively homogeneous across tissues relating to male-specific reproductive functions. In that case, we would expect roughly similar contributions of *de novo* genes to testis and AG transcriptomes, and roughly similar dynamics (selected or neutral) of *de novo* gene spread. Alternatively, differences between the tissues could reflect the strength and/or nature of selection. For example, greater fixation rates of AG-expressed *de novo* genes might indicate stronger selection on novelty associated with male–male or male–female postcopulatory phenotypes, while greater fixation rates of testis-expressed *de novo* genes could be indicative of greater selection favoring novelty in germline phenomena, perhaps due to genomic conflicts associated with sex-ratio *X* chromosomes (Levine *et al.* 2006) or transposable elements. Differences between the regulatory environments of AG *vs* testis cells could affect the rate at which novel transcripts originate. Finally, functional differences between cells and tissues in the testis *vs* AG could lead to differences in the universe of novel proteins or RNAs that are not strongly deleterious and thus could be exposed to positive selection in different cellular milieus. Here, we begin addressing some of these questions through an investigation of young AG-expressed *de novo* genes that originated in *D. melanogaster* since the split from its sibling species, *D. simulans*.

Materials and methods

Fly strains, datasets, and sequencing

Most of the data used here are described in detail in Cridland *et al.* (2020). Briefly, we dissected AG + anterior ejaculatory duct (referred to throughout as AG) of 2-day-old virgin males from six highly inbred *D. melanogaster* strains established by the *Drosophila* Genetic Reference Panel (DGRP; Mackay *et al.* 2012): RAL-304, RAL-307, RAL-357, RAL-360, RAL-399, and RAL-517. We carried out an allelic imbalance assay using F1 flies generated from crosses among DGRP lines (RAL-307 male × RAL-304 female, RAL-357 male × RAL-399 female, RAL-360 male × RAL-517 female). Two inbred strains of *D. simulans* were used. One strain, w^{501}, was the strain used for the *D. simulans* reference sequence (Begun *et al.* 2007); a second strain, Lara10, was established from flies collected September 2011 in Homestead, FL, and was sibmated for 10 generations in our laboratory (Zhao *et al.* 2014). We used the *D. yakuba* reference sequence strain, Tai18E2 (Begun *et al.* 2007) as our second outgroup for most analyses (existing data from a third outgroup, *Drosophila ananassae* strain 14021-0371.13, was also used for some analyses—Yang *et al.* 2018). All flies were reared on standard cornmeal medium at 25°C under a 12:12 light/dark cycle. RNA was extracted using Trizol (Invitrogen). RNA-seq libraries were made and sequenced as described in Cridland *et al.* (2020), resulting in paired-end, 100 bp reads.

De novo transcriptome assemblies

Trinity (Grabherr *et al.* 2011; v2.11) was used to create *de novo* transcriptome assemblies for each species, using both individual strains and pooled data across strains from each species (Supplementary Table S1). Based on the k-mer distribution generated by Jellyfish V1.1.5, a k-mer of 25 was used for assembly. We also used Trinity using default parameters to carry out *de novo* transcriptome assemblies for eight tissues and both sexes for *D. yakuba* and *D. ananassae* (Yang *et al.* 2018).

Criteria for calling *D. melanogaster* AG-expressed *de novo* genes

We used BLAST (v. 2.10.1+, Altschul *et al.* 1990) to screen all transcripts from our *D. melanogaster* AG *de novo* transcriptome assemblies that were greater than 300 bp long for matches to known genes or transcripts from *D. melanogaster* (v. 6.34), *D. simulans* (v. 2.02), and *D. yakuba* (v. 1.05). The files used for screening were fasta files from each species containing records for CDS, exon, 5′UTR, miRNA, miscRNA, ncRNA, pseudogene, transcript, transposon, tRNA, and 3′UTR (downloaded from www.flybase.org, July 23, 2020; Thurmond *et al.* 2019). To consider a transcript, a match we required 80% identity over at least 100 bp. We also separately aligned transcripts to *D. melanogaster* introns to identify potential intronic *de novo* gene candidates. Transcripts that matched only intronic or intergenic sequences were retained for further analysis. To further reduce the likelihood of erroneously inferring *de novo* gene status for ancestral genes unannotated in *D. melanogaster* we also screened our *D. melanogaster* AG transcripts against Trinity-generated *de novo* transcriptome assemblies from our outgroup AG RNA-seq data, as well as against Trinity-generated *de novo* assemblies derived from *D. yakuba* and *D. ananassae* libraries from eight different tissues (Yang *et al.* 2018, files downloaded from SRA January 2021). Thus, all *D. melanogaster* transcripts matching existing *D. melanogaster* or outgroup gene annotations, or any outgroup transcripts we assembled, were removed from further consideration.

To reduce the likelihood of mistaking an unannotated *D. melanogaster* exon of an ancestral gene for a *de novo* gene we required each candidate, intergenic or intronic, to be at least 500 bp from any annotated exon boundary. We then generated a GTF file for this set of candidates, including all transcripts of each candidate *de novo* gene, combined these new records with the *D. melanogaster* v6.34 GTF file, and estimated TPMs for each of the six inbred RAL lines separately. Candidates with a TPM $\geq$ 1 in one or more RAL lines were retained.

To confirm that the remaining *de novo* gene candidates reside in orthologous DNA in all three main species we performed a microsynteny analysis by identifying the nearest neighbor genes of each candidate and identified their orthologs in *D. simulans* and *D. yakuba*. The location of these orthologs in the outgroup genomes was used to confirm that the candidate gene plus 5 kb upstream and downstream of the transcript start and stop aligned to the syntenic region. Most synteny analyses were carried out using a perl script to compare the positions of the candidate to the syntenic region. The remainder were checked manually, largely due to small deletions in the outgroup(s) in the syntenic region that resulted in the perl script flagging the candidate for a manual check.

The final list of *D. melanogaster* candidates have the following attributes: they do not overlap existing exon annotations in *D. melanogaster*, *D. simulans*, or *D. yakuba* and are at least 500 bp from known exons, they are expressed at TPM $\geq$ 1 in at least one inbred *D. melanogaster* genotype, they reside in orthologous regions of all three species, and they exhibit no evidence of expression in any tissue of any of the three outgroup species (*D. simulans*, *D. yakuba*, or *D. ananassae*). Notably, this approach is expected to be more conservative than the one previously used for our analysis of the testis (Zhao *et al.* 2014), as we impose no minimum expression level on outgroup TPM estimates and we include a substantial amount of new outgroup transcriptome data. Segregating genes are defined as those for which at least one line expresses at TPM $\geq$ 1 and at least one expresses at TPM $<$ 1 (Cridland *et al.* 2020). "Fixed" genes are those expressed at TPM $\geq$ 1 in all six Raleigh inbred lines. Once a final list of *de novo* genes was generated, we estimated TPMs in the three Raleigh F1s.

