Title

New insights into the genetic, epigenetic, and genomic bases of de novo gene expression in Drosophila melanogaster accessory glands

Author

Blair, Logan Kevin

Publication Date

2022

Peer reviewed|Thesis/dissertation

New insights into the genetic, epigenetic, and genomic bases of *de novo* gene expression in *Drosophila melanogaster* accessory glands

By

LOGAN KEVIN BLAIR
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Integrative Genetics and Genomics

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Artyom Kopp, Chair

David Begun

Neelima Sinha

Committee in Charge

2022

TABLE OF CONTENTS

Project Abstract

Species differ in what genes they have, therefore understanding the processes underlying gene origin is of great importance to the field of evolutionary genetics. There are interesting cases where gene origin can be traced back to noncoding DNA instead of previously genic sequence. This unique mechanism of generating new genes has been coined a "*de novo"* origin. Evidence suggests that *de novo* genes are typically tissue-restricted, by several metrics less genic than older genes, and undergo rapid turnover over evolutionary time. Yet how they initially become expressed a mystery, and we know especially little about the youngest *de novo* genes. In these two studies, I investigate how very young *de novo* genes become transcribed and may spread across the population. I use male accessory gland tissue from the model organism *Drospophila melanogaster*. In chapter one, I show that most *de novo* genes are very rare in the population. Yet I find assessment of *de novo* origin is complicated by the degree to which some unannotated genes persist over evolutionary time. I also show *de novo* genes do not frequently reuse existing regulatory sites near older genes, but there is evidence for convergent evolution in their regulatory sequences. In chapter two, I show *de novo* genes have accessible chromatin that is not present in the ancestral state, further suggesting that *de novo* genes draw their expression from novel regulatory regions. And yet, the expression basis for few *de novo* genes could be mapped to nearby regions, suggesting a complex genetic basis or that additional changes are also necessary for their expression. Together, these studies greatly expand on our knowledge of the first stages of *de novo* gene origin.

Acknowledgements

I owe a great debt to those in my life. Graduate school can be a wonderful experience, but it can also be humbling and challenging. No person exists in isolation, so accomplishments are shared.

First, there are many people to thank in the Kopp lab. I would like to thank my advisor Artyom Kopp for his support and guidance. His creativity and intellect were always an inspiration. I am certainly a better scientist because of his mentorship. Our lab manager Olga Barimina, herself, is an encyclopedia of lab technique. Her cheer always made lab work more enjoyable (and was an effective counterbalance to my pre-coffee surliness). I would like to also thank my graduate student, postdoc, and project scientist peers: Ammon Thompson, Emily Delany, Kohtaro Tanaka, Ben Hopkins, Yuichi Fukutomi, Gavin Rice, David Leucke, Judy Wexler, Zach Farrow, Giovani Hanna, and Yige Luo. This experience would not have been the same without your advice, mentorship, and commiseration. I would like to thank the undergraduates I worked with for pushing me to be a better teacher and communicator.

Next, I would like to thank my committee members Neelima Sinha and David Begun for their service and edits. Work in the Begun lab paved the way for this dissertation project. In particular, I would like to thank Li Zhao and Julie Cridland for their advisement and knowledge of *de novo* gene evolution. Julie was a wonderful teacher and was responsible for much of the bioinformatic analysis.

Lastly, I would like to thank my friends and family for shaping me and my life. I would not be where I am today without their love and support. In particular, my wife, Shannon, was an integral part of my graduate school experience. Her presence is forever associated with its best parts, because she was directly responsible for them.

Introduction

Many new genes come from sequence that, in some previous form, also used to be genic. It was initially seen as a surprise when some genes were discovered to originate "*de novo*" from sequences that did not used to be genes. Despite this, both the number of newly annotated *de novo* genes and number of species where they have been found continue to rise.

After the novelty of their existence began to wane, attention shifted to just exactly how nongenic sequence can be converted into genes and why the early stages of their transcription escape selection. Past research has indicated that some *de novo* genes may rapidly be gained and then subsequently lost, which could indicate a selective disadvantage of these sequences. Some *de novo* genes may also contain some genic traits, including ORFs and signal sequences, but fewer do compared to older genes. Whether these genes encode function has been challenging to assess. It seems likely that a high number of newly transcribed sequences may be required before one gains traction and spreads.

There is much we do not understand about how *de novo* genes become transcribed. Since these are new genes, it seems logical to conclude that they may use new regulatory sequences. Yet elsewhere in evolution there are many examples of the "cooption" of existing elements to use in new contexts. May it be possible that *de novo* genes reuse certain aspects of the existing regulatory structure? This question leads to several testable predictions. Since regulatory elements are enriched near genes, and since many *de novo* genes are tissue-restricted, we may expect to see more *de novo* genes evolving near older tissue-specific genes. Likewise, regulatory cooption would suggest that the epigenetic signatures of regulatory sequence would predate the expression of *de novo* genes. The evolutionary scale at which these questions are tested is a very important consideration. Ideally, few differences will separate expressed and non-expressed alleles so that

the causative change is more apparent. Therefore, I choose to examine very young *de novo* genes that are still segregating in the population.

In chapter I, I use RNA-seq in a core set of 29 different *Drosophila melanogaster* genotypes to identify *de novo* genes that are expressed in accessory gland tissue. Within this extensive sampling of genotypes, I better show that most potential *de novo* genes are at very low frequency in the population. I also present similar unannotated gene sequences that, while similar to *D. melanogaster de novo* genes in some respects, are expressed in multiple species. These sequences are also hard to sample outside of *D. melanogaster*, raising questions about the potential false-positive error rate for *de novo* gene origin. I also present a few other key findings regarding the locations where candidate species-specific *de novo* genes evolve: few *de novo* genes use bidirectional promoters, and *de novo* genes are less close to other tissue-specific genes than they are to themselves. These points do not support that *de novo* genes reuse existing regulatory sites. However, I also find similarities between *de novo* genes and older tissue-specific genes regarding the transcription factor binding sites upstream of their transcription start sites. This may suggest some convergent evolution in how *de novo* genes are transcribed.

In chapter II, I investigate whether *de novo* genes reuse ancestral regulatory regions and, if so, whether there are genetic changes that occur in these regions to explain why. I use ATAC-seq in five *D. melanogaster* genotypes, as well as two related outgroup species, and show open chromatin is often a derived state unique to the genotypes that express *de novo* genes. Next, I use *cis* association mapping to identify the genetic basis of several *de novo* genes. Few *de novo* genes have variants that are highly associated with their expression. Among three genes that did have significantly-associated variants, all had at least one SNP located within the boundaries of differentially-accessible chromatin peaks. These results suggest a potential for enhancer evolution

to spark the transcription of *de novo* genes, though for the typical *de novo* gene it may be challenging to determine the contribution of individual genetic variants.

Chapter I: New insight into the dynamics of *de novo* gene origin from increased population sampling of *Drosophila melanogaster* accessory glands

Logan Blair, Julie Cridland, Yige Luo, David Begun, and Artyom Kopp

Abstract

The evolution of genes *de novo* from nongenic sequence has become more appreciated as a mechanism of gene origin. Most studies have used few genotypes or distantly related species to infer *de novo* genes, which may lead to undersampling short-lived or uncommon alleles. Here, we examine the messy beginnings of *de novo* genes as they arise in the population. We use a comparative transcriptomics approach with an expansive pool of tissue and genotype-specific sequences to capture the characteristics of rarer alleles. To do so, we used RNA-seq in accessory gland tissue from a core set of 29 *D. melanogaster* genotypes, as well as additional genotypes from two of its close relatives. We show the majority of unannotated genes are lowly transcribed and lack fixed expression within species. Some rare alleles can be found in multiple species, which is hard to reconcile with models of a discrete, single new allele origin and/or fast rates of loss. As a technical consideration, the rarity of many unannotated genes may inflate the *de novo* classification rate, as evidence of their expression in outgroups is likely to be missed. Next, we explored whether the location of these genes may suggest a mechanism for their origin. Surprisingly, we found within tissue-specific genes, gene age was inversely correlated with proximity to other tissue-specific genes. This does not support the idea that the youngest *de novo* genes simply reutilize the tissue-specific regulatory elements near annotated genes. Despite this, younger *de novo* genes showed an enrichment for some tissue-specific binding motifs, which could indicate convergent evolution of regulatory sequence elsewhere in intergenic space.

Introduction

Differences in gene complement contribute to variation between species. With widespread sequencing, the basis for some new genes can be traced to lineage-specific gains in the expression of DNA sequences not previously known to be transcribed (Begun et al., 2006). It was seen as a surprise when genes were found that did not originate from other genes. For many years the paradigm was that most came from duplications (Ohno, 2013) or horizontal gene transfer (Soucy et al., 2015). But since the first descriptions of "*de novo"* gene origins, *de novo* genes have been documented in a diverse array of taxa across the tree of life (Neil et al., 2009, Ruiz-Orera et al., 2015, Li et al., 2016, Zhao et al., 2014).

The study of *de novo* gene origin poses tricky methodological challenges. Calling *de novo* genes requires the absence of signal in outgroups. Yet the "absence of a negative" is not as definitive a result as the "presence of a positive". The origin status for several potential candidates has later been revised after incorporating new sequence information or by using different methods (Casola, 2018). The nature of gene expression itself is a complicating factor. Calling *de novo* gene origin relies on inferring discrete character states (the presence of a gene) from a continuous variable (gene expression). Often these two are in agreement. In macroevolutionary comparisons between distant species, the many fixed differences may lend themselves to an "on" or "off" binary state applied to whole species. But less is known about the initial stages *de novo* genes, when a nongenic region is first transcribed. If genes are only expressed 1) under very specific conditions, 2) in small amounts, or 3) variably between genotypes (due to null alleles segregating in the population), it may be more likely the evidence of their transcription is missed. And yet, evidence suggests that *de novo* genes are predisposed towards all three of these complications. There is a need for more study on expression variation of very young *de novo* genes in light of these issues.

The quantity of *de novo* genes is influenced by the rate of their birth via novel transcription and the rate of their death via drift and selection (Oss and Carvunis, 2019). A holistic understanding of *de novo* origin requires understanding how features of the genome may impact both. Genomic structures hypothesized to facilitate transcription include bidirectional promoters (Blevins et al., 2021), open chromatin (Werner et al., 2018), and preexisting regulatory elements. It is possible that *de novo* genes tend to originate near other genes, where these regulatory features will be more common. In terms of selection, expression patterns of extant *de novo* genes may reflect a survivorship bias towards where they are less likely to be deleterious. Studies indicate most extant *de novo* genes are expressed in low abundance and are tissue-restricted. Both these restrictions may reduce the exposure of *de novo* genes, blunting the fitness consequence of suboptimal sequence. Another fitness consideration is that transcriptionally permissive, noisy regions may have low-level selection against maladaptive open reading frames already, suggesting some intergenic sequence may be more fit than sequence that is completely random (Werner et al., 2018). Taking all these points into consideration, tissue-specific regulatory elements are an attractive starting point to facilitate *de novo* gene origin. Such structures may both increase the expression of the surrounding area and convey the tissue-restrictions observed in many young genes.

In this report, we use transcriptomes of *Drosophila melanogaster* accessory glands (AGs), a model tissue to study *de novo* gene evolution. Hypothetically more *de novo* genes could be found using whole-animal tissue, but in practice the low abundance of *de novo* genes and their tissue-specificity mean that they may not be sampled outside of single-tissue preps. The AGs make and secrete seminal fluid, which can elicit post-mating responses when transferred to females (Wigby et al., 2020). There are several reasons that they are a useful model for *de novo* genes. The presence of AG-specific *de novo* genes has already been found, being one of the first tissues where *de novo*

genes were described (Begun et al., 2006). AGs also have the advantage of being a simple tissue, consisting of two secretory cells types (though within the primary cell type there is substructure) and a muscle sheath (Majane et al., 2022). Past work has described some regulatory properties of AG *de novo* genes and how they may differ from other *D. melanogaster de novo* genes. In AGs, *de novo* genes tended to be shorter and contain fewer exons in comparison to the testis (Cridland et al., 2022). The AG *de novo* genes also exhibited a *cis*-regulatory basis to a lesser degree than the testis (Zhao et al., 2014), suggesting local sequence change may interact with natural variation in *trans*-regulatory factors. There was a positional effect of AG *de novo* genes being clustered around older AG-specific genes. This could suggest that the older gene's tissue-specific regulatory elements may facilitate *de novo* gene evolution.

In this study, we explored two questions to better our understanding of *de novo* gene evolution. Our first question is one of population biology, with methodological implications for the assessment of *de novo* genes. We ask to what extent does *de novo* gene expression vary in the very early stages of gene birth, and whether that may affect how *de novo* genes are sampled. To answer this, we sequenced 31 *D. melanogaster* AG-specific transcriptomes to explore intraspecific polymorphism in *D. melanogaster de novo* gene expression. While we use weak minimum expression filters and have no requirement for genes to have ORFs, our outgroup filtering procedures are strict and include an expanded pool of outgroup sequences. To determine if polymorphism is constrained to individual species, we also used both publicly available sequences of related species and our own RNAseq libraries to explore how our candidate *de novo* genes vary from other non *D. melanogaster*-specific unannotated sequences. Second, we explored the mechanism of *de novo* gene origin by asking to what extent preexisting regulatory structures increase the likelihood of finding nearby *de novo* genes. We tested whether *de novo* genes are more

likely to evolve near older tissue-specific genes, in a bidirectional promoter orientation, and near motifs for the binding sites of tissue-specific regulatory elements.

Materials and Methods

RNA preparation

RNA libraries used for unannotated transcript discovery were constructed separately from each of 29 lines of the Drosophila genetic resource panel (DGRP; Mackay et al., 2012). The DGRP consists of 205 sequenced isofemale lines derived from a single *D. melanogaster* population in Raleigh, NC. To assess the presence of unannotated transcripts in other *D. melanogaster* populations and to better establish patterns of outgroup expression, RNA was also collected from two non-DGRP *D. melanogaster* lines, two *D. simulans* lines, and one *D. yakuba* line (see Table 1 for lines used in this study, including previously published data). For each line, RNA was extracted from pooled accessory glands of 30 unmated, two day old males using TRIZOL (Invitrogen) followed by on column cleanup with DNAse digestion (Zymo). RNA libraries were prepared with Illumina Truseq stranded mRNA kit (Illumina), which uses polyT beads to capture polyadenylated RNA sequences. Libraries where sequenced using 150 bp paired-end reads on an Illumina Hiseq4000.

Species	Genotypes
D melanogaster DGRP	153, 217, 229, 287, 304, 320, 338, 352, 357, 359, 360, 370, 380, 399, 517, 530, 563, 630, 703, 761, 805, 812, 822, 850, 85, 88, 900, 911, 93
D melanogaster non Raleigh	ED10, iso-1
D simulans	w501, 116, Lara10
D yakuba	tai18

Table 1: Lines used in this study. Establishment of lines from DGRP described in Mackay 2012. Parent population for line Lara10 was collected in Homestead, FL, and was sib-mated for 10 generations in Begun lab (Zhao et al., 2014). We also used the *D. yakuba* reference sequence strain, Tai18E2 (Begun et al., 2007).

Identification and quantification of de novo genes

To identify *D. melanogaster*-specific *de novo* genes expressed in the accessory glands, 29 DGRP transcriptome sequences were aligned using STAR v2.6.1d to the Dm6 genome assembly. A mean of 28,213,740 reads were mapped per sample, with a mean rate of 90.09% (Table S1). Reference-guided transcript assembly was performed for individual lines using stringtie version 1.3.4d. Next, assemblies from individual lines were merged into a unified file using TACO v0.7.3 to create a preliminary set of transcripts expressed in accessory glands of DGRP lines. Two other *D. melanogaster* genotypes were sequenced from other populations (Table 1), but these were not included in the *de novo* gene assembly.

We then applied a series of filters to the assembled transcripts to identify the candidates most likely to have evolved *de novo* in *D. melanogaster* following the split from *D. simulans.* This procedure largely followed Cridland 2021, though with our focus on the effect of intergenic regulatory elements, we included no minimum distance cutoff between unannotated and annotated genes but excluded intronic sequences. See Figure 1 for description of filters used and overall workflow. We defined *de novo* genes as transcripts that: were longer than 200 bp; were expressed >1 TPM in at least one DGRP line; aligned to genomic sequence of *D. simulans* or *D. yakuba*; did not contain transposable element sequences; were located on chromosomes 2, 3, or X; and did not align to any transcripts (annotated or unannotated) from any of the 9 tested outgroup species. In total, these filtration steps yielded transcripts for 119 *D. melanogaster*-specific candidate *de novo*

genes. We also defined a second set of 140 "unannotated ancestral" (UA) transcripts whose origin likely predated the *melanogaster*/*simulans* split. These transcripts aligned to unannotated transcripts from *D. simulans* and/or *D. yakuba* but passed all other criteria.

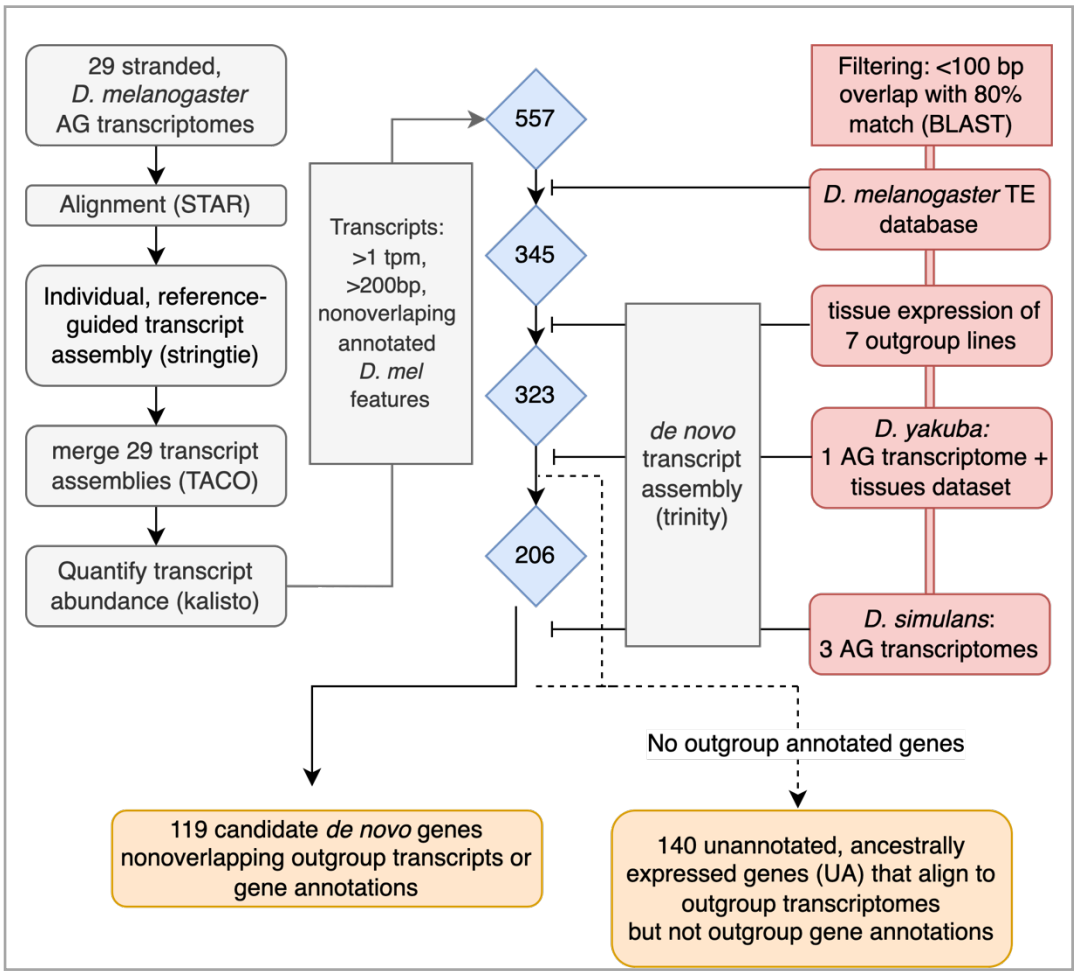


Figure 1: Workflow of candidate *de novo* gene discovery. Left column show steps for identifying genotype-specific transcripts expressed in AGs. In brief, individually aligned transciptomes were also individually assembled and then merged, across 29 DGRP genotypes, into one unified set. This assembly was appended to the *melanogaster* reference genome annotation 6.41 and then used to re-quantify transcript abundance across all individual *melanogaster* lines. Three preliminary filters were applied to the set of unannotated transcripts: they shared no overlapping bases with any annotated transcripts from *D. melanogaster* genome v6.41, were expressed at least 1 tpm in at least one *D. melanogaster* DGRP line and were at least 200 bp long. Next, we sequentially filtered out transcripts that showed >100 bp alignment and 80% sequence identity with any of the following expressed features, arguing against *de novo* origin in *D. melanogaster* (red boxes on

right): transposable element sequences from the *D. melanogaster* v6.41 annotation; reads mapping to unannotated transcripts expressed in the outgroup 7 species sequenced in Yang 2018 (see table S2 for list of species and tissues in this dataset); *D. yakuba* r1.3 annotated transcripts, unannotated transcripts from Yang et al, and unannotated transcripts from our dataset; *D. simulans* r2.02 annotated transcripts and transcripts from our dataset. For outgroup species, we constructed *de novo* transcriptome assemblies using Trinity (v2.9.0) for each separate genotype or tissue expression sample (see table S1 for list of genotypes). After all filtering steps, we identified 119 candidate *de novo* genes and 140 unannotated transcripts that aligned to *D. simulans* or *D. yakuba* transcriptomes but did not align to annotated genes in any outgroup species.