Sequence alignments, variant calling, and allelic imbalance

Our methods closely follow Cridland *et al.* (2020) which generally follows McManus *et al.* (2010). Briefly, parental RAL TPM estimates and corresponding estimates from their F1s were used to partition variation into *cis* and *trans* effects. We used a fold-change cutoff of 1.25 to call differences in (1) expression between RAL parents, (2) between parent-specific estimates in hybrids, and (3) between the observed overall F1 expression and the expected F1 expression assuming additivity. To reduce the influence of noise on inferring *cis*- and *trans*-effects we restricted the analysis to genes for which at least one parent expressed at TPM $\geq$ 1 and the other parent expressed at TPM $<$ 0.2. We further required at least 10 unique fragments from the F1s to include the observation. Genes were categorized as exhibiting *cis*- or *trans*-effects as described in Cridland *et al.* (2020).

Coding potential and signal peptide prediction

We used the Coding Potential Assessment Tool (CPAT; Wang *et al.* 2013) to estimate the probability that a transcript was derived from a protein-coding gene *vs* noncoding gene, generating sets of the top five ORFs per transcript. We used SignalP 5.0 (Armenteros *et al.* 2019) to determine the probability of a signal sequence for each of the most likely ORFs predicted by CPAT.

Ancestral AG-biased genes

To identify ancestral AG-biased genes, we used the male data for several tissues from FlyAtlas 2 (Leader *et al.* 2018). We defined AG-biased genes as those that: (1) had FPKM $\geq$ 1 in the AG, (2) exhibited the highest expression in the AG relative to other male tissues, and (3) exhibited strong AG bias, with estimated *tau* (Yanai *et al.* 2005) $\geq$ 0.9.

Physical distribution of *de novo* genes

We investigated the physical distribution of *de novo* genes at several scales, ranging from entire chromosome arms to regions of a few kilobases. To determine whether *de novo* genes tend to be co-localized with ancestral AG-biased genes we segmented the genome into 500- and 100-kb nonoverlapping windows and asked whether *de novo* genes are more likely than expected to fall in windows harboring ancestral AG-biased genes. Some *de novo* genes are tandemly located, defined here as being adjacent and $\leq$10 kb apart. To ascertain whether such *de novo* genes tend to show correlated expression across genotypes, we compared the frequency with which genes in clusters tend to be expressed or not expressed together across genotypes. We first converted the TPMs of each *de novo* gene in a cluster to 0 (if TPM $<$ 1) or 1 (if TPM $\geq$ 1). We then permuted the expression of each gene over the six RAL lines to generate 1000 sets of *de novo* genes with expression randomized over lines but with the number of lines expressing each *de novo* gene preserved. We then calculated an index for each gene cluster that summarized the number of RAL lines that had the same expression value, 0 for not expressed or 1 for expressed, for all the genes within a given cluster. The mean for this index was then calculated across all *de novo* gene clusters for observed and permuted data. To further examine if expressed genes in a cluster were more likely to be expressed in the same RAL line(s) than would be expected, we calculated a second index where we determined the number of RAL lines per cluster for which all *de novo* genes were expressed and then compared the mean across all clusters to the distribution of means from the permuted data.

Comparison to testis *de novo* genes

To compare AG *de novo* genes to previously identified testis-expressed *de novo* genes from the same inbred lines (Zhao *et al.* 2014) we applied the methods described above to the previously reported data from the testis (Zhao *et al.* 2014), which enables direct comparison of the putative *de novo* genes expressed in the two tissues. We used BLAST to compare the testis candidates to *D. simulans* and *D. yakuba* annotations to transcript assemblies made from our *D. simulans* and *D. yakuba* AG + ejaculatory duct libraries, and to outgroup *de novo* transcript assemblies made from RNA-seq data from several *D. yakuba* and *D. ananassae* tissues (Yang *et al.* 2018), as described above. To enable direct comparisons of *cis*- and *trans*- effects on *de novo* gene expression for testis and AG, we subjected the set of previously identified testis-expressed *de novo* gene candidates to the same pipeline described above used for candidate AG-expressed *de novo* genes.

Results

Basic attributes of AG-expressed *de novo* genes

We identified a total of 133 candidate *de novo* genes (49 intergenic and 84 intronic; Supplementary Table S2) in the six DGRP strains, of which 131 were segregating and two (both intronic) were "fixed" (expressed at TPM >1 in all six Raleigh inbred lines); 99 genes were expressed in only one strain, while 34 were expressed in more than one strain. For convenience, we sometimes refer to these genes as "de novo genes" rather than "candidate de novo genes," despite the absence of evidence for genic function. While ancestral genes nested in introns of annotated genes are strongly biased (71.3%) toward being on the opposite strand in *Drosophila* (Lee and Chang 2013), multiexonic candidate *de novo* genes located in introns of ancestral genes were roughly equally likely to be on the same strand *vs* opposite strand ($n = 10$ and 9, respectively), which is significantly different from ancestral genes (binomial probability, $P = 0.025$). The mean (median) number of AG-expressed *de novo* genes per line across the six inbred lines was 33.5 (32.5). The greatest and least number of candidates were expressed in line RAL 517 (54 expressed genes) and RAL 307 (21 expressed genes), respectively. We observed no significant correlation between the mean TPM of ancestral AG-biased genes for each line and the number of candidate *de novo* genes it expressed.

To compare the number of AG-expressed *de novo* genes to that observed in the testis we first reassessed the candidate genes reported in Zhao *et al.* (2014), all of which were intergenic as constrained by the filtering process used. Of the 106 fixed and 142 segregating genes reported in Zhao *et al.* (2014) we discovered apparently homologous transcripts, defined as BLAST matches to outgroup databases of ≥80% over ≥100 bp in one or more outgroups, for 53 fixed and 27 segregating candidates (Supplementary Table S3). Therefore, for the same strains used here for the AG, our current conservative estimate for fixed and segregating testis-expressed *de novo* genes are 53 and 115, respectively. Comparing the number of intergenic *de novo* genes expressed in the two tissues relative to the total number of annotated genes expressed in those tissues at TPM ≥ 1 reveals greater than twofold more testis-expressed than AG-expressed *de novo* genes. We observed a median/mean of 10/12 intergenic AG-expressed genes per strain, while the median number of intergenic testis-expressed *de novo* genes expressed per strain (in the same six RAL strains) was 99. Testis-expressed *de novo* genes were more likely to be expressed in more than one RAL strain (70% of candidates) than AG *de novo* genes (26% of candidates; Fisher's exact test; $P = 7.6 \times e^{-15}$). Overall then, it seems safe to conclude that intergenic *de novo* genes make a substantially smaller contribution to AG transcriptome complexity than to testis transcriptome complexity.

The mean length of the longest transcript for AG intergenic candidate genes, 701 bp, was shorter than the mean for intergenic testis-expressed *de novo* genes (935 bp, Zhao *et al.* 2014; *t*-test, $P = 0.001$). Nineteen AG-expressed genes were associated with a transcript > 1000 bp; the longest observed transcript was 2214 bp. There was no significant difference in the length of the longest transcript per gene for intergenic *vs* intronic candidates (Wilcoxon test; $P = 0.52$). Most genes (106/133, 80%) were single-exon; the maximum exon-number for any transcript was three. While the majority (94%) of intron splice junctions were canonical GT/AG, this proportion is significantly smaller than that observed for ancestral genes (binomial; $P < 0.001$; Crosby *et al.* 2015). Twenty-eight genes (including six single-exon genes) exhibited multiple transcripts, while the maximum number of transcripts for any gene was four. The maximum number of transcripts/gene and exons/gene were both positively, though weakly correlated with expression levels (mean of expressing RAL line TPM estimates; $P = 3.5 \times e^{-4}$ and $P = 9.4 \times e^{-3}$, respectively), similar to Zhao *et al.* (2014).

As expected, the expression of many candidate genes (TPMs: expressed mean 2.67, median 1.39) is low relative to that of ancestrally expressed genes (TPMs: mean 106, median 6.2). However, Supplementary Figure S1, which shows the distribution of max TPM across strains for all candidates, reveals that a considerable number of genes show relatively high maximum across-strain TPMs. For example, 18 genes exhibited a maximum TPM ≥ 5, with the most highly expressed gene expressed at maximum TPM = 53 (though this gene is only expressed in two strains). Genes expressed at a higher level (mean of nonzero TPM estimates) tended to be expressed in more lines (Kruskal–Wallis test, $P = 1.6 \times e^{-5}$). Comparing singleton genes (expressed in only one genotype) *vs* nonsingleton genes, we find that nonsingletons have a greater median longest transcript (696 *vs* 485, $P = 1.36 \times e^{-6}$) and greater median maximum TPM (3.26 *vs* 1.25, $P = 4.46 \times e^{-8}$). Thus, the overall picture is consistent with the literature—candidate *de novo* genes tend to be short, simple, and lowly expressed, some candidate genes, nevertheless, exhibit high expression and multiple transcripts, and there is a general trend for longer transcripts and greater expression to be associated with genes expressed in more genotypes (Zhao *et al.* 2014).

Coding potential

Using CPAT with default settings based on a *D. melanogaster* training set (Wang *et al.* 2013) we identified the top five ORFs for each of the 170 *de novo* transcripts (corresponding to 133 genes) and for each ORF determined the coding *vs* noncoding likelihood. An ORF was identified by CPAT for 165 of 170 transcripts. This analysis revealed that 98% of genes (131 of 133) and 99% of transcripts (163/165) were predicted to be noncoding. Both predicted coding transcripts were X-linked and expressed in more than one RAL strain. To investigate whether the unusually short transcript length of *de novo* gene candidates relative to the lengths of ancestral annotated protein-coding genes used to train CPAT was potentially biasing this conclusion, we used CPAT to categorize all annotated *D. melanogaster* protein-coding transcripts as either coding or noncoding, and then binned these annotated protein-coding genes by transcript length in 100-bp increments up to 2 kb (Supplementary Figure S2). Assuming that annotated protein-coding status in Flybase is correct, this analysis revealed that very short protein-coding genes are more likely than other protein-coding genes to be predicted by CPAT to be noncoding. To determine whether this possible bias could influence our conclusion about *de novo* gene coding probability, we used a resampling procedure. We made random draws without replacement from annotated *D. melanogaster* transcripts to generate sets of 165 transcripts with the same length distribution (based on length bins in 100-bp increments) as the observed *de novo* gene candidates, and repeated this 1000 times. We then compared the predicted coding probability of the highest scoring ORF for each annotated transcript to that of each *de novo* gene transcript. In none of the 1000 permutations did we observe a fraction of coding genes as small or smaller than the observed value, with the most extreme permutation still exhibiting 112/165 transcripts as compared to only 2/165 observed. On average 130/165 transcripts from these draws exhibited a coding probability score ≥0.39 (the cutoff for calling a *D. melanogaster* gene as coding), yielding a binomial test $P < 1 \times e^{-7}$. This suggests that the finding that a substantial proportion of candidate *de novo* genes is likely noncoding is probably not attributable solely to artifacts or biases associated with CPAT.

The five most likely ORFs predicted for each transcript by CPAT were also used to determine the probability that the corresponding predicted proteins harbored a signal sequence (SignalP 5.0; Armenteros *et al.* 2019). Only six were predicted to be secreted; none of the predicted proteins showed similarity to known proteins. Thus, if these genes are in fact protein coding, it seems unlikely for most that their products are transferred to females during mating. Using male data from FlyAtlas 2 and a cutoff of *tau* > 0.9 (Yanai *et al.* 2005) revealed 538 ancestral strongly AG-biased genes. Of these, 324 were identified as coding based on the FlyBase annotation, 239 of which (73.8%) contained predicted signal sequences (Armenteros *et al.* 2019). In contrast, 209 ancestral AG-biased genes were annotated as noncoding in FlyBase. To investigate the possibility that some of these putatively noncoding genes are actually coding genes producing secreted proteins we used CPAT to identify associated ORFs. For those genes for which CPAT identified at least one ORF, we used SignalP to determine whether the most likely ORF contained a predicted signal sequence. Of the 209 AG-biased genes annotated as noncoding, 38 (18%) contained predicted signal sequences (Supplementary Tables S4 and S5). It is quite plausible that these noncoding genes are misannotated and in reality code for seminal fluid proteins. In addition, ORFs of 12 AG-biased ncRNAs match FlyBase polypeptide sequences (Supplementary Tables S4 and S5), including three of the 38 with predicted signal sequences. Thus, we speculate that 47 of the 209 AG-biased genes annotated as noncoding are likely to be coding. Nevertheless, the proportion of candidate *de novo* genes predicted to be noncoding (0.98) is much greater than the proportion of ancestral AG-biased genes likely to be noncoding (162/538 = 0.30; binomial $P = 1.45 \times e^{-67}$). Thus, AG-expressed *de novo* genes appear much more likely to be noncoding relative to ancestral AG-biased genes.