To quantify the abundance of candidate *de novo* and annotated transcripts, we appended the unannotated sequences to the *D. melanogaster* reference annotation v6.41 and used kallisto (Bray et al., 2016) v0.46.2 to quantify transcript abundance in all DGRP lines, an African *melanogaster* line (ed10), and the *melanogaster* genome strain (iso-1). We used the tximport R package to summarize transcript counts across genes and collect TPM measurements. To compare our calculated expression values to those of replicates from the same tissue but with slight differences in rearing and RNA library preparation methods, we also quantified transcript abundance in the data from Cridland 2022 and Zhao 2022 (see Table 1 for list of genotypes) using the same custom annotation file containing the appended unannotated transcripts (*i.e.* we did not separately call *de novo* genes using these libraries).

Comparison of de novo candidates to annotated genes

To compare *de novo* genes to older genes with similar tissue expression profiles, we selected *D. melanogaster* annotated genes (v6.41 melanogaster release) that demonstrated strong AG bias. This procedure was similar to that used in Cridland 2022. We used the FlyAtlas2 data (Leader et al., 2018) to identify genes that showed the highest expression in the *D. melanogaster* AG and showed tissue specificity index (Yanai et al., 2005) $\tau > 0.9$. Next, genes with mean expression <1 TPM across DGRP lines were filtered. We also removed genes that we could not

verify originated prior to the *D. melanogaster-D.simulans* split. To accomplish this, we processed these annotated genes through the same BLAST pipeline used to identify *de novo* genes. Fifteen annotated AG genes, all noncoding RNAs, which did not have any BLAST matches >100 bp to annotated genes or *de novo* transcriptome assemblies in the outgroup species were removed. In total, 452 AG-enriched *D. melanogaster* genes passed these criteria (Figure 2A).

Analysis of bidirectional promoters

First, we identified all genes occurring in bidirectional pairs in the genome using a prior definition of two –/+ oriented genes within 1 kb (Behura and David W. Severson, 2015). Next, we binned genes by annotation status, expression pattern consistent with *de novo* origin, and expression specificity to AGs and counted cooccurrences of each gene class in bidirectional pairs. We than sought to determine whether certain gene class pairs were more likely to be coregulated. To do this, we then transformed transcript-level RNA counts using edgeR (Robinson et al., 2010) to correct for sequencing differences between libraries and calculated spearman correlation coefficient between genes in divergently transcribed pairs.

Motif analysis of promoter and upstream regions

To determine if the upstream regulatory regions of unannotated genes contained similar binding motifs to ancestral genes, we pulled sequences 1 kb upstream of all genes using bedtools getfasta command. We used the find individual motif occurrence utility from MEME suite (Bailey et al., 2015) under default settings to determine motif locations of position weight matrices from the iDMMPMM (improved Drosophila Melanogaster Major Position Matrix Motifs) collection. The iDMMPMM contains 39 PWMs with experimental support from DNase I footprinting,

SELEX, ChIP-chip, and the bacterial-one-hybrid experiments. When calculating occurrences per kb of regulatory sequence, we merged these regions (per gene class) so sequences were not sampled multiple times when promoter regions overlapped. To create samples of intergenic sequence, we downloaded intergenic fasta file from Flybase, subtracted regions overlapping TEs, repeats, and heterochromatic regions, then randomly selected 1000 non overlapping sequences 1000bp in length. HOMER was used to identify *de novo* motifs from the 1 kb promoter regions of candidate *de novo* transcripts, relative to annotated, AG enriched genes and to all annotated genes using command findMotifs.pl with option -fasta. We repeated this process with a core set of promoter motifs, based on the region -50 to + 50 relative to the TSS, using the set of core Drosophila motifs included with the HOMER software (Heinz et al., 2010).

Results

Identification and frequency distribution of unannotated transcripts

We sequenced 29 *D. melanogaster* genotypes in order to identify previously-unannotated transcripts expressed in AGs from the DGRP population. We looked to classify transcripts by whether they were present only in *D. melanogaster* and, if so, whether they were present in all *D. melanogaster* genotypes. To do so, we aligned reference-guided transcript assemblies to previously available and newly-generated sequencing data from outgroup species and TE databases. Given our focus on *D. melanogaster de no*vo genes, the asymmetric sampling regime, and non-AG specific sequencing data of most outgroup species, we choose not to take a fully phylostratigraphic approach (Domazet-Loso et al., 2007) by mapping point of origin. Instead, we sorted unannotated transcripts into two classes: *D. melanogaster*-specific unannotated transcripts,

and other unannotated transcripts expressed both within *D. melanogaster* and elsewhere in *D. simulans* and/or *D. yakuba*.

We found 119 full-length transcripts that were unique to *D. melanogaster* and expressed at 1 tpm or higher in least one DGRP line (Figure 2A; Figure 2B). We refer to these full-length transcripts as "candidate *de novo* genes", given the absence of contrary evidence. These 119 intergenic genes are more than the 49 intergenic genes found previously by using only 6 genotypes (Cridland et al., 2022). This sampling reiterated the fact that the average candidate *de novo* gene is uncommon within accessory gland tissue, and therefore sampling the rarest alleles requires many libraries from many genotypes.

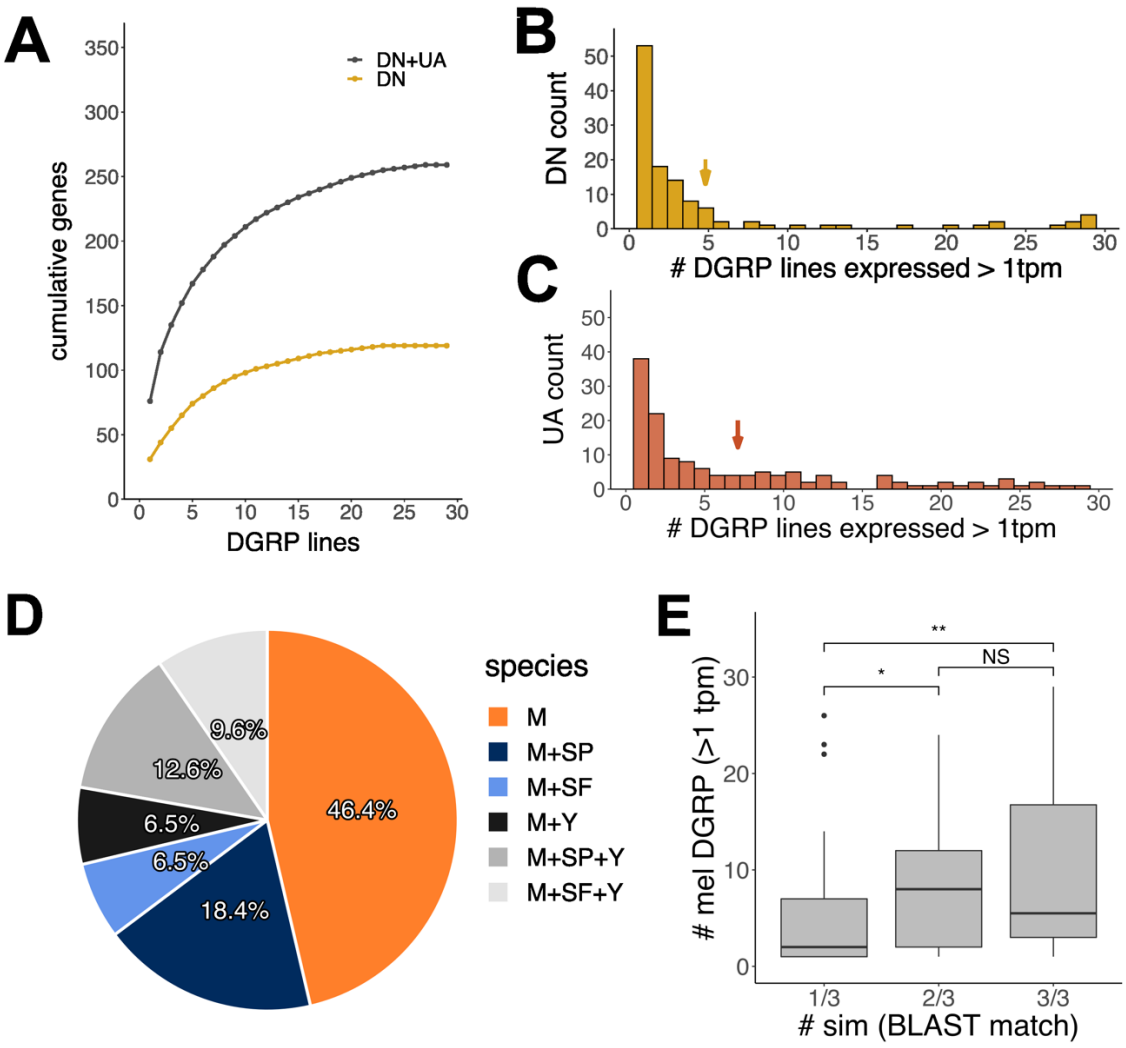


Figure 2: Most unannotated genes are at low frequency in the population. (A) Rarefaction plot depicting unannotated genes discovered per new genotype. "All unannotated genes" include both *de novo* and unannotated, ancestrally expressed genes (UA) in *D. simulans* or *D. yakuba.* Of 119 de novo genes, we found only 4 fixed genes that were present in all 29 DGRP lines. Lines with more unique *de novo* genes are plotted first. (B) Comparison of *de novo* and (C) UA gene frequencies in the *DGRP*. Expressed alleles at cutoff >1 TPM are at significantly higher frequency for UA genes than for DN genes (Wilcoxon rank sum test with continuity correction; $p<0.001$). Arrows indicate mean values. (D) Expression distribution UA genes in *D. simulans* and *D. yakuba*. Presence in outgroup genotype was established through transcript overlap of candidate *de novo* genes with *de novo* transcript assemblies of AG transcriptomes. Abbreviations: M=melanogaster, Y=yakuba, SP=simulans polymorphic (1/3 or 2/3 lines), SF=simulans fixed (3/3 lines). (E) Unannotated transcripts expressed in *D. simulans* genotypes are more likely to be rare in DGRP

when they are also rare in *D. simulans* (Wilcoxon rank sum test with continuity correction; *p<0.05, **=p<0.01).

Of these 119 de novo genes, we found only 4 fixed genes that were present in all 29 DGRP lines at >1 TPM, plus the two non-DGRP *D. melanogaster* sequences. This low proportion of fixed genes parallel the very low proportion found in accessory glands in Cridland 2022 (2/133, though this included intronic genes). Though we used 29 replicates from a single North American population to identify *de novo* genes, comparatively high transcription of candidate *de novo* genes in the African line ED10 suggest that the identified genes likely are not specific to the DGRP population (Figures S1-S2).

We found 140 unannotated, ancestrally-expressed transcripts (UA) that aligned to unannotated transcripts from either *D. simulans* or *D. yakuba* (Figures 2C-D). Rapid sequence divergence makes it difficult to pinpoint the time of their origin; these transcripts may be a mix of relatively recent *de novo* genes and older rapidly evolving genes. Much like the candidate *de novo* genes, the older UA genes also did not appear to be fixed within species. Only 1/140 UA gene was fixed in *D. melanogaster*, with >1 tpm in all 29 DGRP lines (and the other 2 *D. melanogaster* genotypes). Likewise, of the UA genes with transcriptional evidence in *D. simulans*, only 34.2% had alignable transcripts in all three tested genotypes. In addition, we also found some variation within replicates of the same lines. We had replicate libraries from our study and Cridland 2022 for *D. yakuba* strain tai18 and *D. simulans* strain w501. When *D. yakuba* replicate libraries aligned to an unannotated *melanogaster* transcript, both libraries identified the transcript 56/80 times (70%). Replicates were in greater concordance for *D. simulans* line w501, at 102/108 (94.4%).

We found UA genes that were expressed at lower frequency in the DGRP were also expressed at lower frequency in *D. simulans* (Figure 2E). Since so many candidate *de novo* genes

were rare in *D. melanogaster*, it is logical to conclude that some outgroup-expressed alleles were unsampled in our study. Therefore, some of our candidate *D. melanogaster de novo* genes may predate the *D. melanogaster*/*simulans* split. We also found several UA genes that were both polymorphic within *D. simulans* and found in *D. yakuba* (Figure 2D). Overall, the extent of polymorphism observed across multiple species demonstrate a contributor to the imprecision of phylostratigraphic methods to map gene origin.

Properties of de novo candidates and other unannotated transcripts

Next, we looked at several features of candidate *de novo* genes to see to what extent gene age may be associated with some metrics of genic function. As points of comparison, we chose UA genes and a set of 452 annotated genes enriched for expression in AGs and present outside of *D. melanogaster* (Figure 3A). Candidate *de novo* genes shared several properties that distinguish them from other gene classes, including UA genes. Candidate *de novo* genes were expressed at low levels and with high degree of expression variance between lines (Figure 3B). They were also shorter (Figure 3C) and contained fewer exons (Figure 3D) than other gene classes (pairwise Wilcoxon rank sum test; $p<0.001$). Interestingly, the depletion of *de novo* genes on the X was not as strong as UA or AG-specific annotated genes (Figure 3E ; pairwise Fisher exact tests with Holm correction for multiple comparisons; DN vs annotated $p=0.19$; UA vs annotated $p=0.011$; AG vs other annotated $p<0.001$). Candidate *de novo* genes generally lack ORFs, which has been previously shown to be particularly true for *de novo* genes in AGs. Five *de novo* genes had signal sequences. Thus our data are consistent with previous studies of *D. melanogaster* in that *de novo* genes exhibit reduced features associated with protein coding genes (Zhao et al., 2014, Cridland et al., 2022).

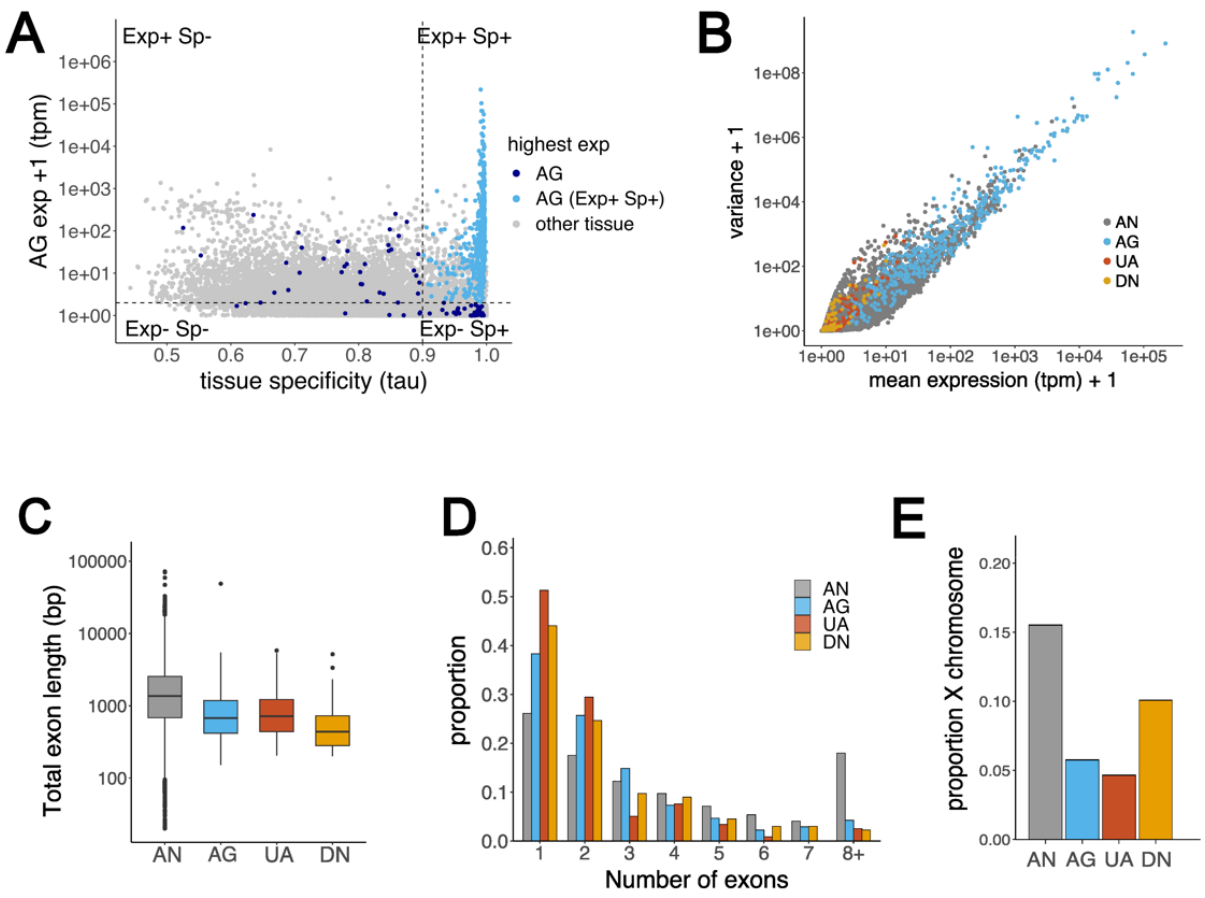


Figure 3: Properties of accessory gland (AG) enriched transcripts. (A) Data from our study showing mean AG expression (Exp) of all annotated genes on X axis vs data from FlyAtlas 2 (Leader et al., 2018) showing tissue specificity (Sp) on y axis, based on index tau. Light blue points in top right quadrant comprise annotated, AG-enriched genes in this study. (B) Most candidate *de novo* transcripts exhibit low mean expression with high relative expression variance compared to annotated genes. Abbreviations: AN (-AG): all annotated genes excluding set of AG enriched genes; AG: Accessory gland enriched annotated genes (same as light blue from figure A); UA: unannotated, expressed in *simulans* or *yakuba*; and DN: *de novo*. (C) Gene length distribution. Candidate *de novo* transcripts are shorter than other gene types, including UA genes (pairwise Wilcoxon rank sum test with holm adjustment; $p<0.001$ for all contrasts). Non-AG-enriched, annotated genes are longer than other gene classes ($p<0.001$ for all contrasts), but annotated AG-specific genes are not significantly longer than ancestrally-expressed unannotated transcripts ($p=0.3$). (D) Exon numbers of full-length transcripts. *De novo* transcripts have fewer exons than annotated genes and unannotated, ancestral genes (Pairwise Wilcoxon rank sum test with Holm

adjustment). Annotated, AG-specific genes do not contain significantly more exons than unannotated ancestrally-expressed genes. Arithmetic means, per gene class, of the number of exons were: AN = 4.67 , AG= 2.65, UA = 2.68, and DN= 1.98 (E) *De novo* genes are not as depleted on the X as other AG gene classes. Though at a lower frequency, *de novo* genes were not significantly depleted on the X relative to annotated genes with less restricted tissue expression patterns after correcting for multiple comparisons (pairwise Fisher exact tests with Holm correction for multiple comparisons; DN vs annotated $p=0.19$; UA vs annotated $p=0.011$; AG vs other annotated $p<0.001$).

It has been posed that very young *de novo* genes start with little function, but this changes as transcripts become refined by selection and spread across the population. In particular, short ORFs could gain new functions after mutations cause the formation of longer coding sequences (Bornberg-Bauer et al., 2015). While these *de novo* genes were typically ORF poor, it is still possible that ncRNAs alone might show a similar effect if length is correlated with function, and previous associations have been drawn between ncRNA length and evolutionary conservation (Sang et al., 2021). We explored how three metrics potentially associated with "functionality" correlated with frequency in the DGRP population, in both candidate *de novo* and UA genes. We found that gene length (Figure 4A-B) and exon number (Figure 4C-4D) did not correlate with expressed allele frequency as strongly as highest measured TPM value (Figure 4E-F). Since *de novo* genes and UA genes shared similar relationships for all three metrics, we did not see any effect of phylostratigraphic age (limited to *D. melanogaster* only vs *D melanogaster + D. simulans* and/or *D.yakuba*) on how proportion affects these characteristics.

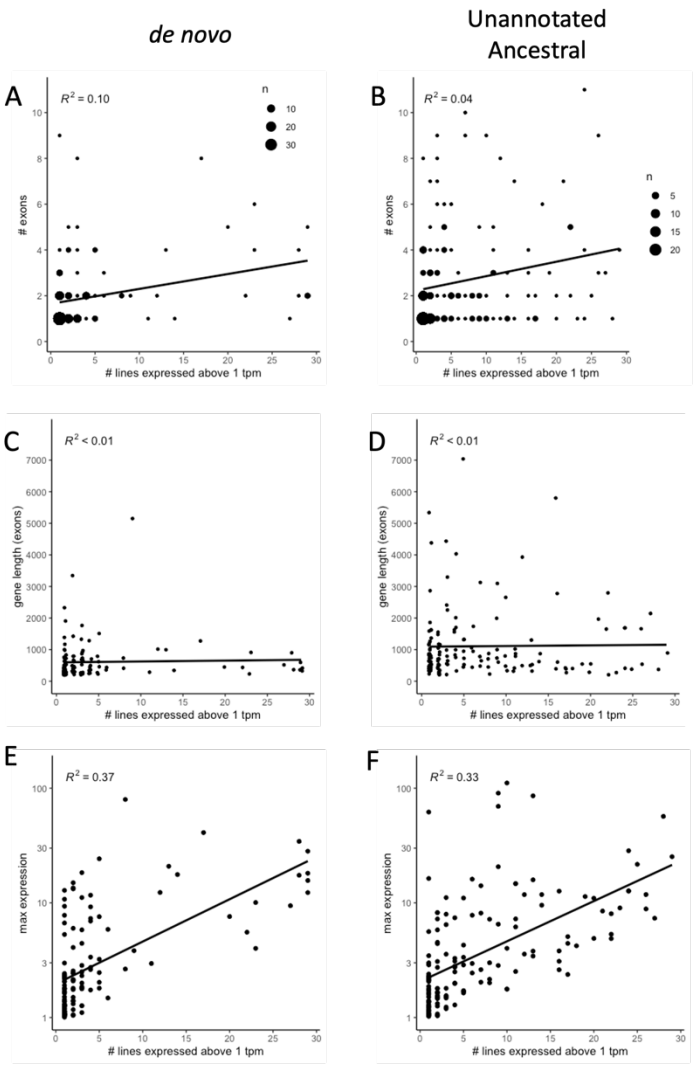


Figure 4: Properties of unannotated genes by frequency of expressed alleles. There is a weak correlation between exon number (A and B), no correlation with gene length (C and D), and a strong correlation between highest measured expression of unannotated transcripts (E and F) and the rarity of expressed alleles at >1 tpm. Both classes of unannotated transcripts (*de novo* – first column, UA – second column) exhibited similar correlations across categories.

Location of unannotated genes

Next, we explored whether existing genomic structures may facilitate *de novo* gene evolution. We first re-examined the effect of gene proximity using our increased sample size of *candidate de novo* genes (119 genes here vs 49 in Cridland 2022). If tissue-specific *de novo* genes are more likely to evolve in regions with existing tissue-specific regulatory elements, it would

reason that they occur more frequently near other genes expressed specifically in that tissue. We found that on a broad, genomic scale, *de novo* genes are more likely to occur near tissue-specific genes. However, this effect was stronger for older genes than for *de novo* genes. First, we broke the genome into 100 kb windows and tested whether AG-specific genes and *de novo* genes were more likely to occur in the same windows, which we found to be the case. Next, we looked at the locations of *de novo* gene TSSs to determine whether they were more likely to occur near TSSs of other gene classes (Figure 5A). In the broader windowing analysis, *de novo* genes were significantly more likely to be close to AG-annotated genes. However, when looking a TSS distance alone, it is clear this proximity effect to annotated AG-specific genes is smaller for the young candidate *de novo* genes, and greater for the older UA genes and AG annotated genes. There also appeared to be some DN genes that were very close to each other (Figure 5A), which could indicate changes in the region may spawn multiple *de novo* genes at once.

Bidirectional transcription of AG-expressed genes

We next looked at the effect of bidirectional promoters as a mechanism of *de novo* gene origin. Since tissue-specific promoter regions are enriched for tissue-specific regulatory elements, it has been hypothesized that *de novo* gene evolution in these regions may require fewer nucleotide changes to modify unidirectional promoter to accommodate transcription in both directions. First, it should be noted that the proportion of genes fulfilling this arbitrary cutoff for bidirectional transcription was a minority in any class of genes (Figure 5B). Yet within the genes that did fulfill this criteria, we found interesting trends between gene classes. Bidirectional transcription did not appear to be a common mechanism of *de novo* gene origin when paired with tissue-specific annotated genes. Like in the case of gene proximity, the raw values of DN-AG bidirectional pairs

were not as impressive in comparison to UA-AG and AG-AG pairs. Only 4.2% of *de novo* genes occurred in a bidirectional pair with annotated, AG-specific genes. In contrast, 12.5% of UA genes were in UA-AG pairs and 23.1% AG genes were in AG-AG pairs. This effect was significant for both contrasts (DN-AG vs AG-AG $p<0.001$; DN-AG vs UA-AG $p<0.028$; pairwise Fisher's exact tests with Holm correction). We also found a small number of candidate *de novo* genes were in bidirectional pairs with another *de novo* gene (6.6%), which could reflect single regulatory changes spawning multiple transcripts.

If bidirectionally-transcribed genes utilizing aspects of the same regulatory structure, genetic differences could foreseeably affect the amount of both genes in the same manner. To test this, we measured the extent of correlated expression between bidirectional gene pairs across the 29 DGRP lines (Figure 5D). The few DN – annotated AG enriched gene bidirectional pairs exhibited weaker correlations than UA – DN and DN – DN pairs. This does not suggest stronger regulatory activity around annotated, tissue-specific genes often spawns *de novo* genes as a byproduct. Yet on a larger scale there was some effect of correlated expression differences. Most AG–AG pairs were positively correlated, and correlations between AG-AG pairs were significantly higher than annotated broadly expressed – broadly expressed pairs. So while correlated pairs are possible, the lack within AG-DN did not generally seem to support a mechanism of *de novo* gene origin by reusing the promoters of tissue-specific genes.

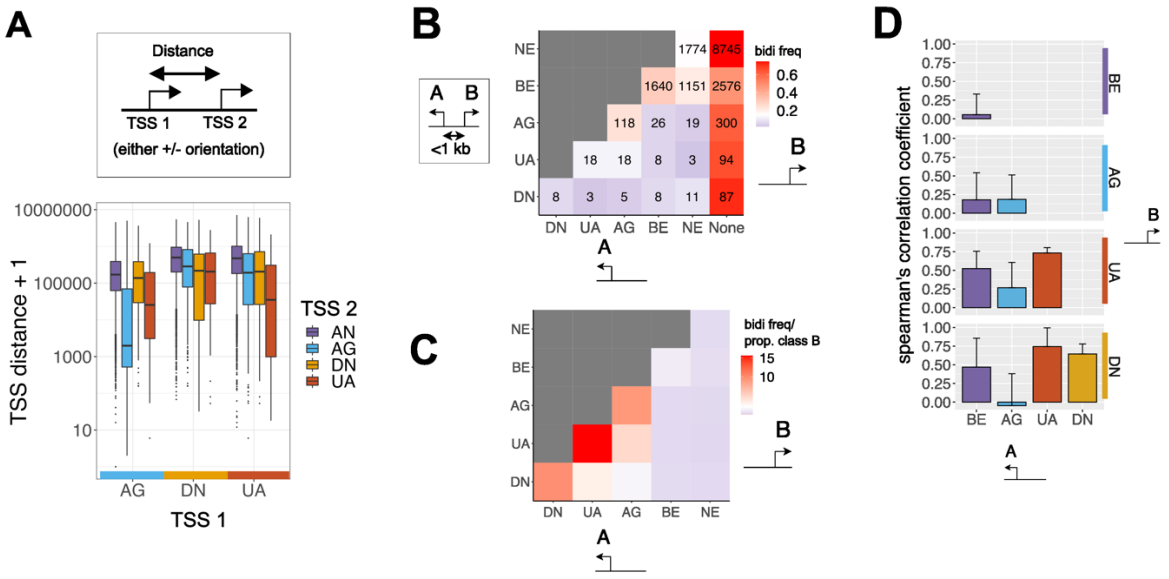


Figure 5: Positional biases in *de novo* gene origin. Stats (A) Closest TSS, by class of gene. Orientation of either transcript was not factored into calculation. (B) Frequencies of each gene class in divergently transcribed gene pairs. Each cell gives the proportion of genes in category A that occur in a bidirectional pair with a gene in category B ("none" indicates the proportion of genes that are not part of bidirectional pairs). In this figure, each gene in a bidirectional pair is counted once as part of category A and once as part of category B (*e.g.* eight *de novo* transcripts stemming from *de novo-de novo* bidirectional promoters, though only four promoters exist). Abbreviations: DN – *de novo*; UA – unannotated ancestrally-expressed; AG – accessory gland specific annotated genes; BE – subset of more broadly expressed annotated genes that are expressed, but not enriched within AGs; NE – annotated but not expressed in AGs. (C) Same data as figure 5b, but values are adjusted to account for the different numbers of genes per gene class. The proportions of genes per row in 5b are divided by total number of genes in each column's class, as a fraction of total genes. (D) Correlation (Spearman's ρ) between first and second gene in bidirectional promoter, by gene class. Error bars denote +/- 1 SD. DN-AG pairs significantly less correlated than DN-DN and DN-UA pairs, but not DN-BE pairs (ANOVA with Tukey HSD).

Motif analysis of nearby cis regulatory regions

We next tested whether the composition of *de novo* gene regulatory regions resembled that of AG-specific genes, since both could draw a tissue-restricted expression domain from a similar regulatory source. We had an *a priori* expectation that the transcription factor *paired* may be involved in *de novo* gene regulation, since it has a well-characterized role in AG gene development

and gene regulation (Xue and Noll, 2002). We analyzed the 1kb upstream regions of all genes to find occurrences of known TF binding motifs, adjusted by a baseline rate measured from intergenic sequence. First, we found significantly more binding sites for *paired* in *de novo* genes and AG enriched genes than the baseline rate within intergenic sequence (figure 5A; pairwise fisher-exact tests). These are not explainable by GC content, since candidate *de novo* genes did not differ significantly from any other gene class or intergenic sequence (one way ANOVA with Tukey HSD, figure S6). We then compared rates of different 39 motifs between gene classes. Genes specific to AGs appear to be depleted for binding sites of the gene *serpent*, which is important for fat cell identity. Though few other transcription factors stand out as likely candidates, we found AG-enriched genes and *de novo* genes clustered the closest together.

Next, we focused on occurrences of core promoter motifs (restricted to sequences -50 to +50, relative to the TSS). In general, genes with AG-enriched expression were depleted for core promoter motifs (Figure 6C). Candidate *de novo* genes were particularly depleted, even compared to randomly sampled intergenic sequence. Motifs in exception include *caudal*, *bicoid*, and *unknown6*, which were all enriched in *de novo* gene promoters relative to all annotated genes. While annotated, AG-specific genes were generally also promoter motif poor, they did have large number of TATA box motifs. These results may be consistent with selection against promoter motifs conferring broader expression domains to *de novo* genes.

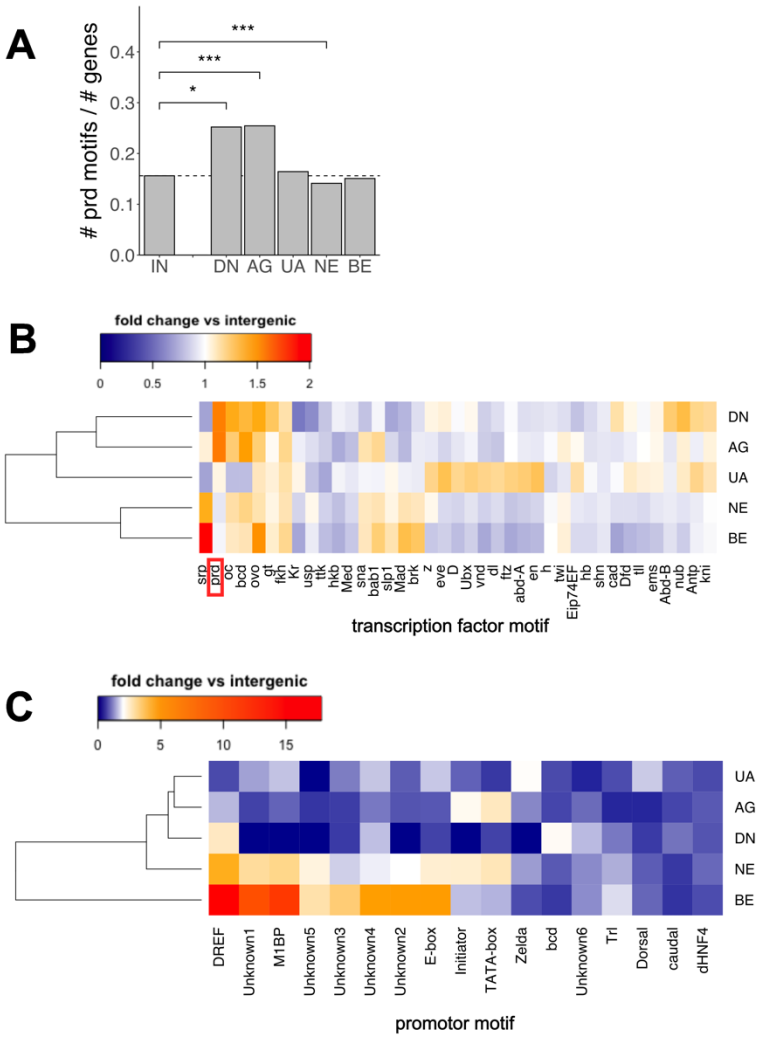


Figure 6: Similarities between regulatory regions of *de novo* genes and annotated, AG-enriched genes. (A) Relative occurrences of the motif prd in sequence 1 kb upstream of TSS. Each gene class is compared to a baseline rate calculated from 10,000 total 1000 bp regions of intergenic sequence, which did not overlap annotated genes, TEs, or repeat regions. (B) Heatmap of the occurrence rate for TF binding motifs in sequence 1kb upstream of TSS, relative to their frequency in intergenic space. Individual TF binding motifs obtained from iDMMPMM database (n=39). Hierarchical Euclidean clustering algorithm (left of heatmap) shows *de novo* genes and AG enriched annotated genes are most similar in terms of motif occurrence rates. Abbreviations-DN: candidate *de novo* genes; AG: accessory gland enriched genes; UA: unnanotated, ancestrally expressed genes; NE: annotated genes not expressed in accessory glands; BE: broadly expressed annotated genes that are expressed in accessory glands. (C) Heatmap of promoter motif occurrence rates per kb core promoter sequence. The *de novo* genes are particularly motif poor in comparison to broadly-expressed, annotated genes.

Discussion

In this study, we examined how additional population sampling affects our understanding of *de novo* genes. Some of our findings offer new insight into how *de novo* genes evolve while simultaneously confirming many previous patterns. Study systems and identification methodology have varied, yet common qualities of *de novo* genes have been largely consistent: *de novo* genes are shorter, contain fewer exons, and are less likely to contain ORFs than other annotated genes (when noncoding RNAs are also considered). When multiple individuals within a population are sampled, studies indicate that *de novo* genes are likely to be polymorphic and rare. Other properties seem to be less universal, varying more depending on methodology and particular study system. For instance, there have been conflicting results on the intrinsic stability of *de novo* genes and their depletion on sex chromosomes. We found that *de novo* genes were not as depleted on the X as older, genes enriched for expression within the same tissue.