Physical distribution of candidate genes at different scales

To investigate the genomic distribution of *de novo* genes we first asked whether the proportion of genes on each chromosome arm differs from that observed for all ancestral AG-biased genes. Consistent with previous reports for male-biased genes in general (Sturgill *et al.* 2007) and seminal fluid protein genes and AG-biased genes specifically (Findlay *et al.* 2008; Meisel *et al.* 2012), ancestral AG-biased genes are underrepresented on the *X* ($n = 33$) relative to the major autosomes ($n = 504$, or 126 per major autosomal arm; Fisher's exact test $P = 3.2 \times e^{-10}$). Across autosomes, arm *2L* ($n = 159$ genes) is significantly enriched for these genes, also consistent with the literature (Findlay *et al.* 2008; Table 1). Candidate *de novo* genes exhibit similar chromosomal patterns; relatively few are located on the *X*, while *2L* harbors the greatest number. There is no significant deficit of AG-expressed *X*-linked *de novo* genes relative to the expected value based on the fraction of all annotated genes that are *X*-linked. However, direct comparison of *de novo* genes and ancestral AG-biased genes on the *X* chromosome *vs* autosomes reveals that the *X/A* ratio for *de novo* genes (16/117 = 0.14) is about twice that of ancestral genes (33/505 = 0.065; Fisher's exact test; $P = 0.025$). Thus, whatever processes lead to the strong autosomal bias of ancestral AG-biased genes (Meisel *et al.* 2012) are weaker for *de novo* genes. Because these *de novo* genes are polymorphic and expressed at a low level, reduced expression constraints associated with *X*-linkage and dosage compensation (Meisel *et al.* 2012) might contribute to this pattern. We then determined the *X/A* ratio for testis-expressed *de novo* genes (Zhao *et al.* 2014) and ancestral testis-biased genes (tau > 0.9 and highest male expression in the testis in male FlyAtlas 2 data). For testis-expressed *de novo* genes the *X/A* ratio (14/154 = 0.09) is smaller than the ratio for ancestral testis-biased genes (411/2580 = 0.16; Fisher's exact test, $P = 0.048$). Thus, while the two classes of *de novo* genes, AG- and testis-expressed, are very similar in their *X/A* distributions (Fisher's exact test; $P = 0.33$), deviations from comparable ancestral-biased genes *X/A* distributions are very different, with testis-biased candidate *de novo* genes showing X underrepresentation and AG-biased candidate genes showing *X* overrepresentation. Interestingly, while both testis- and AG-biased ancestral genes are underrepresented on the *X*, the effect is roughly twofold greater for AG-biased genes (*cf.* Meisel *et al.* 2012).

To investigate within chromosome-arm heterogeneity and the possible connection between the locations of AG-expressed *de novo* genes and ancestral AG-biased genes, we segmented each chromosome arm into 500-kb windows and then asked whether windows harboring an ancestral AG-biased gene were also more likely to harbor a *de novo* gene. We found that consistently across all arms (genome-wide Fisher's exact test; $P = 2.6 \times e^{-8}$), *de novo* genes were much more likely to reside in windows containing an

Table 1 AG *de novo* candidate genes

Chromosome	AG *de novo* genes	*De novo* enrichment	AG-biased genes	AG-biased enrichment	Total genes
2L	43 (32.3%)	7.14E–04	159 (29.6%)	3.32E–08	3559 (20%)
2R	20 (15%)	1.33E–01	92 (17.1%)	2.18E–02	3673 (20.6%)
3L	15 (11.3%)	1.58E–02	119 (22.1%)	7.41E–02	3501 (19.6%)
3R	37 (27.8%)	3.09E–01	134 (24.9%)	2.71E–01	4262 (23.9%)
X	16 (12%)	3.97E–01	33 (6.1%)	5.70E–11	2706 (15.2%)
4	2 (1.5%)	2.15E–01	1 (0.2%)	8.62E–01	116 (0.6%)

Table 2 AG-biased *vs de novo* genes in 500-kb windows

Chromosome	AG-biased genes absent		AG-biased genes present		AG *de novo* genes/AG-biased genes absent (%)	AG *de novo* genes/AG-biased genes present (%)	Fisher's exact test
	De novo genes absent	*De novo* genes present	*De novo* genes absent	*De novo* genes present			
All	104	19	84	73	5.45	46.50	2.60E−08
2L	9	1	19	19	10.00	50.00	3.10E−02
2R	19	4	15	13	17.39	46.43	3.90E−02
3L	19	1	26	11	5.00	29.73	4.10E−02
3R	22	6	15	22	21.43	59.46	2.60E−03
X	26	6	9	7	18.75	43.75	9.00E−02
4	1	1	0	1	50.00	100.00	1.00E+00
Y	8	0	0	0	NA	NA	NA

ancestral AG-biased gene, with enrichments ranging from about two to fivefold (Table 2). This pattern remained when we considered only ancestral AG-biased genes annotated as noncoding (Supplementary Table S6). Thus, there is strong evidence of chromosomal domains of correlated expression for *de novo* genes and ancestral AG-biased genes, consistent with previous studies of correlated patterns of gene expression along *Drosophila* chromosomes (Spellman and Rubin 2002; Boutanaev *et al.* 2002; Parisi *et al.* 2004).

At a smaller scale, we observed 14 clusters (six in intergenic regions, eight in intronic regions) of either two or three adjacent *de novo* genes exhibiting less than 10 kb between genes. To investigate whether these very tightly linked *de novo* genes tend to be expressed in a correlated manner across genotypes (*i.e.*, if the first gene in a cluster is expressed in a line it is likely the second gene is also expressed in that line) we compared the observed expression of clustered *de novo* genes to that expected under the null hypothesis that the *de novo* genes in each cluster are independently expressed. We found that for the mean cluster, 4.86 out of six lines exhibited consistent expression of genes in that cluster—that is, across lines either all genes in the cluster were expressed or all were unexpressed. This was substantially more consistent than the permuted data, which exhibited a mean consistency of 3.57 out of 6 (z-score 6.4, P 9.9 × e^{-11}). Focusing just on the correlation for expression (omitting nonexpression as an observation), we found that across all clusters, 0.86 out of six lines on average express all genes in a cluster, which was substantially higher than the correlation in the permuted data, 0.27 out of six lines (z-score 6.1, P 4.9 × e^{-10}). These data support the notion that the regulatory processes underlying the expression of *de novo* genes in the AG are spatially heterogeneous on multiple scales, ranging from a few kilobases to entire chromosome arms, and contribute to the physical distribution of candidate *de novo* genes across chromosome arms.