While our study supported known properties of *de novo* genes, the quantity of tissue-specific transcriptomes facilitated unique insights into *de novo* gene evolution. We found that while *de novo* genes are located closer to annotated, tissue-specific genes than by random chance, this enrichment is to a lesser degree than proximity exhibited by tissue-specific genes to each other. We also found that regulatory environments around *de novo* genes exhibit an enrichment of TF binding sites associated with tissue-specific gene expression, but they are depleted for the binding sites present in more broadly-expressed genes. Together, these results suggest that the youngest *de novo* genes are not directly piggybacking on the most proximal tissue-specific regulatory elements, but use some of the same regulatory sequences. These results raise three possibilities for how *de novo* genes acquire tissue-specific regulatory patterns. It is possible that *de novo* genes use: (1) tissue-specific enhancers located in intergenic space, (2) tissue-specific enhancers near annotated

genes that must fold to reach further away *de novo* genes, or (3) entirely novel enhancer regions with convergently-evolved motif binging sites. Yet that *de novo* genes do not evolve nearby the closer, old regulatory regions could logically point to them not using the more distant old, regulatory regions either.

Further work is needed to determine the mechanism by which *de novo* genes become expressed and are able to establish expression patterns distinct from neighboring genes. Unlike in humans, promoters in *Drosophila* are not thought to be intrinsically bidirectional, and seen "bidirectional" activity of promoter regions may instead be caused by separate core promoter motifs in each transcriptional direction (Behura and David W. Severson, 2015). Though one previous analysis of bidirectional promoters in insects showed that the frequency of bidirectional promoters was largely a function of genome compactness, some divergently transcribed gene pairs exhibit correlated expression. One possibility is that the relative depletion AG-DN bidirectional gene pairs may reflect functional constraint at the gene regulatory level. When *de novo* genes do evolve near ancestral, annotated genes, they don't typically exhibit strong coregulation. It is possible that selection maintains expression levels of the preexisting gene. If this is the case, new variants working to increase the expression of the surrounding region may cause maladaptive changes to the older, annotated gene. Conversely, when two *de novo* genes do evolve in bidirectional pairs, their expression levels tend to be highly correlated. Since it seems unlikely to converge on separate expressed alleles simultaneously, this indicates individual regulatory changes can spawn multiple new transcripts. It is possible that such genes are eRNAs, since transcription from enhancers in *Drosophila* does exhibit some bidirectional activity. Positionally, there may be less "cost" to regulatory changes in intergenic space (around bidirectional *de novo* genes) where changes may be less likely to affect established genes. Even the exceptions to these

patterns may prove informative. Some cases where *de novo* genes and annotated genes do have correlated expression patterns occur when the annotated gene is polymorphic for expression within the accessory gland, and therefore does not appear essential to tissue function (figure S2).

We found some key similarities between *de novo* genes and accessory gland specific in terms of their regulatory sequence composition. Aside from the enrichment of the TATA box in AG-specific genes, neither were particularly enriched for common promoter motifs. One possibility is that these sequences contribute to broad expression across tissues and therefore are disfavored by selection. We found both AG-specific genes and *de novo* genes are enriched for *PRD* binding sites, a transcription factor with both high expression abundance and functional evidence of regulatory activity in the accessory gland. Though prd is perhaps the most likely candidate for encoding AG-specific gene expression, prd sites are still not particularly abundant in the 1 kb region upstream of *de novo* genes. More work is needed to determine other causes of expression specificity, both for *D. melanogaster* AGs and for other *de novo* genes in other systems.

Sources of false negatives for de novo origin

In terms of accurately assessing *de novo* gene origin, it is a concern that rare transcripts in *D. melanogaster* are also more likely to be rare in *D. simulans*. It suggests a proportion of false-positives that have not been sampled in outgroups, since outgroups were not sampled as extensively. Current practice in most studies is to only sample a few genotypes per species. These regimes are probably inadequate if our finding is replicated in other systems. However, many studies have required ORF evidence, which may not be as polymorphic as the transcription-based approach used here. Yet these studies come with the drawback of ignoring one source of novel transcripts.

The false-positive rate for *de novo* gene origin has been discussed at length in other studies, but an unsampled polymorphism present in multiple species has not factored strongly in these discussions. In Figure 7, we outline four potential scenarios for a transcript to be erroneously classified as a *de novo* gene. Prior discussion has revolved more around the potential for transcripts to undergo a fixed loss in multiple lineages (Figure 7 "fixed loss" - scenario 1) and the potential to misidentify orthologous genes if intergenic sequence evolved rapidly (Figure 7 "sequence divergence" - scenario 2) or if the two lineages are very old (Vakirlis et al., 2018, Weisman et al., 2020). In this study, we show evidence that these false positives may also be due to either longstanding genetic polymorphism (Figure 7 "old polymorphism" - scenario 3) and the same genotype not always being able to sample a transcript (Figure 7 "expression variance" - scenario 4). All these reasons why a gene appears taxonomically-restricted are not necessarily mutually exclusive. For instance, a gene expressed in the common ancestor of *D. melanogaster*, *D. simulans*, *D. yakuba*, and *D. ananasae* could be expressed but not identifiable in *D. ananasae*, completely lost in *D. yakuba*, and very rare in *D. simulans*. In the next section, we discuss how our results fit in with these sources of false positives.

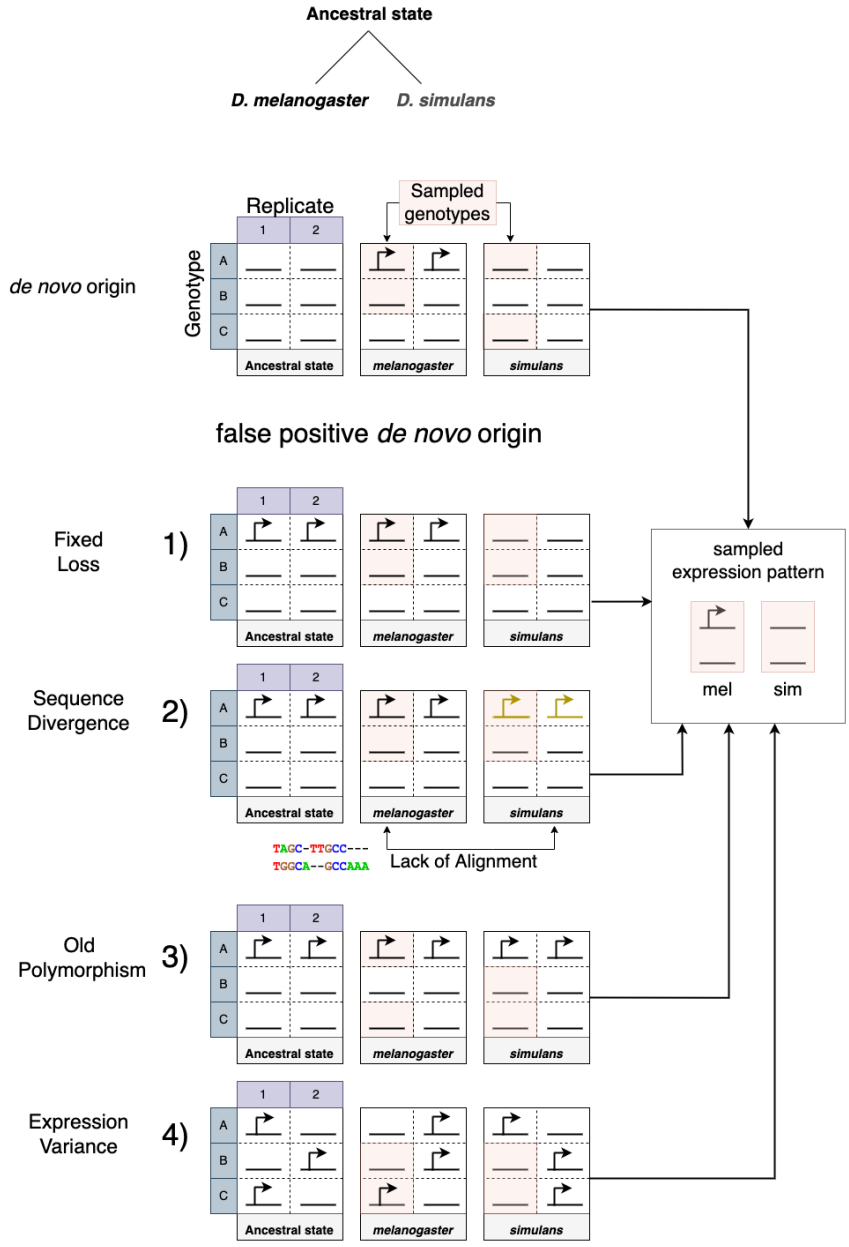


Figure 7: Patterns leading to appearance of polymorphic taxonomically restricted genes. The *de novo* origin of a gene is contrasted with four other scenarios leading to a pattern of polymorphic, lineage-restricted gene expression. In these examples, there are 3 genotypes (A-C) with 2 technical replicates used to call expression. *De novo* origin: gene expression is a derived state sampled only in one extant *D. melanogaster* line. False positive scenarios: 1) Fixed loss: ancestrally polymorphic gene is lost in all genotypes in *D. simulans* lineage. 2) Sequence divergence: despite gene being present in both *D. melanogaster* and *D. simulans,* failure to detect orthology leads to appearance of uniquely-transcribed sequence. 3) Old Polymorphism: the ancestral state (the presence in only some individuals of the population) is maintained in both *D. melanogaster* and *D. simulans* lineages. Since gene may be rare, it is possible to miss evidence of its expression in one or more

lineages. 4) Expression variance: highly-variable expression of *de novo* gene causes disagreement between replicates. Chance sampling leads to appearance of gene loss in *D. simulans.*

A fixed loss in multiple species (Figure 7 scenario 1) becomes less parsimonious the more species are sampled. Here, we used 9 species to account for this possibility. We controlled for the second scenario (Figure 7 scenario 2) by using minimum cutoffs for the genomic sequence alignment of candidate genes to *D. simulans* and *D. yakuba,* however we did not filter *de novo* genes that did not meet these for species more distantly-related to *D. melanogaster*. Despite the range of tissues and number of species sampled, we found transcriptomes from Yang et al. 2018 identified very few unannotated genes present in DGRP accessory glands (Figure 1). If sequences are too diverged, no amount of sampling in these species could show that the *D. melanogaster* gene did not originate *de novo.* However, all unannotated transcripts identified from these distant outgroup libraries were also found in our AG tissue-specific datasets of *D. yakuba* and/or *D. simulans,* which does not suggest many unique hidden orthologs are only be found in these distant species. One other technical reason for this pattern may be coverage dropout of AG-specific genes, since they included male gonad but not AG-specific libraries. Or it is possible that the longer read lengths in our study facilitated longer transcript assemblies.

Distinguishing between false positives due to rare polymorphic alleles (Figure 7 scenario 3) and variable expression (Figure 7 scenario 4) is a challenge when exploring very short timescales. Expression variance is difficult to compare between *de novo* genes and annotated genes, since any class-specific differences are confounded by the lower expression of *de novo* genes (Figure S5). And yet, our data suggests both may be possible. The fact that different *D. simulans* genotypes identified different unannotated transcripts suggest an effect of scenario 3, as does the fact that most polymorphic alleles are rare. That the same genotypes, in different

experiments, identified different unannotated transcripts also suggests an effect of scenario 4. These may be real biological properties of *de novo* genes. Yet from an origin agnostic standpoint, genes with these properties would be more likely to be classified as phylogenetically restricted.

There is no consensus as to the extent a transcript must be absent in outgroups. Indeed, the preadaptation hypothesis (Wilson et al., 2017) suggests that *de novo* genes will be more likely to evolve in leaky regions. But with the strictest filtering criteria, the existence of transcriptional noise in outgroups would preclude the identification of a gene as *de novo*. In our study, we found that the older UA transcripts were more likely be found near annotated, AG genes than *de novo* genes – a regulatory environment hypothetically conducive to their origin. Without the sampling of recent common ancestors, these UA genes may eventually become a fixed difference with no alignable orthologs, making it challenging to resolve their origin time.

One major take-home from these additional genotypes is that *de novo* gene expression within species can be variable, and differences between species may not be clear cut. Relying on transcriptional evidence is necessary for fully understanding all avenues of *de novo* gene origin, yet studies tracing ORFs do not run into all of the same expression-threshold challenges. If the goal is strict criteria of species-specific *de novo* origin, our results suggest the most efficient due diligence is to sample the same tissue and focus on species where intergenic sequences remain mostly alignable.

References:

Bailey, T.L., Johnson, J., Grant, C.E., Noble, W.S., 2015. The MEME Suite. Nucleic Acids Res. 43, W39-49. https://doi.org/10.1093/nar/gkv416

Begun, D.J., Lindfors, H.A., Kern, A.D., Jones, C.D., 2007. Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade. Genetics 176, 1131–1137. https://doi.org/10.1534/genetics.106.069245

Begun, D.J., Lindfors, H.A., Thompson, M.E., Holloway, A.K., 2006. Recently evolved genes identified from Drosophila yakuba and D. erecta accessory gland expressed sequence tags. Genetics 172, 1675–1681. https://doi.org/10.1534/genetics.105.050336

Behura, S.K., David W. Severson, 2015. Bidirectional promoters of insects: genome-wide comparison, evolutionary implication and influence on gene expression. J. Mol. Biol. 427, 521–536. https://doi.org/10.1016/j.jmb.2014.11.008

Blevins, W.R., Ruiz-Orera, J., Messeguer, X., Blasco-Moreno, B., Villanueva-Cañas, J.L., Espinar, L., Díez, J., Carey, L.B., Albà, M.M., 2021. Uncovering de novo gene birth in yeast using deep transcriptomics. Nat. Commun. 12, 604. https://doi.org/10.1038/s41467-021-20911-3

Bornberg-Bauer, E., Schmitz, J., Heberlein, M., 2015. Emergence of de novo proteins from 'dark genomic matter' by 'grow slow and moult.' Biochem. Soc. Trans. 43, 867–873. https://doi.org/10.1042/BST20150089

Bray, N.L., Pimentel, H., Melsted, P., Pachter, L., 2016. Near-optimal probabilistic RNA-seq quantification. Nat. Biotechnol. 34, 525–527. https://doi.org/10.1038/nbt.3519

Casola, C., 2018. From De Novo to "De Nono": The Majority of Novel Protein-Coding Genes Identified with Phylostratigraphy Are Old Genes or Recent Duplicates. Genome Biol. Evol. 10, 2906–2918. https://doi.org/10.1093/gbe/evy231

Cridland, J.M., Majane, A.C., Zhao, L., Begun, D.J., 2022. Population biology of accessory gland-expressed de novo genes in Drosophila melanogaster. Genetics 220, iyab207. https://doi.org/10.1093/genetics/iyab207

Domazet-Loso, T., Brajković, J., Tautz, D., 2007. A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages. Trends Genet. TIG 23, 533–539. https://doi.org/10.1016/j.tig.2007.08.014

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., Glass, C.K., 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 38, 576–589. https://doi.org/10.1016/j.molcel.2010.05.004

Leader, D.P., Krause, S.A., Pandit, A., Davies, S.A., Dow, J.A.T., 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. https://doi.org/10.1093/nar/gkx976

Li, Z.-W., Chen, X., Wu, Q., Hagmann, J., Han, T.-S., Zou, Y.-P., Ge, S., Guo, Y.-L., 2016. On the Origin of De Novo Genes in Arabidopsis thaliana Populations. Genome Biol. Evol. 8, 2190–2202. https://doi.org/10.1093/gbe/evw164

Mackay, T.F.C., Richards, S., Stone, E.A., Barbadilla, A., Ayroles, J.F., Zhu, D., Casillas, S., Han, Y., Magwire, M.M., Cridland, J.M., Richardson, M.F., Anholt, R.R.H., Barrón, M., Bess, C., Blankenburg, K.P., Carbone, M.A., Castellano, D., Chaboub, L., Duncan, L.,

Harris, Z., Javaid, M., Jayaseelan, J.C., Jhangiani, S.N., Jordan, K.W., Lara, F., Lawrence, F., Lee, S.L., Librado, P., Linheiro, R.S., Lyman, R.F., Mackey, A.J., Munidasa, M., Muzny, D.M., Nazareth, L., Newsham, I., Perales, L., Pu, L.-L., Qu, C., Ràmia, M., Reid, J.G., Rollmann, S.M., Rozas, J., Saada, N., Turlapati, L., Worley, K.C., Wu, Y.-Q., Yamamoto, A., Zhu, Y., Bergman, C.M., Thornton, K.R., Mittelman, D., Gibbs, R.A., 2012. The Drosophila melanogaster Genetic Reference Panel. Nature 482, 173–178. https://doi.org/10.1038/nature10811

Majane, A.C., Cridland, J.M., Begun, D.J., 2022. Single-nucleus transcriptomes reveal evolutionary and functional properties of cell types in the Drosophila accessory gland. Genetics 220, iyab213. https://doi.org/10.1093/genetics/iyab213

Neil, H., Malabat, C., d'Aubenton-Carafa, Y., Xu, Z., Steinmetz, L.M., Jacquier, A., 2009. Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. Nature 457, 1038–1042. https://doi.org/10.1038/nature07747

Ohno, S., 2013. Evolution by Gene Duplication. Springer Science & Business Media.

Oss, S.B.V., Carvunis, A.-R., 2019. De novo gene birth. PLOS Genet. 15, e1008160. https://doi.org/10.1371/journal.pgen.1008160

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140. https://doi.org/10.1093/bioinformatics/btp616

Ruiz-Orera, J., Hernandez-Rodriguez, J., Chiva, C., Sabidó, E., Kondova, I., Bontrop, R., Marqués-Bonet, T., Albà, M.M., 2015. Origins of De Novo Genes in Human and Chimpanzee. PLOS Genet. 11, e1005721. https://doi.org/10.1371/journal.pgen.1005721

Sang, S., Chen, W., Zhang, D., Zhang, X., Yang, W., Liu, C., 2021. Data integration and evolutionary analysis of long non-coding RNAs in 25 flowering plants. BMC Genomics 22, 739. https://doi.org/10.1186/s12864-021-08047-6

Soucy, S.M., Huang, J., Gogarten, J.P., 2015. Horizontal gene transfer: building the web of life. Nat. Rev. Genet. 16, 472–482. https://doi.org/10.1038/nrg3962

Vakirlis, N., Hebert, A.S., Opulente, D.A., Achaz, G., Hittinger, C.T., Fischer, G., Coon, J.J., Lafontaine, I., 2018. A Molecular Portrait of De Novo Genes in Yeasts. Mol. Biol. Evol. 35, 631–645. https://doi.org/10.1093/molbev/msx315

Weisman, C.M., Murray, A.W., Eddy, S.R., 2020. Many, but not all, lineage-specific genes can be explained by homology detection failure. PLOS Biol. 18, e3000862. https://doi.org/10.1371/journal.pbio.3000862

Werner, M.S., Sieriebriennikov, B., Prabh, N., Loschko, T., Lanz, C., Sommer, R.J., 2018. Young genes have distinct gene structure, epigenetic profiles, and transcriptional regulation. Genome Res. 28, 1675–1687. https://doi.org/10.1101/gr.234872.118

Wigby, S., Brown, N.C., Allen, S.E., Misra, S., Sitnik, J.L., Sepil, I., Clark, A.G., Wolfner, M.F., 2020. The Drosophila seminal proteome and its role in postcopulatory sexual selection. Philos. Trans. R. Soc. B Biol. Sci. 375, 20200072. https://doi.org/10.1098/rstb.2020.0072

Wilson, B.A., Foy, S.G., Neme, R., Masel, J., 2017. Young Genes are Highly Disordered as Predicted by the Preadaptation Hypothesis of De Novo Gene Birth. Nat. Ecol. Evol. 1, 0146. https://doi.org/10.1038/s41559-017-0146

Xue, L., Noll, M., 2002. Dual role of the Pax gene paired in accessory gland development of Drosophila. Development 129, 339–346. https://doi.org/10.1242/dev.129.2.339

Yanai, I., Benjamin, H., Shmoish, M., Chalifa-Caspi, V., Shklar, M., Ophir, R., Bar-Even, A., Horn-Saban, S., Safran, M., Domany, E., Lancet, D., Shmueli, O., 2005. Genome-wide

midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics 21, 650–659. https://doi.org/10.1093/bioinformatics/bti042
Zhao, L., Saelao, P., Jones, C.D., Begun, D.J., 2014. Origin and spread of de novo genes in Drosophila melanogaster populations. Science 343, 769–772. https://doi.org/10.1126/science.1248286

Chapter II: Open chromatin near *de novo* genes is typically a derived change without a simple genetic explanation

Logan Blair, Julie Cridland, David Begun, and Artyom Kopp

Abstract

The evolution of genes *de novo* from a non-transcribed state has been reported across a wide range of taxa. While *de novo* gene transcription could imply novel regulatory activation, in most cases what specifically has changed from the ancestral sequence is unknown. In this study, we sought to test whether *de novo* genes evolve near conserved enhancer sequences, or alternatively, whether *de novo* genes use new regulatory regions. We focused on regions near young *Drosophila melanogaster de novo* genes which are still segregating in the species. First, we used ATAC-seq in *D. melanogaster* and found a general trend towards increased chromatin accessibility near actively-transcribed *de novo* genes, but not within genotypes that do not express them. Second, we generated ATAC-seq for the related species *D. simulans* and *D. yakuba.* We found the most *D. melanogaster* open regions closed in these outgroups, suggesting open chromatin near *ne novo* genes is not the ancestral state. Finally, we used a set of 29 genotypes to map *cis*-variants associated with *de novo* gene expression in accessory gland tissue. Following additional validation, we found three that harbored variants explaining most variance. We conclude that *de novo* genes often use new regulatory sequences, but rarely is this explained by a single genetic change within an ATAC-seq peak boundary.