To compare the genomic distribution of AG-expressed *de novo* genes to testis-expressed *de novo* genes we repeated the windowing analysis for the testis-expressed *de novo* genes identified in Zhao *et al.* (2014; modified as described above) and strongly testis-biased ancestral genes (as described above; Supplementary Table S7). Because the number of testis-biased genes on the major chromosome arms is much higher than the number of AG-biased genes, 3049 *vs* 538, we compared the tissues using 100-kb windows so that for both tissues we had a sufficient number of windows in each of two categories (with one or more ancestral tissue-biased genes and with zero ancestral tissue-biased genes) for a reasonably powered analysis. At the whole genome level, we see a consistent pattern for both tissues—windows that contain *de novo* genes tend also to harbor ancestral genes exhibiting that tissue bias (Fisher's exact test; $P = 1.1 \times e^{-4}$ for AG; $P = 3.9 \times e^{-7}$ for testis). The degree of this enrichment at the 100 kb scale differed between tissues for the *X* chromosome, however, with a much smaller percentage of windows with testis-biased genes also containing testis *de novo* genes (6.4%) compared to the AG (18%), even though the total number of windows with testis-biased genes is substantially greater. Thus, it appears that whatever local regulatory phenomena are facilitating or driving *de novo* gene expression on the *X* chromosome, these effects are weaker for the testis than the AG, consistent with the greater underrepresentation of X-linked testis-expressed *de novo* genes.

Finally, to investigate the possible correlation between AG- and testis-expressed genes we first compared the number of ancestral testis-biased genes in windows with *vs* without ancestral AG-biased genes. We found that windows with AG-biased genes had significantly more testis-biased genes (mean = 3.2) than windows without AG-biased genes (mean = 2; Wilcoxon rank sum test; $P = 1.4 \times e^{-13}$), which supports the notion that our previous observation of correlated expression in the testis and AG (Cridland *et al.* 2020) could be explained in part by correlated chromosomal gene locations. Similarly, we found that windows with AG-expressed *de novo* genes contained more testis-expressed *de novo* genes (mean = 0.3) than windows without AG-expressed *de novo* genes (mean = 0.1; Wilcoxon rank sum test; $P = 1.6 \times e^{-6}$), suggesting that the *de novo* gene origination process is physically correlated across the genome for these two tissues.

Regulatory mechanism

We investigated the mechanisms of *de novo* gene expression using allelic imbalance experiments following Cridland *et al.* (2020). We had sufficient numbers of observations for only 13 genes. Of these, nine (69%) exhibit both *cis* and *trans* effects, three (23%) exhibit only *trans* effects, and one (7.6%) exhibits only *cis* effects (Table 3 and Supplementary Table S8). To compare these patterns to those of testis-expressed *de novo* genes, we reanalyzed 47 segregating testis-expressed *de novo* genes (Zhao *et al.* 2014) using the same pipeline and found that 30 (64%) exhibited both *cis*- and *trans*-effects, three (6%) exhibited only *trans*-effects, and 14 (30%) exhibited only *cis*-effects (Supplementary Table S8). While the proportion of *cis*-only testis-expressed *de novo* genes is smaller here than reported in Zhao *et al.* (2014) as a result of methodological differences, the important point is that for the candidate *de novo* genes subjected to the same allelic imbalance analysis, those expressed in the AG exhibit dramatically less *cis*-only regulation compared to those expressed in the testis (Fisher's Exact test; $P = 1.6 \times e^{-09}$).

Table 3 Regulation of *de novo* genes

Regulatory mechanism	AG	Testis
Cis	1	14
Cis and *trans*	9	30
Trans	3	3

Expression in different cell types

We used AG single-nucleus RNA-seq data from RAL 517 (Majane *et al.* 2021) to investigate expression of *de novo* genes in the three major cell types of our dissected bulk tissue: main cells, secondary cells, and ejaculatory duct cells. Of the 13 candidate genes called as expressed (TPM > 1) in RAL 517, 10 were also called as expressed in single-nucleus data. This independent validation of our *de novo* gene candidates in a different experiment using different technology suggests that our approach for identifying such genes is robust. Imposing a much lower cutoff of TPM > 0.1 to categorize a *de novo* gene as expressed in bulk transcriptome data from RAL 517 results in 86 expressed genes, of which 39 were also expressed in the single-nucleus data, strongly suggesting that the TPM > 1 criterion is quite conservative. The conservative nature of *de novo* gene TPM criterion would be especially pronounced for genes that tend to be expressed at a higher level in secondary or ejaculatory duct cells, as those cells are considerably less abundant than main cells in bulk tissue used in the experiment (Majane *et al.* 2021). It also suggests that we may have underestimated the proportion of RAL strains expressing a given *de novo* gene candidate.

Majane *et al.* (2021) used the single-nucleus data to identify marker *de novo* genes—those that exhibited biased expression across the three major cell types. Of the 43 total *de novo* gene candidates identified here that were defined as expressed in the single-cell data from RAL 517 (Majane *et al.* 2021), five were classified as marker genes, all of which showed ejaculatory duct biased expression. This represents a significant enrichment of *de novo* genes expressed at a high level in this cell type relative to the other two types (hypergeometric, $P < 0.0005$). Interestingly, four of the five ejaculatory duct marker *de novo* genes are nonsingletons, which represents a significant enrichment of higher frequency genes (hypergeometric, $P = 0.016$), though the small sample size precludes strong conclusions about this pattern. Majane *et al.* (2021) reported that ejaculatory duct-cell transcriptomes tend to evolve more quickly than main cell and secondary cell transcriptomes. Our finding that expression-biased *de novo* in the AG tends to be ejaculatory duct-biased is consistent with the notion that this cell type may be prone to higher rates of transcriptome turnover compared to main and secondary cells.

Discussion

Our comparisons of *de novo* gene candidates expressed in the AG and testis revealed several differences that illuminate variation in the processes underlying their origin and evolution. While there were minor differences between the starting material and analyses used in our investigation here of the AG and our previous investigation of the testis, such methodological differences cannot plausibly explain the stark differences between the two tissues in the abundance, chromosomal distributions, regulatory mechanisms, and population frequencies of *de novo* genes.

The AG expresses many fewer intergenic *de novo* genes than the testis, often at lower levels. Moreover, most AG-expressed *de novo* genes are expressed in only one genotype, and only two were fixed in our sample. The testis, in contrast, expresses many more intermediate and high frequency/fixed genes (Zhao *et al.* 2014), most of which are germline expressed (Witt *et al.* 2019). Thus, at the population level, the contribution of *de novo* genes to the somatic male reproductive tissue transcriptome appears much smaller than their contribution to the germline transcriptome.

There are also apparent differences between the two tissues in the way *de novo* gene candidates are regulated. While strictly *cis*-acting variation is common for testis-expressed *de novo* genes, the AG tends to exhibit more complex regulatory variation, including a more substantial *trans*-acting component. The selective spread or removal of a *de novo* gene would be more efficient if its expression resulted from a novel, tightly linked *cis*-regulatory element that co-opted existing *trans*-acting regulatory factors, as appears to be the case for testis (Zhao *et al.* 2014), compared to a situation where the underlying genetics of novel expression is more complex, as appears to be the case for the AG. This difference may contribute to the lower fixation rate for AG- *vs* testis-biased *de novo* genes and the relative paucity of singleton testis-expressed *de novo* genes. Larger samples from both tissues would shed light on their potential differences in the relative proportion of rarely expressed *de novo* genes.

The physical distribution of *de novo* gene candidates expressed in the AG or testis show some similarities, but also important differences. Both types of *de novo* genes tend to be found in chromosomal regions harboring ancestral genes exhibiting the same tissue-biased expression, and the locations of testis- and AG-expressed ancestral and *de novo* genes are correlated. Furthermore, very tightly linked AG-expressed *de novo* genes, which show no evidence of origination by duplication, show strongly correlated expression patterns across genotypes, also supporting a physically correlated origination process. The fact that ancient genes and young *de novo* genes show correlated physical distributions for both testis and AG suggests that regulatory phenomena underlying the origination process and/or the selective retention of such genes during evolution play a role in generating the distribution. However, this effect is diminished for X-linked testis-expressed *de novo* genes, which could contribute to the lower (though not significantly so) *X*/*A* ratio for testis- than for AG-expressed *de novo* genes. Both testis- and AG-expressed *de novo* genes are underrepresented on the *X* chromosome, as expected for strongly male-biased genes (Sturgill *et al.* 2007; Meisel *et al.* 2012). However, relative to the *X*/*A* distributions for their corresponding ancestral tissue-biased genes, AG-expressed *de novo* genes are roughly twice as likely to be *X*-linked, while testis-expressed *de novo* genes are roughly twice as likely to be autosomal. The disparate deviations of young *de novo* gene *X*/*A* distributions from that of their ancestral counterparts could result from effects of heterogeneous origin processes and/or heterogeneous selective processes. For example, *X*-chromosome inactivation in the male germline (Lifschytz and Lindsley 1972; Kemkemer *et al.* 2011; Landeen *et al.* 2016; Mahadevaraju *et al.* 2021) could lead to reduced birth-rates for *X*-linked testis-expressed *de novo* genes. Alternatively, assuming the very strong autosomal enrichment of ancestral AG-biased genes is shaped by selection, the relative enrichment of *X*-linkage for AG-expressed *de novo* genes is consistent with a weaker selective effect on their chromosomal distribution, consistent with their very young age. The increased proportion of noncanonical splice junctions in these genes also supports the view that for at least some, their properties have not been optimized by natural selection. The observation that ancestral and *de novo* testis- and AG-biased genes reside in shared chromosomal domains of expression yet differ in the importance of *cis*-acting regulatory variation is suggestive of underlying hetergeneous processes, but the relative contributions of differences in origination processes *vs* selective effects are difficult to discern with existing data.

The general question of how frequently *de novo* genes originate as coding *vs* noncoding is of great importance, but currently unresolved (*e.g.*, Ruiz-Orera and Alba 2019), and cannot be addressed using bioinformatic approaches that start with the premise that these genes are coding. Our computational analyses suggest that most young AG-expressed *de novo* genes are noncoding, though direct inferences from proteomic or ribo-profiling data would be required to put this conclusion on firmer ground. Similarly, the testis-expressed *de novo* candidates discussed here are also predicted by CPAT to be noncoding. Even if the small ORFs associated with these AG-expressed transcripts are translated, the vast majority of potential predicted proteins do not carry signal sequences. Thus, if several of the novelties described here have functions, they are unlikely to be directly related to processes requiring conventional protein secretion, though it remains possible that noncoding RNAs derived from *de novo* genes are transferred to females during mating (Bono *et al.* 2011; Ahmed-Braimah *et al.* 2021). Nevertheless, it follows from their properties that if AG-expressed *de novo* gene products have biological functions, they are likely biased toward those occurring inside the gland.

The complex regulatory variation influencing expression of *de novo* genes in the AG, the small sample of genotypes investigated here, the fact that many candidate genes are expressed in only one or a few genotypes, and the likelihood that most candidates are predicted to be noncoding, all conspire to compromise population genetic investigation into the possible influence of selection on these sequences. For example, approaches that seek evidence for protein functional constraint would have no value for noncoding genes. Moreover, the observation that the majority of candidates, which are predicted to be noncoding, are not associated with homologous ORFs in *D. simulans* (not shown) would not speak to the possible existence of a *D. melanogaster*-specific ORF, and in any case would also be entirely consistent with the hypothesis that a gene was *D. melanogaster* specific and noncoding. Similarly, the complex regulation and expression in one or a few genotypes for most candidates make it challenging to seek evidence of hitchhiking effects. Under the premise that *trans*-acting variants are likely to be more strongly deleterious than *cis*-acting variants, the influence of *trans*-acting variation on *de novo* gene expression in the AG would be consistent with the apparently low frequency of these genes in the population. However, without information on the population genetics of the *trans*-acting variants themselves, this idea cannot be evaluated. Thus, it seems to us that the question of the influence of selection on these candidate genes, while vitally important, cannot be adequately addressed with existing data.