Introduction

Evolution has been described as a process more likely to "tinker" with existing than to "engineer" from scratch (Jacob, 1977). Many case studies show new phenotypes occurring through changes in existing units (such as genes or *cis*-regulatory elements). Yet the *de novo* origin of genes, by definition, starts from sequences that were not functionally transcribed (Begun et al., 2006). How do we resolve the disconnect between the existence of *de nov*o genes across the tree of life (Ruiz-Orera et al., 2015, Li et al., 2016, Vakirlis et al., 2018) with the supposed improbability their origin (Weisman, 2022)?

One conceptual roadblock to *de novo* gene origin lies in the number of steps that may need to occur. To persist as a new gene, intergenic sequence may: (1) contain open reading frames (ORFs) to create protein product, (2) not be removed from the gene pool by natural selection, and (3) be transcribed. Of these steps, the origin of ORFs is perhaps the most straightforward to examine. A large fraction of *de novo* gene candidates have short or nonexisting ORFs (Zhao et al., 2014). However, many noncoding RNAs still affect phenotype, and it is possible for noncoding *de novo* genes to serve important roles even without coding sequence. Other data show that when *de novo* genes do contain ORFs, there is no one path for how they are acquired. Some *de novo* genes are born from newly-transcribed intergenic ORFs, whereas others start as transcribed, noncoding sequence that may gain ORFs later (Reinhardt et al., 2013).

The nature of *de novo* gene expression makes it challenging to measure their fitness impacts. Most *de novo* genes are not at high frequency in the population, and evolutionary comparisons suggest most will be lost. However, some *de novo* genes become fixed, and a few expressed *de novo* gene alleles have lower nucleotide diversity than their inert counterparts that are consistent with hard selective sweeps (Zhao et al., 2014). Since selection will work to remove

deleterious alleles, extant *de novo* genes at a high frequency in the population are either very unlikely occurrences or they have a net neutral-positive effect on fitness. It is challenging to show why - protein evolution metrics, like DN/DS, fail in cases where open reading frames are short. But the behavior of some extant *de novo* genes help explain why they may stick around, even in cases where they do not currently have function. First, prior to widespread transcription, the "preadaptation" hypothesis suggests that deleterious ORFs can be purged from transcriptionally noisy regions (Wilson et al., 2017). Sequences could be drawn from intergenic sequence less likely to form toxic aggregations (Kosinski et al., 2022). Second, *de novo* genes tend to be tissue-restricted, limiting their transcriptional exposure. This may reflect a survivorship bias where more broadly expressed *de novo* genes are quickly purged. Third, a constant influx of new, weakly-deleterious transcripts can still lead to the fixation of new genes due to drift or subsequent beneficial mutations (Moutinho et al., 2022).

In this report, we focus on the transcriptional origins of de novo genes. Though little is known about the mechanisms through which *de novo* genes become transcribed, existing promoters and enhancers are an attractive starting point for *de novo* gene transcription. Hypothetically, existing promoters with new bidirectional activity could transcribe new sequences in the opposite direction. Their association with *de novo* genes has differed between studies (Vakirlis et al., 2018, Blevins et al., 2021). In *Drosophila melanogaster* accessory glands, only a few *de novo* genes were divergently-transcribed from same promoter region as other same-tissue-specific genes (Blair et al. 2022, unpublished). Bidirectional pairs with just a single unannotated gene did not show highly-correlated expression, suggesting each transcriptional direction may be regulated separately. The promiscuity of many enhancers may facilitate nearby *de novo* gene evolution. Use of tissue-specific enhancers would explain the why many *de novo* genes are also

tissue-restricted in their expression. Obviously, promoters are located near older genes, but enhancers, on average, are more likely to be as well. Therefore, hypothetically there may be a proximity effect for *de novo* genes and older genes.

It is first worth asking how closely the transcriptional environment of *de novo* genes resembles that of older genes. A common answer is that young *de novo* genes often have epigenetic signatures associated with active transcription, though possibly to a lesser degree than older genes. In *Arabidopsis thaliana*, *de novo* genes exhibited DNA methylation signatures between annotated genes and non-transcribed intergenic space (Li et al., 2016). The chromatin environment surrounding *de novo* genes may resemble enhancers or promoters (Majic and Payne, 2020, Werner et al., 2018). Other indirect evidence also exists. In *Saccharomyces*, *de novo* genes were enriched near nucleosomes-depleted recombination hotspots (Vakirlis et al., 2018), and in *Drosophila*, the permissible germline transcriptional environment is thought be a large factor in the high number of *de novo* genes expressed in the testis (Witt et al., 2019). Together, these studies suggest an endpoint of transcriptional activity, though in most cases it remains unclear how this differs from the ancestral state.

Two of the chief concerns in studying *de novo* gene evolution are what evolutionary timescales to use and how conservatively to call transcripts *de novo* genes. Here, we use as recent a timescale as possible – looking within a single species – in order to minimize distance between expressed and non-expressed alleles. This very recent timescale, however, necessitates relaxing the likelihood that these transcripts affect phenotype, since most of the youngest transcripts are the least "genic". As such, we do not require these genes to have ORFs, and use a very relaxed expression cutoff of >1 TPM in at least one line. Using these criteria, we previously sequenced accessory gland (AG) tissue of 29 DGRP lines within *D. melanogaster* in order to better assess

their natural variation in the population. We identified 119 *de novo* gene candidates unique to *D. melanogaster*. Most of these (117/119) appeared to be polymorphic (only expressed in some genotypes of the species). The use of AG tissue in our study is of particular use here, since the lower than usual heterogeneity of the tissue (Majane et al., 2022) increases the signal-to-noise for the average gene. Prior allelic imbalance assays have shown many AG genes draw their expression basis from both *cis* and *trans*-regulatory factors (Cridland et al., 2022), suggesting local enhancer or promoter changes may play a part, but do not entirely explain the expression basis for many *de novo* genes.

In this study, we explore several questions relating to the origin of *de novo* gene transcription. First, we ask to what extent *de novo* gene regulatory sequences are also engineered from scratch. To answer this question we use the chromatin accessibility profiling technique of ATAC-seq (Buenrostro et al., 2015), which shows signal at promoter and enhancer sequences. If only *D. melanogaster* lines expressing *de novo* genes show chromatin accessibility – and not the non-expressed other genotypes from *D. melanogaster* or the closely-related outgroup species – it would suggest the use of new regulatory sequences within the tissue. Second, we investigate whether specific genetic changes leading to *de novo* gene transcription can be mapped. To do this, we use *cis*-regulatory association mapping to explore the extent to which local variants may be associated with *de novo* gene expression. Finally, we explore the interplay between these two factors. If *de novo* genes use old enhancers, may changes directly be traced to these enhancer sequences? If *de novo* genes rely on *de novo* regulatory sites, are there simple genetic variants associated with these new enhancers or promoters?

Materials and Methods

RNA materials and de novo genes

We used a set of 119 previously-described *de novo* genes expressed in *D. melanogaster* accessory glands (Blair et al; Table 1). These were identified from 29 RNA-seq libraries from individual Drosophila genetic resource panel (DGRP - Mackay et al., 2012) isofemale lines. These lines were originally drawn from a single location in Raleigh, NC.

AG-specific annotated genes

Since *de novo* genes are typically enriched for expression within a single tissue, the genomic regions near older, tissue-specific annotated genes provide a convenient point of comparison for the acquisition of tissue-specific regulation. We previously identified a set of 452 tissue-enriched genes that 1) showed evidence that they expressed more in accessory glands than any other tissues, 2) most expression is specific to AGs (tissue specificity index $\tau > 0.9$), 3) were expressed in the lines we used for RNA libraries (mean tpm>1), and 4) have identifiable orthologs in *D. simulans* or *D. yakuba*.

ATAC-seq collection

We performed ATAC-seq experiments in whole accessory gland tissue, using 3 technical replicates of 5 DGRP lines in *D. melanogaster* and 1 line for the related species *D. simulans* (3 technical replicates) and *D. yakuba* (1 technical replicate; Table 1). Each sample was collected from 15 males 48 +/- 2 hours post eclosion. Tissues were lysed in 200 µl of ATAC-seq lysis buffer (10 mM Tris-HCl, 10 mM NaCl, 3 mM MgCl2, 0.1% IGEPAL CA-630) and manually homogenized 25 times with a plastic pestle, followed by a 1-min incubation on ice, and repeated twice. Following homogenization, the samples were pelleted at 4 °C (100 g for 10 min) to recover

the nuclei. We removed the supernatant and washed the pellet in 200 µl of the lysis buffer. The nuclei preparation was filtered through a 40-µm cell strainer and washed with another 200 µl of lysis buffer. The purified nuclei were isolated by centrifugation at 1,000 × g for 10 min at 4 °C. After removing the supernatant, 12.5 µl Nextera Tagment DNA Buffer (Illumina), 11.25 µl ddH2O and 1.25 µl Tn5 Transposase were added to the purified nuclei to tagment the DNA. Libraries were then processed using the workflow of Buenrostro et al., 2015 with a final upper and lower size-selection using SPRI beads with bead-to-sample ratios of 0.4× and 1.7×, respectively. An aliquot of the purified library was analyzed on a Bioanalyzer (Agilent) to ensure the characteristic nucleosome periodicity of ATAC-seq libraries. Then sample concentration was determined using Qubit fluorometer and sequenced using 150 bp paired-end reads on an Illumina Hiseq4000 (Table S4).

Species	Data type (reps)	Lines
D melanogaster	RNA (1x)	Blair 2022: 153, 217, 229, 287, 304, 320, 338, 352, 357, 359, 360, 370, 380, **399**, **517**, 530, 563, 630, 703, 761, 805, 812, 820, 822, 85, 900, 911, 93 Zhang 2022: 208, 379, **399**, 427, **517**, 799
D melanogaster	ATAC (3x)	304, 357, 360, **399**, **517**
D simulans	ATAC (3x)	w501
D yakuba	ATAC (1x)	tai18

Table 1: Lines used in study. Lines in bold are replicated between this study and Zhang 2022.

ATAC-seq data processing

Reads were examined with fastQC. Raw ATAC-seq reads were trimmed using cutadapt. Alignment was performed using *bowtie2* on using parameters -X 2000 --local --very-sensitive-local. We observed that *Wolbachia*-infected strains had, on average, worse mapping statistics than other strains (Table S4). We examined the fragment length distributions of reads to verify the

characteristic signal of nucleosome periodicity (Figure S7). Peak calling was initially performed on each genotype individually using Genrich v.0.5 206 (https://github.com/jsh58/Genrich) parameters -j -q 0.001 -d 200. To create a set of global peak coordinates, peaks from all individual *D. melanogaster* genotypes were merged using bedtools mergeBed. This strategy retained peaks present in any single line, resulting in 13,748 peaks across all genotypes. Finally, ATAC-seq enrichment for individual samples from the unified set of peaks was quantified using featureCounts.

ATAC-seq analysis in D. melanogaster

To quantify differences in ATAC-seq signal between lines, we used the likelihood ratio test in Deseq2 (Love et al., 2014). This test compares the full model of all genotypes to the reduced model of any individual sample removed, assigning differential accessibility when one or more genotypes are substantially different in peak coverage. To select peaks exhibiting high differential enrichment between samples, we used Benjamini Hochberg (Benjamini and Hochberg, 1995) adjusted p value <0.05 as a cutoff.

To match ATAC-seq peaks with the transcripts likely affected by their accessibility, we identified peaks in genomic regions whose coordinates exhibited any partial overlap with previously annotated, or unannotated (including *de novo*) transcripts expressed at >1tpm. These peaks were most frequently centered near promoter regions. This conservative approach excludes ATAC-seq peaks located distantly upstream or distantly downstream from the transcripts, although such peaks could in principle have an effect on transcription. We also retrieved the first proceeding peak and the first following peak of genes.

To examine the location of the open chromatin, with respect to the TSS, we generated a bedfile of the genomic region 5 kb upstream and 5 kb downstream of the each genes TSS. We then calculated per bp ATAC coverage in these regions, per sample, using bedtools. Next, we normalized these counts by per-replicate read depth in all 10 kb regions centered on the TSSs. We subdivided these regions into 250 bp windows and calculated total ATAC-seq coverage. Since the highest ATAC signal was located within 250 bp upstream and downstream of the TSS, we repeated this procedure for a signal 500 bp window centered on the TSS.

To test the interaction between RNA seq and ATAC seq, we first log transformed RNA counts values+0.1 TPM to better fit ANOVA assumptions. We compared of slopes least-square means using Tukey post hoc tests with the "lsmeans" package (Lenth, 2016).

Inter-species ATAC-seq comparison

To test whether the observed chromatin state was unique to *D. melanogaster*, we used a custom script to identify genomic regions of *D. simulans* and *D. yakuba* orthologous to *D. melanogaster* ATAC-seq peaks. First, we filtered *D. melanogaster* peaks through a reciprocal BLASTn search to each outgroup genome. We retained alignments at e-value cutoff of 10e-10 with a minimum of 150 bp per alignment. Next, we filtered regions with less than 80% of the peak area aligned to the outgroup genome, though we did not require the entire alignment to be contiguous. Regions with fragmented alignments were filtered if the mapped locations were not colinear with respect to the original *D. melanogaster* positions, or if there were insertions or deletions between alignments greater than 10% of the length of the *D. melanogaster* peak. Initial alignments were extended to the boundaries of the *D. melanogaster* peak (*e.g.* a 400 bp *D. melanogaster* peak for which bp 26-360 aligned with the *D. simulans* genome would be extended

25 bp upstream and 40 bp downstream in *D. simulans*). Following this extension, we filtered regions greater than 10% longer than the initial *D. melanogaster* peak.

Next, we quantified the ATAC signal in *D. simulans* and *D. yakuba* in regions orthologous to *D. melanogaster* peaks using the program featureCounts (R package subread). However, coverage to these orthologous regions only constitute a fraction of the total depth of reads from the species' respective ATAC-seq alignments. Since *melanogaster*-specific reads were normalized to the total coverage *D. melanogaster* peaks, we normalized the depth of *D. melanogaster* peak orthologous regions, in *D. simulans* and *D. yakuba*, to the read depth of species-specific peaks. To do this, we called species-specific peaks with default genrich settings, then normalized the coverage to *D. melanogaster* peak orthologous regions by the total depth of coverage in the species-specific peaks.

Association testing

To determine genetic variants significantly associated with *de novo* gene expression, we used variants from the DGRP freeze 2 variant calls (Huang et al., 2014) and calculated associations with Accessory gland libraries from 29 separate DGRP lines. To calculate associations, we used the R package matrixEQTL (Shabalin, 2012) using the additive "modelLINEAR". This conventional linear model was:

$\mathbf{y}=\mu+\mathbf{m}u$

Where y = expression, μ=the intercept, m=the value from the genotype matrix (0 is homozygous reference allele and 2 is homozygous alternative allele; since DGRP lines are near-homozygous, we did not consider any heterozygous sites), and u= the SNP marker effect. The significance of u

is calculated from a t-distribution. We used within 5kb of the gene sequence as a cutoff for *cis* variants.

After an initial cutoff of $p<1*10^{-6}$, we then used available public data to further test the strength of association between expression and genotype. To do so we quantified the expression of our previously-identified *de novo* genes from the sequence data of another study that also measured RNA expression from AG tissue, also from the same population From Zhang et al 2022, there 4 unique and 2 previously quantified RAL (Table 1). We required rank-ordered expression to reaffirm the association to the variant (all lines with variant must be expressed higher than all lines without).

As additional verification, we used qRT-PCR assays on two additional DGRP genotypes not used in any RNA-seq studies (see Table S5 for primers used). Genotypes were selected such that one would have the variant and one would not. RNA was collected as described previously (Blair et al., unpublished). All qRT-PCR assays were conducted using the SsoAdvanced Universal SYBR Green Supermix (Bio-Rad) on a CFX96 qPCR machine (Bio-Rad). Program specifications were: step 1 - 95° initial denaturation (2 min); step 2 - 95° denaturation (30 sec), 55° annealing (15 sec), 60° extension (15 sec), and repeat 40x cycles. Measured expression was relativized to the reference gene RPL20.

Results

ATAC-seq in DGRP

Previously, we sequenced 29 lines from D. melanogaster accessory gland tissue to identify *de novo* genes expressed within the tissue. We identified a total of 119 candidate *de novo* genes.

Most had very low expression and were expressed in very few genotypes. This suggests many *de novo* genes are rare and that very few new genes spread to fixation in the tissue. We also found *de novo* genes were not located particularly close to other annotated tissue-specific genes, yet regions upstream of *de novo* genes are enriched for some of the same transcription factor binding sites. This could indicate that *de novo* genes draw their expression from nearby intergenic enhancers.

In this study, we first explored whether *de novo* genes had open chromatin near their transcription start sites, as suggested by their transcription and the enrichment of nearby tissue-specific transcription factor binding sites in our previous study. To determine the relationship between chromatin accessibility and *de novo* gene origin, we collected ATAC-seq data for 5 DGRP lines (R304, R357, R360, R399, R517). We observed a mean 11,679 peaks per individual line (range: 10,631-12,308). We combined these ATAC-seq peak regions, resulting in a union set of 13,748 peaks. Of these, 4064 exhibited differentially accessibility in at least one line (indicating at least one line is different from the others) according to the log ratio test from R package DEseq2 (Love et al., 2014) ($q<0.01$).