Nevertheless, the relationships between gene frequency, size and expression may provide some information about evolutionary mechanisms. If AG-expressed *de novo* genes were on average deleterious, either due to the cost of transcription and/or translation, or because of deleterious interactions of their products (RNAs or proteins) in the cell, then we might expect singletons to be longer and expressed at a higher level than nonsingletons (Zhao *et al.* 2014), as deleterious genes should be overrepresented in the singleton class. Instead, we observe that singletons are shorter and expressed at lower levels than nonsingletons. This pattern provides no support for the idea that the genes described here are, on average, deleterious or strictly neutral (in which case there would be no expectation of heterogeneity of their attributes across frequency classes). The observed differences between singleton and nonsingleton genes, which is similar to that observed for testis-expressed *de novo* genes, would be consistent with an influence of positive selection (Zhao *et al.* 2014). However, the observation that very few AG-expressed *de novo* genes occur at high frequency does not support the simple hypothesis that simple directional selection plays an important role in their dynamics, as in this case, several would have fixed. Moreover, their overrepresentation on the X chromosome relative to ancestral AG-biased genes does not support the idea that directional selection is playing an important role. This does not imply, however, that these genes are uninfluenced by positive selection, as various forms of balancing selection acting on AG function could generate substantial polymorphism but relatively low fixation rates (*e.g.*, Hughes 1997; Brisson 2018). That the molecular biology of the AG also appears to be highly variable in terms of amino acid polymorphism (Coulthart and Singh 1987; Begun *et al.* 2001), presence/absence of intact gene copies (Begun and Lindfors 1995), and gene expression or lack thereof in the organ (Cridland *et al.* 2020), is consistent with this possibility. Testing models of drift or selection acting on novel AG variation would be greatly facilitated by detailed information on the genetic variants underlying the gain of genes (Zhao *et al.* 2014) or gene expression (Cridland *et al.* 2020).

Data availability

Raw sequence data for all experiments are available from SRA, PRJNA575046, PRJNA210329. Supplemental Material available online at figshare: https://doi.org/10.25386/genetics.16934209.

Acknowledgments

We thank the anonymous reviewers for their comments, and A. Kopp and L. Blair for sharing *D. yakuba* sequence data. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Funding

This work was supported by the National Institute of Health, NIH R01GM110258 and NIH R35GM134930 awarded to D.J.B.

Conflicts of interest

The authors declare that there is no conflict of interest.

Literature cited

Ahmed-Braimah YH, Wolfner MF, Clark AG. 2021. Differences in transcriptional responses between conspecific and heterospecific matings in *Drosophila*. Mol Biol Evol. 38:986–989.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Armenteros JJA, Tsirigos KD, Sonderby CK, Petersen TN, Winther O, *et al.* 2019. SignalP5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol. 37:420–423.

Begun DJ, Whitley P, Todd BL, Waldrip-Dail HM, Clark AG. 2001. Molecular population genetics of male accessory gland proteins in *Drosophila*. Genetics. 156:1879–1888.

Begun DJ, Lindfors HA. 2005. Rapid evolution of genomic *Acp* complement in the *melanogaster* subgroup of *Drosophila*. Mol Biol Evol. 22: 2010–2021.

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

Begun DJ, Lindfors HA, Kern AD, Jones CD. 2007. Evidence for *de novo* evolution of testis-expressed genes in the *Drosophila yakuba/Drosophila erecta* clade. Genetics. 176:1131–1137.

Belote JM, Zhong L. 2009. Duplicated proteasome subunit genes in *Drosophila* and their roles in spermatogenesis. Heredity (Edinb). 103:23–31.

Betrán E, Thornton K, Long M. 2002. Retroposed new genes out of the *X* in *Drosophila*. Genome Res. 12:1854–1959.

Bono JM, Matzkin LM, Kelleher E, Markow TA. 2011. Postmating transcriptional changes in reproductive tract of con- and heterospecifically mated *Drosophila mojavensis* females. Proc Natl Acad Sci U S A. 108:7878–7883.

Boutanaev AM, Kalmykova AI, Shevelyov YY, Nurminsky DI. 2002. Large clusters of co-expressed genes in the *Drosophila* genome. Nature. 420:666–669.

Brisson D. 2018. Negative frequency-dependent selection is frequently confounding. Front Ecol Evol. 6:10 doi:10.3389/fevo.2018.00010.

Cai J, Zhao R, Jiang H, Wang W. 2008. *De novo* origination of a new protein-coding gene in *Saccharomyces cerevisiae*. Genetics. 179: 487–496.

Casola C. 2018. From *de novo* to “de nono”: he majority of novel protein-coding genes identified with phylostratigraphy are old genes or recent duplicates. Genome Biol Evol. 10:2906–2918.

Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, *et al.* 2012. Proto-genes and *de novo* gene birth. Nature. 487:370–374.

Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, *et al.* 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature. 450:203–218.

Coulthart MB, Singh RS. 1987. Differing amounts of genetic polymorphism in testes and male accessory glands of *Drosophila melanogaster* and *Drosophila simulans*. Biochem Genet. 26:153–164.

Coulthart MB, Singh RS. 1988. High level of divergence of male reproductive tract proteins between *Drosophila melanogaster* and its sibling species, *D. simulans*. Mol Biol Evol. 5:182–191.

Crosby MA, Gramates LS, Dos Santos G, Matthews BB, St Pierre SE, *et al.*; FlyBase Consortium. 2015. Gene model annotations for *Drosophila melanogaster*: The rule-benders. G3 (Bethesda). 5: 1737–1749.

Cridland JM, Majane AJ, Sheehy HK, Begun DJ. 2020. Polymorphism and divergence of novel gene expression patterns in *Drosophila melanogaster*. Genetics. 216:79–93.

Findlay GD, Yi X, MacCoss MJ, Swanson WJ. 2008. Proteomics reveals novel *Drosophila* seminal proteins transferred during mating. PLoS Biol. 6:e178.

Findlay GD, MacCoss MJ, Swanson WJ. 2009. Proteomic discovery of previously unannotated, rapidly evolving seminal fluid genes in *Drosophila*. Genome Res. 19:886–896.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, *et al.* 2011. Trinity: reconstructing a full length transcriptome without a genome from RNA-seq data. Nat Biotechnol. 29:644–652.

Haerty W, Jagadeeshan S, Kulathinal RJ, Wong A, Ram KR, *et al.* 2007. Evolution in the fast lane: rapidly evolving sex-related genes in *Drosophila*. Genetics. 177:1321–1335.

Heinen TJAJ, Staubach F, Haming D, Tautz D. 2009. Emergence of a new gene from an intergenic region. Curr Biol. 19:1527–1531.

Hughes KA. 1997. Quantitative genetics of sperm precedence in *Drosophila melanogaster*. Genetics. 145:139–151.

Kemkemer C, Hense W, Parsch J. 2011. Fine-scale analysis of X-chromosome inactivation in the male germline of *Drosophila melanogaster*. Mol Biol E. 28:1561–1563.

Knowles DG, McLysaght A. 2009. Recent *de novo* origin of human protein-coding genes. Genome Res. 19:1752–1759.