We then looked across different classes of genes for the presence of an ATAC-seq peak overlapping gene exon regions (minimum 1 bp), limiting this analysis to genes expressed >1 tpm in at least one of the five DGRP lines. Since we did not obtain ATAC libraries for all 29 lines from our previous study, and since most *de novo* genes are not expressed in many genotypes, operationally we were only able to examine ATAC-seq signal in 46 out of the full set of 119 *de novo* genes. We chose two points of comparison for this analysis: 452 expressed, annotated genes with similar tissue specificity to the candidate *de novo* genes, and 5213 annotated genes, that while expressed in AGs, are show a greater degree of expression in other tissues.

We found *de novo* genes were less likely to overlap at least one ATAC-seq peak than these two classes of annotated genes (Figure 1A). However, *de novo* genes were significantly more likely to overlap peaks that were differentially accessible (Figure 1B). Thus, *de novo* genes appear less likely to have regions of open chromatin that reach the threshold of a peak call—potentially due to their low expression—but when they do it tends to be more different between lines. This indicates the high between-line variance in *de novo* gene expression (Blair 2022) is also mirrored in the greater differences in chromatin accessibility.

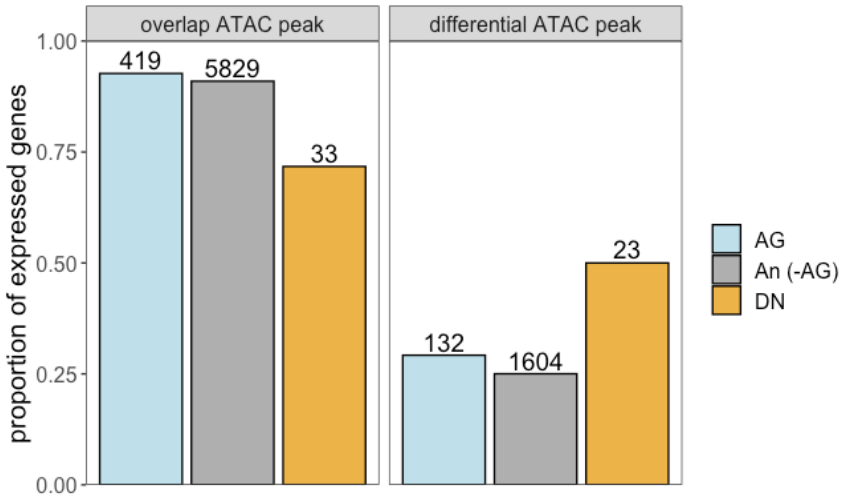


Figure 1*: De novo* genes are less likely to overlap ATAC-seq peaks but these peaks are more likely to be differentially accessible. Numbers above bars indicate counts. Abbreviations – DN: expressed *de novo* genes; AG: accessory gland enriched genes; AN (-AG): other annotated genes expressed in AGs, but more broadly expressed across other tissues.

Location of open chromatin

Many annotated, accessory-gland genes are located in close proximity. This could be consistent with wider genomic pockets of transcriptionally accessible chromatin occurring for older genes than for genes with a *de novo* origin. We found that peaks intersecting annotated AG-

specific genes were significantly wider than peaks intersecting *de novo* genes (Figure 2A; p=0.02952, Wilcoxon rank sum test with continuity correction). Next, we looked at the median bp of these peak with respect to the TSS. Both, on average, were centered near gene TSS. However, ATAC peaks overlapping AG-specific genes had greater variance in the median bp (Figure 2B), consistent with fewer *de novo* genes being located in wide regions of open chromatin accessibility which may be centered far from the gene. We looked at ATAC-seq coverage in nonoverlapping windows near the TSS (Figure 2C). The greatest proportion of the signal was located within 250 bp of the TSS for each gene class, yet the signal in AG annotated genes was greater for the whole region.

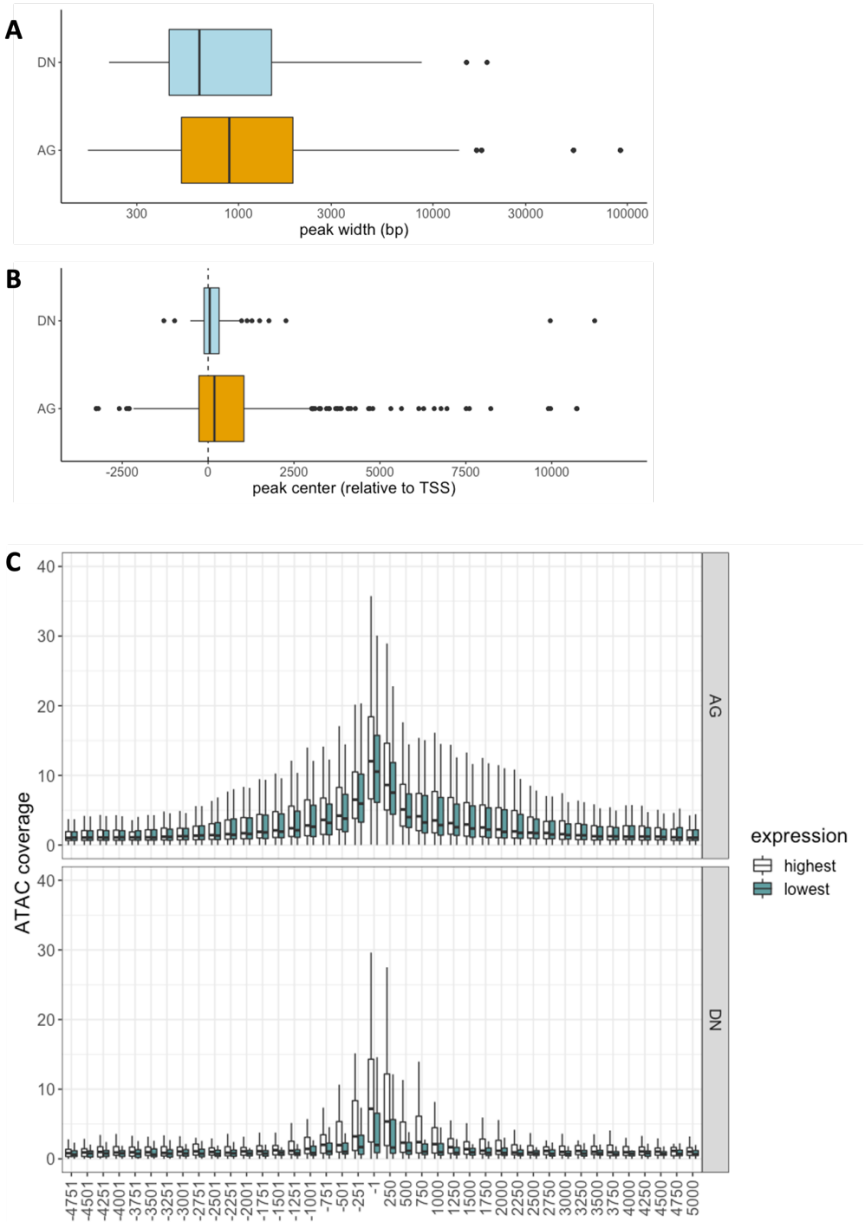


Figure 2: *De novo* gene ATAC-seq peaks are weaker, narrower and localized to TSSs. (A) nonoverlapping windows (250 bp, with upper boundary of window indicated in legend) of ATAC-seq coverage in expressed genes. For visualization, outlier points hidden from graph. Per gene class, ATAC coverage is subdivided between lines the most and least expressed of the five DGRP lines with ATAC-seq data. (B) Width of ATAC-seq peaks intersecting genes, by gene class. AG specific genes (AG) are significantly wider than *de novo* (DN) genes (p=0. 02952, Wilcoxon rank sum test with continuity correction). (C) Location of median bp per ATAC peak, relative to TSS of genes that peak intersects.

Relationship between ATAC-seq and expression between genes

While *de novo* genes have lower chromatin accessibility than AG genes, one possibility is that their signal is similar to that of other annotated genes with similarly-low expression. To test this, we investigated the strength of the relationship between ATAC-seq and expression, as a function of gene class. Unsurprisingly, ATAC-seq and expression were significant correlated for all three gene classes (Pearson's r^2 = AG specific genes: 0.14, annotated, non-AG enriched genes: 0.26, and *de novo* genes: 0.24; $p<0.001$ for all three). The weaker correlations for annotated, tissue-enriched genes are consistent with previous studies (Starks et al., 2019). However, we found differences in the slope of this interaction, which was significantly higher than the other gene types for AG-enriched genes and significantly lower for *de novo* genes. These results suggest *de novo* gene promoter regions have a chromatin that is "more open" than the many of the annotated genes with comparable levels of expression.

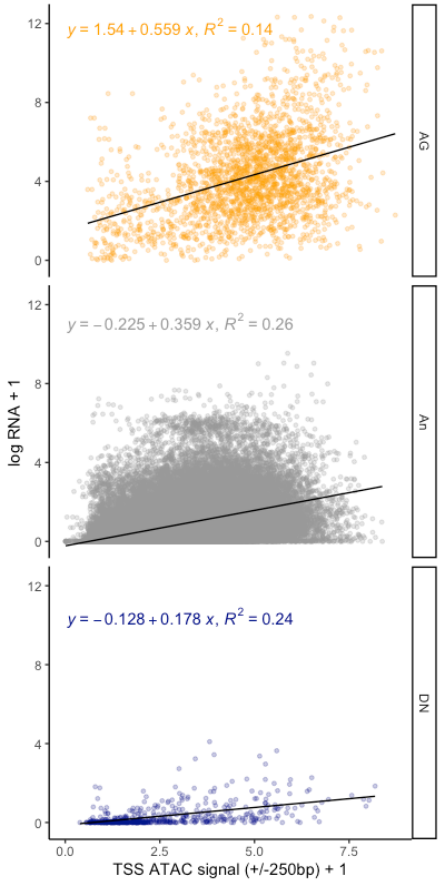


Figure 3: ATAC signal is correlated with expression across all gene classes. Correlations between promoter ATAC-seq (500 bp region centered on TSS) and log RNA-seq counts, by gene class and within line R517. R squared of Pearson correlation coefficients are shown.

Relationship between ATAC-seq and expression within genes, between genotypes

If *de novo* genes use old regulatory regions, it would suggest that genotypes will have open chromatin near *de novo* gene regions regardless of their expression status. We examined the relationship between gene expression and ATAC signal (within 500 bp of TSS) between lines but examining each gene separately. We found most *de novo* genes have a positive relationship between chromatin accessibility and expression, indicating that this is not the case (Figure 4C). This relationship was significantly stronger than that of annotated genes (two-way ANOVA with Tukey HSD $p<0.01$), for which both expression and chromatin accessibility vary less between genotypes. This effect was largely localized to the closest peaks, though a few of the next closest upstream and downstream peaks exhibited a similarly strong positive relationship (Figure 4D).

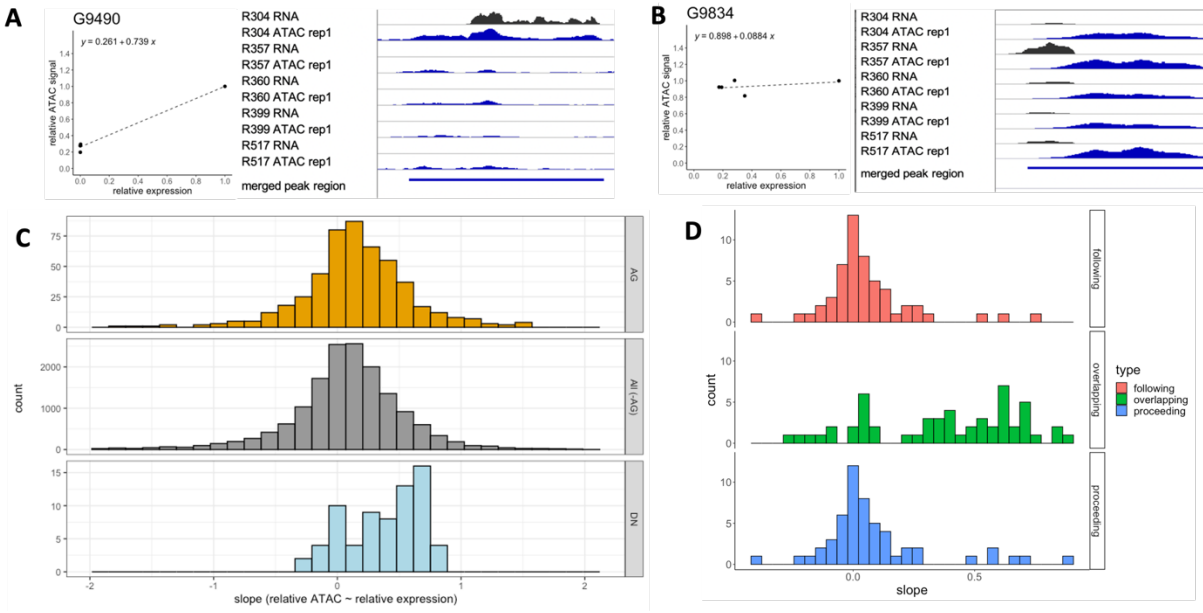


Figure 4: Most *de novo* genes show strong relationship between expression and ATAC-seq signal. (A) First of two example *de novo* gene regions (transcribed on Crick strand). G9490 is expressed in R304, but not the other four lines. Non-expressed lines (R357, R360, R399, and R517) have little ATAC-seq signal in region (shown in blue), therefore there is a positive slope between relative expression and relative ATAC-seq signal in peak region. (B) In contrast, G9834 is a *de novo* gene that does not have a strong relationship between ATAC and RNA expression. Similar ATAC-seq signal is seen regardless of RNA expression in region, corresponding to a lower slope between expression and ATAC signal. (C) Histograms showing distributions of slopes by gene class. Slopes calculated from of relative ATAC signal as a function of relative expression for each gene individually. The slopes for *de novo* genes are significantly higher than other gene classes, though all pairwise contrasts are significant (two-way ANOVA with Tukey HSD $p<0.01$). (D) Slopes of peaks overlapping *de novo* genes (same calculation as C) compared to first non-overlapping peak following gene sequence (top) and proceeding gene sequence (bottom).

ATAC-seq between species

While we found a clear signal of open chromatin near expressed *de novo* genes, yet one possibility is that the openness of these regions constitutes the ancestral state. This may be case if, for instance, *de novo* genes evolve near conserved enhancer region that is subsequently lost in some *D. melanogaster* genotypes. In this scenario, the derived change in *D. melanogaster* would be the loss of open chromatin. To test this, we measured chromatin accessibility of orthlogous peak regions in the closely-related species *Drosophila simulans* and *Drosophila yakuba*. We found that the orthologous sequences to *de novo* gene-intersecting peaks typically had very low accessibility in these two species (Figure 5). This result is consistent with an ancestral state of closed chromatin surrounding *de novo* genes, and does not support a scenario where *de novo* genes reuse conserved enhancer regions directly as promoters.

Since these orthologous regions were the least accessible of those tested, it is tempting to suggest that the non-expressing *D. melanogaster* lines are more open than the ancestral state. However, the design of these contrasts is biased by the *de novo* gene selection procedure. Some of the least expressed melanogaster lines still have some RNA expression, whereas the *de novo* genes are specifically selected to have no transcription in non-*melanogaster* species. Consistent with

greater differences in expression tissue-specific genes, we found ATAC-seq peaks showed less similarity between species in AG genes intersecting peaks than for other annotated genes.

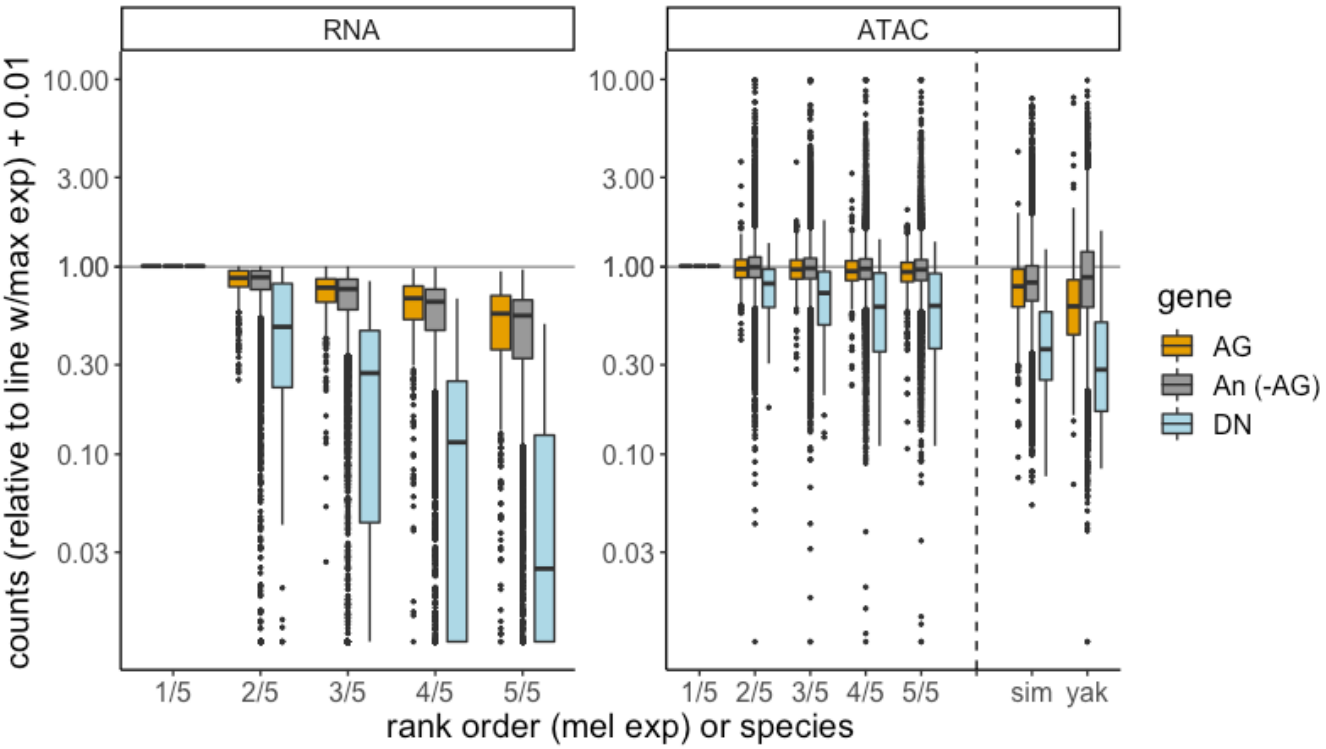


Figure 5: Open chromatin near *D. melanogaster de novo* genes is often unique in *D. melanogaster*-specific expressed genotypes. Left panel shows relative decrease in rank order expression, across different gene types. Right panel shows ATAC signal in the same genotype order as the right panel. Both panels show signal (either RNA or ATAC) relative to which particular genotype has the highest expression of that gene. Also included are relative ATAC signals from the orthologous *D. yakuba* and *D. simulans* peak regions. All ATAC samples have 3x technical replicates except for *D. yakuba*. Some outliers for ATAC signal (>10) are cut off for visualization. The lower relative signal in *simulans* (sim) and *yakuba* (yak) *de novo* gene regions suggest open chromatin is not the ancestral state.