Landeen EL, Muirhead CA, Wright L, Meiklejohn CD, Presgraves DC. 2016. Sex chromosome-wide transcriptional suppression and compensatory *cis*-regulatory evolution mediate gene expression in the *Drosophila* male germline. PLoS Biol. 14:e1002499.

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

Lee YCG, Chang H-H. 2013. The evolution and functional significance of nested gene structures in *Drosophila melanogaster*. Genome Biol Evol. 5:1978–1985.

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

Li D, Dong Y, Jiang Y, Jiang H, Cai J, *et al.* 2010. A *de novo* originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. Cell Res. 20: 408–420.

Lifschytz E, Lindsley DL. 1972. The role of X-chromosome inactivation during spermatogenesis. Proc Natl Acad Sci U S A. 69: 182–186.

Long M, Langley CH. 1993. Natural selection and the origin of Jingwei, a chimeric processed functional gene in *Drosophila*. Science. 260:91–95.

Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, *et al.* 2012. The *Drosophila melanogaster* genetic reference panel. Nature. 482:173–178.

Mahadevaraju S, Fear JM, Akeju M, Galletta BJ, Pinheiro MMLS, *et al.* 2021. Dynamic sex chromosome expression in *Drosophila* male germ cells. Nat Commun. 12:1–16.

Majane AC, Cridland JM, Begun DJ. 2021. Single-nucleus transcriptomes reveal function and evolutionary properties of cell types in the *Drosophila* accessory gland. Genetics. iyab213. doi: 10.1093/genetics/iyab213.

McManus JC, Coolon JD, Duff MO, Eipper-Mains J, Graveley BR, *et al.* 2010. Regulatory divergence in *Drosophila* revealed by mRNA-seq. Genome Res. 20:816–825.

Meiklejohn CD, Parsch J, Ranz JM, Hartl DL. 2003. Rapid evolution of male-biased gene expression in *Drosophila*. Proc Natl Acad Sci U S A. 100:9894–9899.

Meisel RP, Malone JH, Clark AG. 2012. Disentangling the relationship between sex-biased expression and X-linkage. Genome Res. 22: 1255–1265.

Mikhaylova LM, Nguyen K, Nurminsky DI. 2008. Analysis of the *Drosophila melanogaster* testes transcriptome reveals coordinate regulation of paralogous genes. Genetics. 179:305–315.

Murphy DN, McLysaght A. 2012. *De novo* origin of protein-coding genes in murine rodents. PLoS One. 7:e48650.

Neme R, Tautz D. 2013. Phylogenetic patterns of emergence of new genes support a model of frequent *de novo* evolution. BMC Genomics. 14:117.

Nielsen H, Engelbrecht J, Brunak S, Von Heijne G. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. 10:1–6.

Palmieri N, Kosiol C, Schlötterer C. 2014. The life cycle of *Drosophila* orphan genes. eLife. 2014;3:e01311. doi:10.7554/eLife.01311.001.

Parisi M, Nuttall R, Edwards P, Minor J, Naiman D, *et al.* 2004. A survey of ovary-, testis-, and soma-biased gene expression in the *Drosophila melanogaster* adult. Genome Biol. 5:R40.

Ruiz-Orera J, Alba MM. 2019. Translation of small open reading frames: roles in regulation and evolutionary innovation. Trends Genet. 35:186–198.

Sorourian M, Kunte MM, Domingues S, Gallach M, Özdil F, *et al.* 2014. Relocation facilitates the acquisition of short *cis*-regulatory regions that drive the expression of retrogenes during spermatogenesis in *Drosophila*. Mol Biol E. 31:2170–2180.

Spellman PT, Rubin GM. 2002. Evidence for large domains of similarly expressed genes in the *Drosophila* genome. *J* Biol. 1:5.

Sturgill D, Zhang Y, Parisi M, Oliver B. 2007. Demasculinization of X chromosomes in the *Drosophila* genus. Nature. 450:238–237.

Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF. 2001. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. Proc Natl Acad Sci U S A. 98: 7375–7379.

Thurmond J, Goodman JL, Strelets VB, Attrill H, Gramates LS, *et al.*; FlyBase Consortium. 2019. FlyBase 2.0: the next generation. Nucleic Acids Res. 47:D759–D765.

Van Oss SB, Carvunis A-R. 2019. *De novo* gene birth. PLoS Genet. 15: e1008160.

Vakirlis N, Hebert AS, Opulente DA, Achaz G, Hittinger CT, *et al.* 2018. A molecular portrait of *de novo* genes in yeasts. Mol Biol Evol. 35: 631–664.

Wagstaff BJ, Begun DJ. 2005. Molecular population genetics of accessory gland protein genes and testis-expressed genes in *Drosophila mojavensis* and *D. arizonae*. Genetics. 171:1083–1101.

Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, *et al.* 2013. CPAT: Coding-Potential Assessment Tool using an alignment free logistic regression model. Nucleic Acids Res. 41:e74.

Wigby S, Brown NC, Allen SE, Misra S, Sitnik JL, *et al.* 2020. The *Drosophila* seminal proteome and its role in postcopulatory sexual selection. Philos Trans R Soc B. 7;375:20200072. doi: 10.1098/rstb.2020.0072.

Wilson C, Leiblich A, Goberdhan DCI, Hamdy F. 2017. The *Drosophila* accessory gland as a model for prostate cancer and other pathologies. In L. Pick, editor. Fly Models of Human Disease. Cambridge: Academic Press. p. 339–375.

Witt E, Benjamin S, Svetec N, Zhao L. 2019. Testis single-cell RNA-seq reveals the dynamics of *de novo* gene transcription and germline mutation bias in *Drosophila*. eLife. 8:e47138.doi: 10.7554/eLife.47138.

Yanai I, Benjamin H, Shmoish M, Chalifa-Caspi V, Shklar M, *et al.* 2005. Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics. 21:650–659.

Yang H, Jaime M, Polihronakis M, Kanegawa K, Markow T, *et al.* 2018. Re-annotation of eight *Drosophila* genomes. Life Sci Alliance. 1: e201800156.

Zhang L, Ren Y, Yang T, Li G, Chen J, *et al.* 2019. Rapid evolution of protein diversity by *de novo* origination in *Oryza*. Nat Ecol Evol. 3: 679–690.

Zhang Y, Sturgill D, Parisi M, Kumar S, Oliver B. 2007. Constraint and turnover in sex-biased gene expression in the genus *Drosophila*. Nature. 450:233–237.

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

Zhou Q, Zhang G, Zhang Y, Xu S, Zhao R, *et al.* 2008. On the origin of new genes in *Drosophila*. Genome Res. 18:1446–1455.

Communicating editor: M. Hahn