Identification of high-effect variants associated with de novo gene expression

A simple model for *de novo* origin is that a new variant arises in the population, increasing regulatory activity of a genomic region and spawning novel transcripts. Since we found many *de novo* genes have unique regions of accessible chromatin, we sought to determine whether genetic variants associated with expression occurred within these regions. If so, these variants could correspond to sites of regulatory evolution. The scope of this analysis was limited by our sample size – we could only foreseeably find very large-effect SNPs close to *de novo* genes. Therefore, if no significantly-associated variants are found, a *cis*-regulatory basis could still occur if 1) multiple *cis*-haplotypes cause expression or 2) a single *cis*-haplotype causes expression but other required *trans* alleles are segregating in the population.

To test the association between *de novo* gene expression and naturally occurring genetic variants, we measured the association between the *de novo* gene expression of 29 lines and variants within 10 kb of *de novo* genes. Most AG *de novo* genes are uncommon, leading to many imbalanced sample sizes between "expressed" vs "not expressed" alleles. Therefore, despite starting with 119 candidate *de novo* genes, the upper limit of genes in which we could plausibly find genetic associations was much smaller. At a preliminary cutoff of $p<0.0000001$, we found a total of 46 significant variants corresponding to 9 different *de novo* genes.

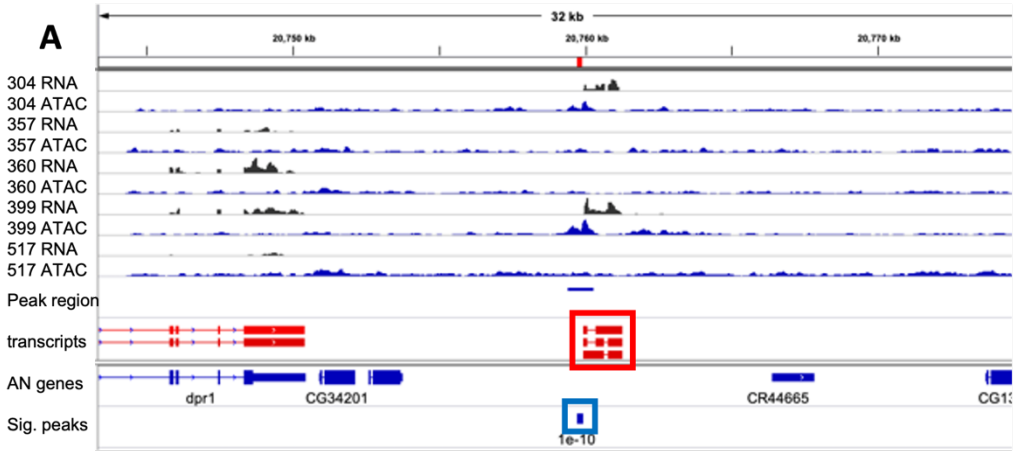
Despite the nominally-significant p-values, low sample sizes can inflate the rate of type I error. Therefore, we then used additional public data from Zhang 2022 to further test the strength of association between expression and genotype, by incorporating 2 more replicate genotypes and 4 more unique genotypes from the same population. We were able to reaffirm the rank order expression in 5/9 of this set (Figure S9). That several of these tests did not completely reaffirm the

effect of the significant variants indicates a fair number of false positives, therefore we independently validated the effect of three more of variants using qRT-PCR in pairs of untested DGRP lines (Figures S10-11). We selected additional pairs of untested DGRP lines: one line contained the derived allele and one line did not. All three of these qRT-PCR tests supported the link between genotype and *de novo* gene expression.

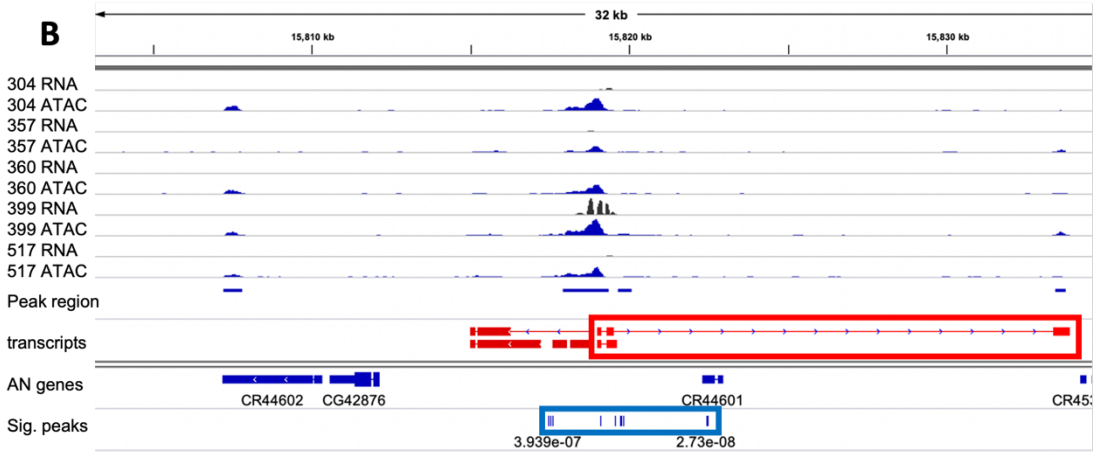
Despite the lack of power in our study, we would be able to tell if the allelic basis was due to a single-*cis* SNP that than spreads through the population. Most genes did not supports a previous allelic-imbalance assay in AG tissue (Cridland et al., 2022), which for many *de novo* genes suggested part of the allelic basis is due to *trans*-regulatory variation. We conclude that expression basis of *de novo* genes usually does not support the simplest model.

Next, we examined the area where these variants were located. All three genes had multiple significant SNPs in the region tested, though the distance between them differed (Figure 5A-C). Each gene had at least one SNP that was located within a differentially-accessible peak region. However, G4151 is a particularly interesting case, since all 13 SNPs were all located within the same peak. Yet since any one SNP, or multiple additive effects between SNPs, could give rise to expression, we were unable to definitively assess what the exact causative change gave rise to expression of these genes.

G4151



G1614



G9490

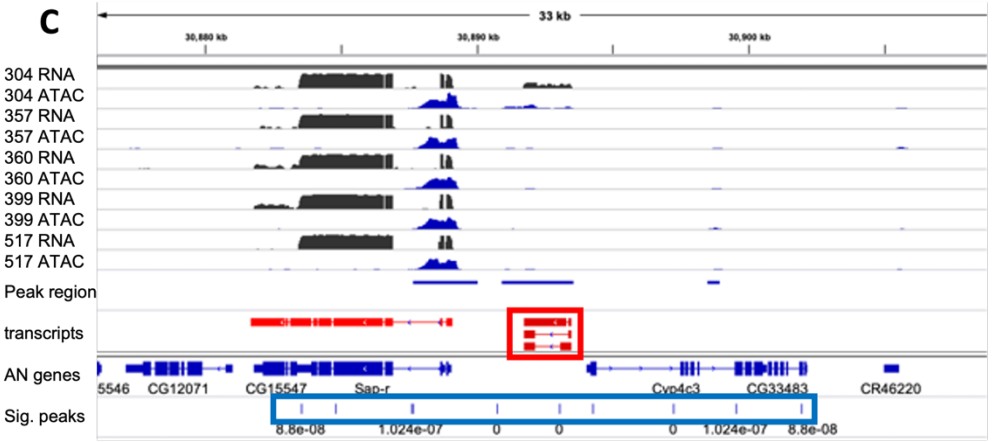


Figure 6: Locations of significant variants differ between *de novo* genes. (A) IGV screenshot showing locations of most significantly-associated SNPs in regions near candidate *de novo* gene G4151. RNA coverage and ATAC coverage tracks (first replicate only) are shown for 5 DGRP genotypes. "Peak region" indicates peaks called from ATAC-seq. "Transcripts" shows output of transcript merge pipeline, which includes annotated + unannotated transcripts expressed in AGs only. Large red box placed around focal de novo gene. "AN genes"- all annotated genes (release Dm6 6.41). Significant SNPs track shows all SNPs within one order of magnitude of lowest association, per gene. G4151 has 13 SNPs within 180bp region (highlighted in blue box). (B) Region for gene G1614, with 12 SNPs in ~2.5 kb region. (C) G9490 has 10 significant SNPs spanning near-entirety of the 20 kb region tested. For G9490, expression shown on log scale due to high transcription of nearby annotated gene.

Discussion

In this report, we examined how *cis*-regulatory sequence and chromatin state changes may cause *de novo* gene expression. We found that *de novo* gene expression is frequently associated with novel regions of open chromatin, but the extent of this association varies. Usually new open regions are most parsimoniously described as a derived change, with the same regions closed in two outgroup species. We found that *de novo* gene expression usually cannot be explained by a single genetic change in regions near the gene, though a few notable exceptions provide evidence that large-effect changes near genes may happen.

Though we found a continuum of chromatin states in outgroup lines, it was perhaps surprising the extent to which *de novo* genes evolve in regions that seem to have a closed ancestral state. This is strong evidence that *de novo* genes do not evolve from slight modifications to existing enhancer sequence. Instead, it seems that unique sequences use engineered, unique regulatory sites. This is an uncommon finding in evolutionary genetics, given the extent to which we find phenotypes arising from modification to existing structures. Locations of open chromatin have been suggested to migrate between species (Maher et al., 2018). These may disproportionately correspond to the locations of *de novo* genes. A tricky "chicken and egg" question follows: did the open chromatin cause transcription, or is some of the ATAC signal attributable to RNA polymerase

activity itself? Such questions require a better mechanistic understanding of the transcriptional cause of each individual *de novo* gene. However, the very low expression of most *de novo* genes might be more suggestive of the former.

Likewise, we found that open chromatin near *de novo* genes does not occur in as wide of "pockets" as for annotated genes. Previously, we found that increases in the expression of nearby older genes largely do not correspond to more *de novo* gene expression, and *de novo* genes are further from AG-specific genes than annotated AG-specific genes are to themselves. One explanation for these findings is that the *de novo* genes that persist may prove to be those less disruptive to the prior regulatory landscape, and thus are more likely to exist further into intergenic space.

We found that *de novo* gene TSSs have a somewhat lower signal of open chromatin compared to tissue-enriched annotated genes (Figure 2C). And yet, expression of these *de novo* genes is typically an order of magnitude lower than their annotated, tissue-enriched counterparts. One possibility is that despite the open chromatin, suboptimal promoter sequence drastically lowers the potential for transcription. We previously found that the regions near *de novo* gene TSSs are depleted for common *D. melanogaster* promoter motifs (Blair et al. 2022). A signal of transcription in enhancer regions could indicate these *de novo* genes function as eRNAs (Harrison and Bose, 2022). Several *de novo* genes have highly correlated, bidirectional transcription from the same promoter regions, another characteristic of eRNAs. However, confidently classifying these open regions as promoters or enhancers would require their associated histone marks to be measured.

Perhaps the simplest model for *de novo* origin is that a new SNP increases the regulatory activity of a genomic region, spawning novel transcripts. However, the general lack of "perfect"

associations between *de novo* genes and local variants indicates that this simple model for *de novo* gene origin is likely not the most common one. Instead, our results are consistent with more complex, quantitative interactions between multiple segregating sites. Our previous allelic imbalance assay in AG tissue showed many AG *de novo* genes have a combination of *cis* and *trans* alleles in hybrids that lead to their expression (Cridland et al., 2022). In theory, few very high-penetrance variants may reflect the end result of easy targets for selection being purged from the population. The preponderance of rare *de novo* gene alleles (Blair et al. 2022) suggests many are selected against before they spread. The three strong associations that we found were not at the higher end of the expression distribution of *de novo* genes. It seems likely the low transcript abundance may correspond to little impact on tissue function.

It is clear that our eQTL design was underpowered, but it is not clear to what extent increasing sample size would lead to better outcomes. The rarity of *de novo* genes leads to imbalanced sample sizes between genotypes that express them and genotypes that do not. One possibility is that other tissues may exhibit a simpler regulatory basis, and thus be easier to map. A higher proportion of genes with *cis*-regulatory basis was found in testis *de novo* genes (Zhao et al., 2014). Furthermore, the expression distribution of testis *de novo* genes appears more bimodal than AG *de novo* genes. Careful model selection may be important for further work into the genetic basis of *de novo* genes, since measuring the requisite number of samples could quickly become costly and time-intensive.

References:

Begun, D.J., Lindfors, H.A., Thompson, M.E., Holloway, A.K., 2006. Recently evolved genes identified from Drosophila yakuba and D. erecta accessory gland expressed sequence tags. Genetics 172, 1675–1681. https://doi.org/10.1534/genetics.105.050336

Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B Methodol. 57, 289–300.

Blevins, W.R., Ruiz-Orera, J., Messeguer, X., Blasco-Moreno, B., Villanueva-Cañas, J.L., Espinar, L., Díez, J., Carey, L.B., Albà, M.M., 2021. Uncovering de novo gene birth in yeast using deep transcriptomics. Nat. Commun. 12, 604. https://doi.org/10.1038/s41467-021-20911-3

Buenrostro, J., Wu, B., Chang, H., Greenleaf, W., 2015. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr. Protoc. Mol. Biol. Ed. Frederick M Ausubel Al 109, 21.29.1-21.29.9. https://doi.org/10.1002/0471142727.mb2129s109

Cridland, J.M., Majane, A.C., Zhao, L., Begun, D.J., 2022. Population biology of accessory gland-expressed de novo genes in Drosophila melanogaster. Genetics 220, iyab207. https://doi.org/10.1093/genetics/iyab207

Harrison, L.J., Bose, D., 2022. Enhancer RNAs step forward: new insights into enhancer function. Development 149, dev200398. https://doi.org/10.1242/dev.200398

Huang, W., Massouras, A., Inoue, Y., Peiffer, J., Ràmia, M., Tarone, A.M., Turlapati, L., Zichner, T., Zhu, D., Lyman, R.F., Magwire, M.M., Blankenburg, K., Carbone, M.A., Chang, K., Ellis, L.L., Fernandez, S., Han, Y., Highnam, G., Hjelmen, C.E., Jack, J.R., Javaid, M., Jayaseelan, J., Kalra, D., Lee, S., Lewis, L., Munidasa, M., Ongeri, F., Patel, S., Perales, L., Perez, A., Pu, L., Rollmann, S.M., Ruth, R., Saada, N., Warner, C., Williams, A., Wu, Y.-Q., Yamamoto, A., Zhang, Y., Zhu, Y., Anholt, R.R.H., Korbel, J.O., Mittelman, D., Muzny, D.M., Gibbs, R.A., Barbadilla, A., Johnston, J.S., Stone, E.A., Richards, S., Deplancke, B., Mackay, T.F.C., 2014. Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. Genome Res. 24, 1193–1208. https://doi.org/10.1101/gr.171546.113

Jacob, F., 1977. Evolution and Tinkering. Science. https://doi.org/10.1126/science.860134

Kosinski, L.J., Aviles, N.R., Gomez, K., Masel, J., 2022. Random Peptides Rich in Small and Disorder-Promoting Amino Acids Are Less Likely to Be Harmful. Genome Biol. Evol. 14, evac085. https://doi.org/10.1093/gbe/evac085

Lenth, R.V., 2016. Least-Squares Means: The R Package lsmeans. J. Stat. Softw. 69, 1–33. https://doi.org/10.18637/jss.v069.i01

Li, Z.-W., Chen, X., Wu, Q., Hagmann, J., Han, T.-S., Zou, Y.-P., Ge, S., Guo, Y.-L., 2016. On the Origin of De Novo Genes in Arabidopsis thaliana Populations. Genome Biol. Evol. 8, 2190–2202. https://doi.org/10.1093/gbe/evw164

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15, 550. https://doi.org/10.1186/s13059-014-0550-8

Mackay, T.F.C., Richards, S., Stone, E.A., Barbadilla, A., Ayroles, J.F., Zhu, D., Casillas, S., Han, Y., Magwire, M.M., Cridland, J.M., Richardson, M.F., Anholt, R.R.H., Barrón, M., Bess, C., Blankenburg, K.P., Carbone, M.A., Castellano, D., Chaboub, L., Duncan, L., Harris, Z., Javaid, M., Jayaseelan, J.C., Jhangiani, S.N., Jordan, K.W., Lara, F., Lawrence, F., Lee, S.L., Librado, P., Linheiro, R.S., Lyman, R.F., Mackey, A.J., Munidasa, M., Muzny, D.M., Nazareth, L., Newsham, I., Perales, L., Pu, L.-L., Qu, C., Ràmia, M., Reid, J.G., Rollmann, S.M., Rozas, J., Saada, N., Turlapati, L., Worley, K.C., Wu, Y.-Q., Yamamoto, A., Zhu, Y., Bergman, C.M., Thornton, K.R., Mittelman, D., Gibbs, R.A., 2012. The Drosophila melanogaster Genetic Reference Panel. Nature 482, 173–178. https://doi.org/10.1038/nature10811

Maher, K.A., Bajic, M., Kajala, K., Reynoso, M., Pauluzzi, G., West, D.A., Zumstein, K., Woodhouse, M., Bubb, K., Dorrity, M.W., Queitsch, C., Bailey-Serres, J., Sinha, N., Brady, S.M., Deal, R.B., 2018. Profiling of Accessible Chromatin Regions across Multiple Plant Species and Cell Types Reveals Common Gene Regulatory Principles and New Control Modules. Plant Cell 30, 15–36. https://doi.org/10.1105/tpc.17.00581

Majane, A.C., Cridland, J.M., Begun, D.J., 2022. Single-nucleus transcriptomes reveal evolutionary and functional properties of cell types in the Drosophila accessory gland. Genetics 220, iyab213. https://doi.org/10.1093/genetics/iyab213

Majic, P., Payne, J.L., 2020. Enhancers Facilitate the Birth of De Novo Genes and Gene Integration into Regulatory Networks. Mol. Biol. Evol. 37, 1165–1178. https://doi.org/10.1093/molbev/msz300

Moutinho, A.F., Eyre-Walker, A., Dutheil, J.Y., 2022. Strong evidence for the adaptive walk model of gene evolution in Drosophila and Arabidopsis. PLOS Biol. 20, e3001775. https://doi.org/10.1371/journal.pbio.3001775

Reinhardt, J.A., Wanjiru, B.M., Brant, A.T., Saelao, P., Begun, D.J., Jones, C.D., 2013. De Novo ORFs in Drosophila Are Important to Organismal Fitness and Evolved Rapidly from Previously Non-coding Sequences. PLOS Genet. 9, e1003860. https://doi.org/10.1371/journal.pgen.1003860

Ruiz-Orera, J., Hernandez-Rodriguez, J., Chiva, C., Sabidó, E., Kondova, I., Bontrop, R., Marqués-Bonet, T., Albà, M.M., 2015. Origins of De Novo Genes in Human and Chimpanzee. PLOS Genet. 11, e1005721. https://doi.org/10.1371/journal.pgen.1005721

Shabalin, A.A., 2012. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. Bioinformatics 28, 1353–1358. https://doi.org/10.1093/bioinformatics/bts163

Starks, R.R., Biswas, A., Jain, A., Tuteja, G., 2019. Combined analysis of dissimilar promoter accessibility and gene expression profiles identifies tissue-specific genes and actively repressed networks. Epigenetics Chromatin 12, 16. https://doi.org/10.1186/s13072-019-0260-2

Vakirlis, N., Hebert, A.S., Opulente, D.A., Achaz, G., Hittinger, C.T., Fischer, G., Coon, J.J., Lafontaine, I., 2018. A Molecular Portrait of De Novo Genes in Yeasts. Mol. Biol. Evol. 35, 631–645. https://doi.org/10.1093/molbev/msx315

Weisman, C.M., 2022. The Origins and Functions of De Novo Genes: Against All Odds? J. Mol. Evol. 90, 244–257. https://doi.org/10.1007/s00239-022-10055-3

Werner, M.S., Sieriebriennikov, B., Prabh, N., Loschko, T., Lanz, C., Sommer, R.J., 2018. Young genes have distinct gene structure, epigenetic profiles, and transcriptional regulation. Genome Res. 28, 1675–1687. https://doi.org/10.1101/gr.234872.118

Wilson, B.A., Foy, S.G., Neme, R., Masel, J., 2017. Young Genes are Highly Disordered as Predicted by the Preadaptation Hypothesis of De Novo Gene Birth. Nat. Ecol. Evol. 1, 0146. https://doi.org/10.1038/s41559-017-0146

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. https://doi.org/10.7554/eLife.47138

Zhao, L., Saelao, P., Jones, C.D., Begun, D.J., 2014. Origin and spread of de novo genes in Drosophila melanogaster populations. Science 343, 769–772. https://doi.org/10.1126/science.1248286

Supplemental figures and tables for Chapter I: New insight into the dynamics of *de novo* gene origin from increased population sampling of *Drosophila melanogaster* accessory glands

Supplemental Tables S1-S3

Lines	input reads	uniquely mapped	% uniquely mapped
R85	27173706	23504060	86.4955998
R88	28832491	27105411	94.0099522
R93	18843593	14582950	77.3894342
R153	24618581	22896608	93.005393
R217	30672531	28550648	93.0821392
R229	38043790	35329027	92.86411
R287	28635444	26680925	93.1744764
R304	34907807	31853520	91.2504186
R320	21796440	18721182	85.8910079
R338	23560470	21787431	92.4745177
R352	25352636	22616457	89.2075167
R357	17438071	15929865	91.3510732
R359	31151731	28255676	90.7033898
R360	35897966	32818709	91.4221965
R370	25762890	23221891	90.13698
R380	24529287	22701910	92.5502237
R399	24465737	21060948	86.0834399
R517	43494521	38814862	89.2408081
R530	23736061	21699710	91.4208554
R563	28840441	25442584	88.2184291
R630	34920671	31259502	89.5157542
R703	22962816	21038925	91.6217114
R761	27106437	24646844	90.9261664
R805	21917359	20342629	92.8151471
R812	24085404	21732996	90.2330557
R822	42252433	37738373	89.3164495
R850	33407564	28408483	85.0360805
R900	27269392	24704063	90.5926432
R911	26522207	24539633	92.5248529

Table S1: Mapping statistics of RNA seq.

Ral line	Blair	Cridland	Zhang (2 replicates)
399	Y	Y	Y
517	Y	Y	Y
304	Y	Y	N
357	Y	Y	N
360	Y	Y	N
208	N	N	Y
379	N	N	Y
427	N	N	Y
799	N	N	Y
85	Y	N	N
88	Y	N	N
93	Y	N	N
153	Y	N	N
217	Y	N	N
229	Y	N	N
287	Y	N	N
304	Y	N	N
320	Y	N	N
338	Y	N	N
352	Y	N	N
357	Y	N	N
359	Y	N	N
360	Y	N	N
370	Y	N	N
380	Y	N	N
530	Y	N	N
563	Y	N	N
630	Y	N	N
703	Y	N	N
761	Y	N	N
805	Y	N	N
812	Y	N	N
822	Y	N	N
850	Y	N	N
900	Y	N	N
911	Y	N	N

Table S2: DGRP lines with AG-specific transcriptomes in three different studies.

Species	lines
D ananassae	Drosophila Species Stock Center (La Jolla, CA):14024-0371.13
D mojavensis	Drosophila Species Stock Center (La Jolla, CA):15081-1352.22
D persimilis	Drosophila Species Stock Center (La Jolla, CA):14011-0111.49
D pseudoobscura	Drosophila Species Stock Center (La Jolla, CA):14011-0121.94
D virilis	Drosophila Species Stock Center (La Jolla, CA):15010-1051.87
D willistoni	Drosophila Species Stock Center (La Jolla, CA):14030-0811.24
D yakuba	Drosophila Species Stock Center (La Jolla, CA):14021-0261.01
D grimshawi	Drosophila Species Stock Center (La Jolla, CA):15287-2541.01

Table S3: Outgroup species and tissues used for transcript screening from Yang 2018. Tissues for all species included: female abdomen without digestive or reproductive system, female digestive plus excretory system, female gonad, female reproductive system without gonad, female thorax without digestive system, female whole body, male abdomen without digestive or reproductive system, male digestive plus excretory system, male gonad, male head, male reproductive system without gonad, male thorax without digestive system, and male whole body.

Supplemental Figures S1-S6

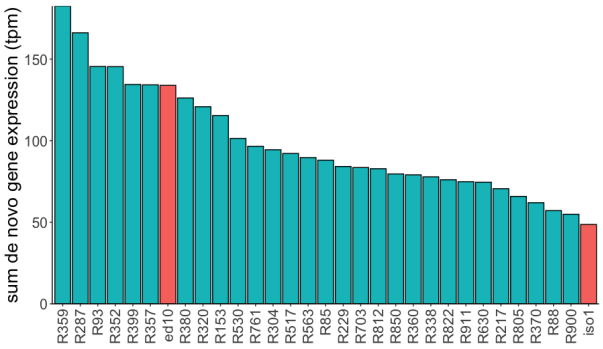


Figure S1: Number of genes expressed >1 tpm per line in *melanogaster*. Red bars indicate non DGRP lines: ed10 originates from Africa, and iso1 is the genome strain.

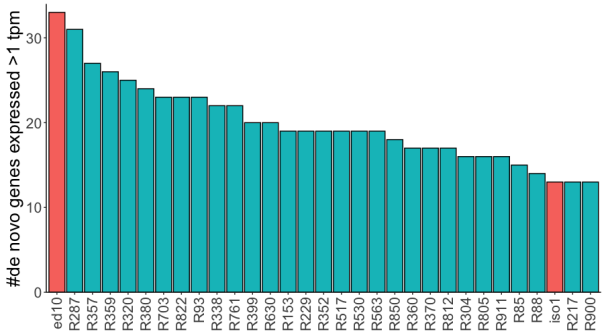


Figure S2: Sum expression of all *de novo* genes per *melanogaster* line. Red bars indicate non DGRP lines: ed10 originates from Africa, and iso1 is the genome strain.

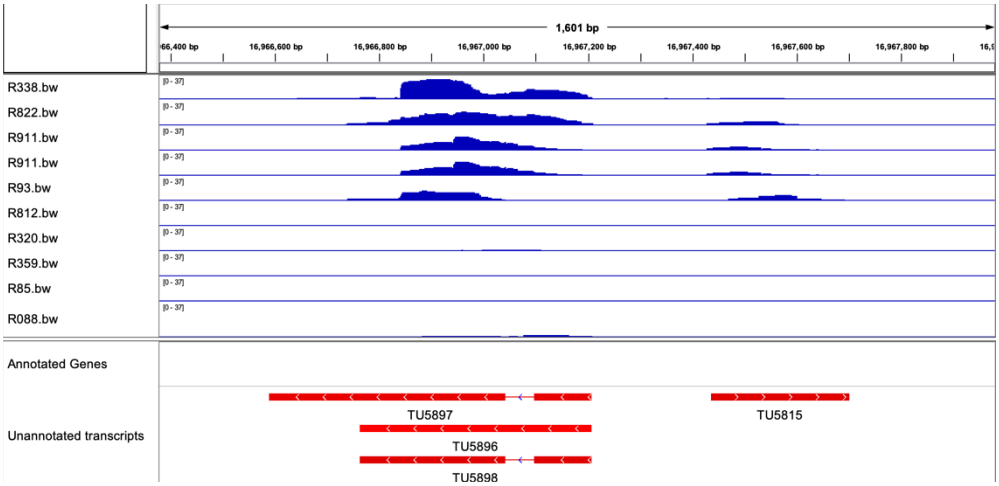


Figure S3: Example DN-DN bidirectional pair. IGV screenshot showing 5 expressed alleles and 5 non expressed alleles (in blue). "Unannotated transcripts" row shows transcripts corresponding to two separate genes (TU5897-TU5898 transcribed on Crick strand and TU5815 transcribed on Watson strand).

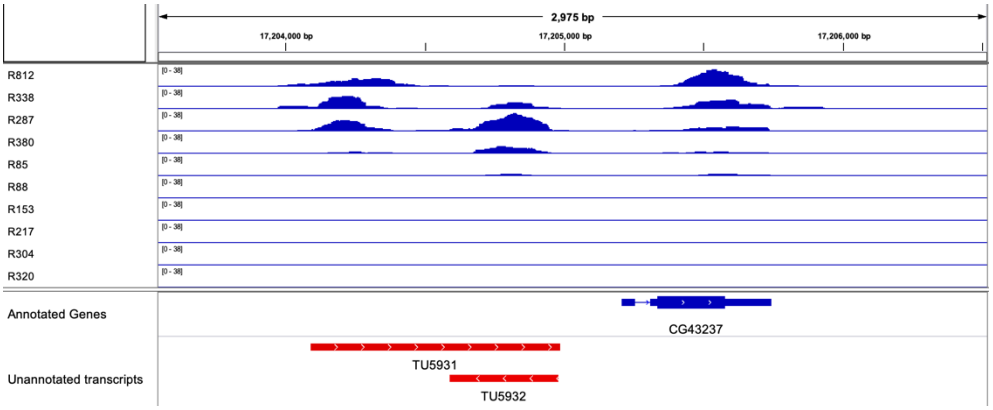
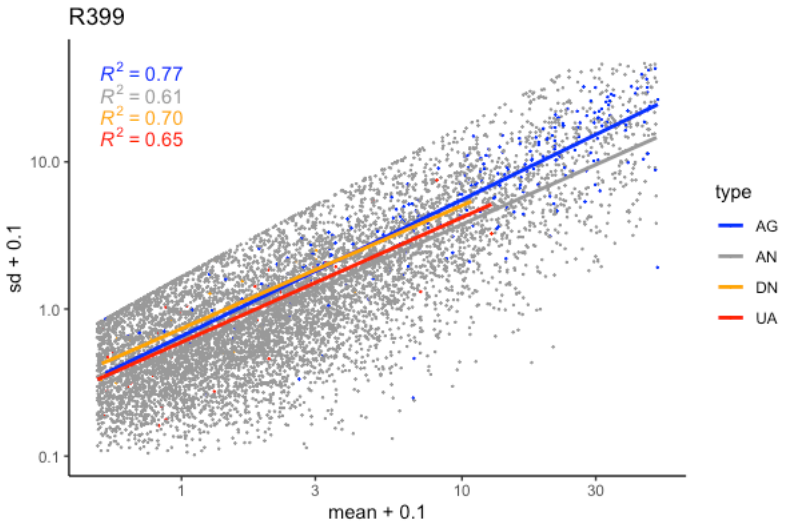
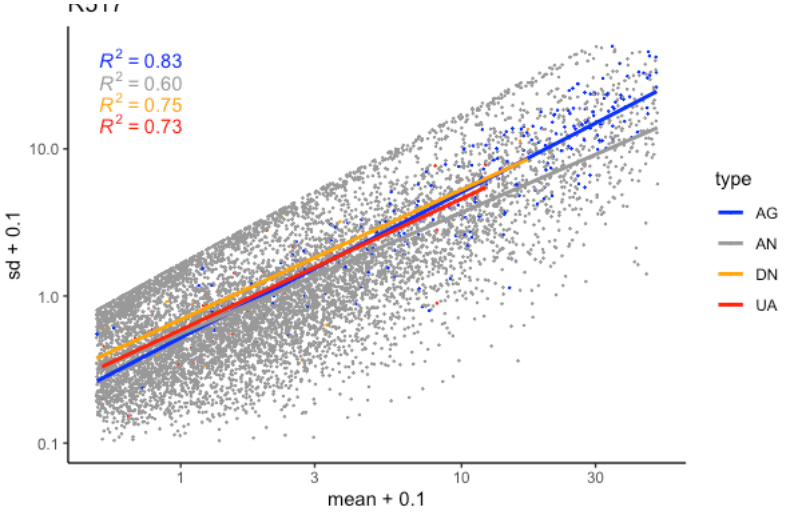


Figure S4: Example DN-annotated gene bidirectional pair. IGV screenshot showing 5 highest-expressed alleles of unannotated transcript TU5932 (expressed on Crick strand). Annotated gene expressed on Watson strand in some, but not all, genotypes.


R399
R^2 = 0.77
R^2 = 0.61
R^2 = 0.70
R^2 = 0.65
sd + 0.1
mean + 0.1
type
AG
AN
DN
UA


R^2 = 0.83
R^2 = 0.60
R^2 = 0.75
R^2 = 0.73
sd + 0.1
mean + 0.1
type
AG
AN
DN
UA

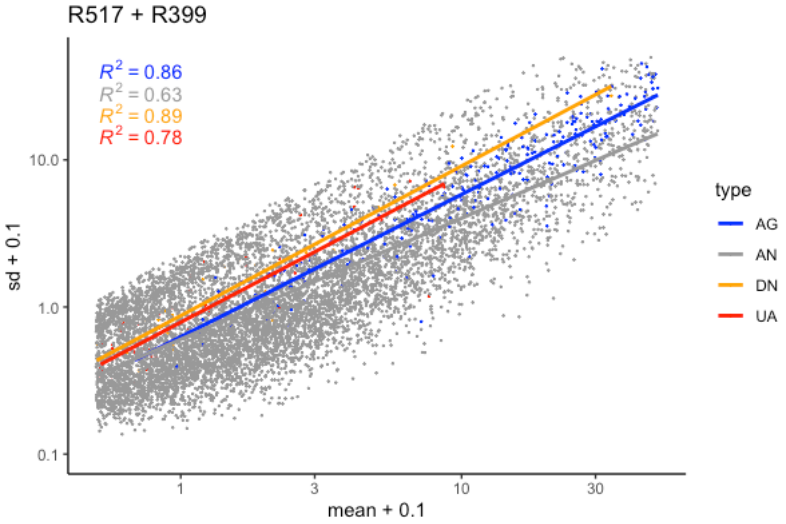

R517 + R399
R^2 = 0.86
R^2 = 0.63
R^2 = 0.89
R^2 = 0.78
sd + 0.1
mean + 0.1
type
AG
AN
DN
UA

Figure S5: Expression variance of unannotated genes is higher than AG annotated genes, when compared across both experiment and genotype. Count data taken from this study, Cridland 2022, and Zhang 2022 (first replicate only) for genotypes R399 (a), R517 (b), and both R399 and R517 (c), with mean and standard deviation calculated per gene and compared across gene classes.

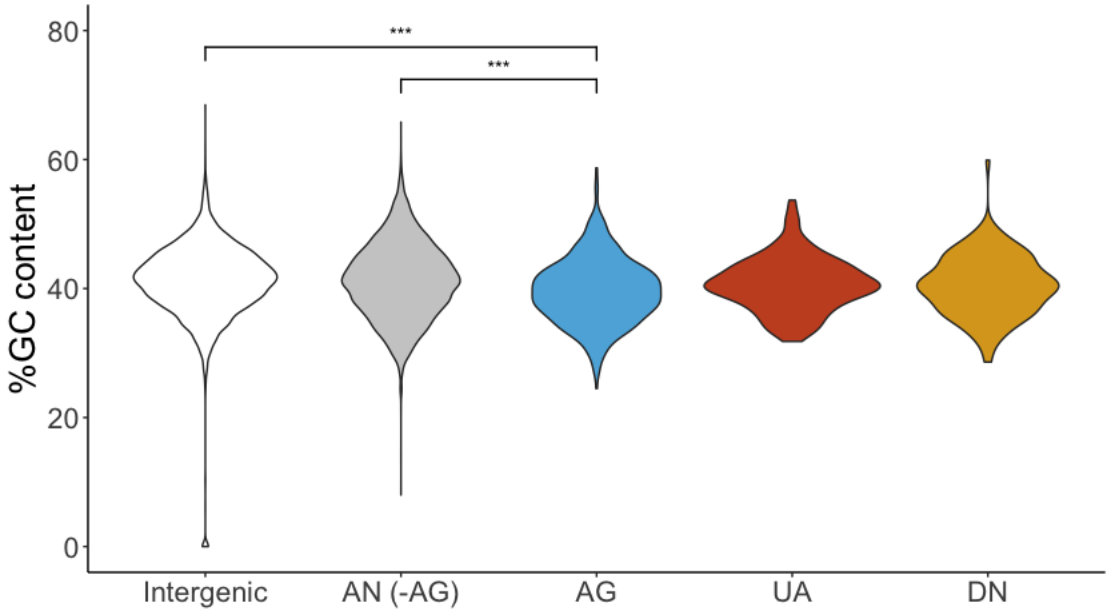


Figure S6: GC content of 1 kb regions upstream of TSS, by gene class. ***=p<0.0001, One way ANOVA with Tukey HSD. No other comparisons significant at p<0.05. Abbreviations- AN (-AG): annotated genes that expressed but enriched for accessory gland expression; AG: accessory gland enriched genes; UA: unannotated transcripts expressed in *D. melanogaster* and *D. simulans* or *D. yakuba*; DN: candidate *de novo* genes.

Supplemental figures and tables for Chapter II: Open chromatin near *de novo* genes is typically a derived change without a simple genetic explanation

Supplemental Tables S4 – S5

Sample	input reads	mapped	% mapped
R304_1*	20009103	13888318	69.41
R304_2*	22085912	14468481	65.51
R304_3*	47296456	35618961	75.31
R357_1	30100842	25046911	83.21
R357_2	42606293	40450415	94.94
R357_3	25273938	22031292	87.17
R360_1*	38939996	26058645	66.92
R360_2*	49081404	20785975	42.35
R360_3*	55213272	36987371	66.99
R399_1	32606773	30653627	94.01
R399_2	43269411	32365519	74.8
R399_3	34001095	23644362	69.54
R517_1	29009019	20451358	70.5
R517_2	28008264	25016981	89.32
R517_3	24436874	23070852	94.41

Table S4: Mapping statistics of ATAC seq. Asterisk indicates lines with Wolbachia infection.

Gene	Left primer	Right primer
rps20	CTGCTGCACCCAAGGATATT	GCGCAAGTTCTGGTTCTTTG
G9490	CTGTGTTGCGATGCTCTTTG	TGCGAGATGCTCGGATATTG
G1614	CATCGCGGATAGGTAACTCAAT	TTGCAATTGTGTGCGAGTATTT
G4151	CATCGCGGATAGGTAACTCAAT	TTGCAATTGTGTGCGAGTATTT
CR45651	AACTCACTTAGTGCCGAGAAA	TTTGTGTCCTGTGTCCTGAG

Table S5: Primers used in this study

Supplemental Figures S7 – S12

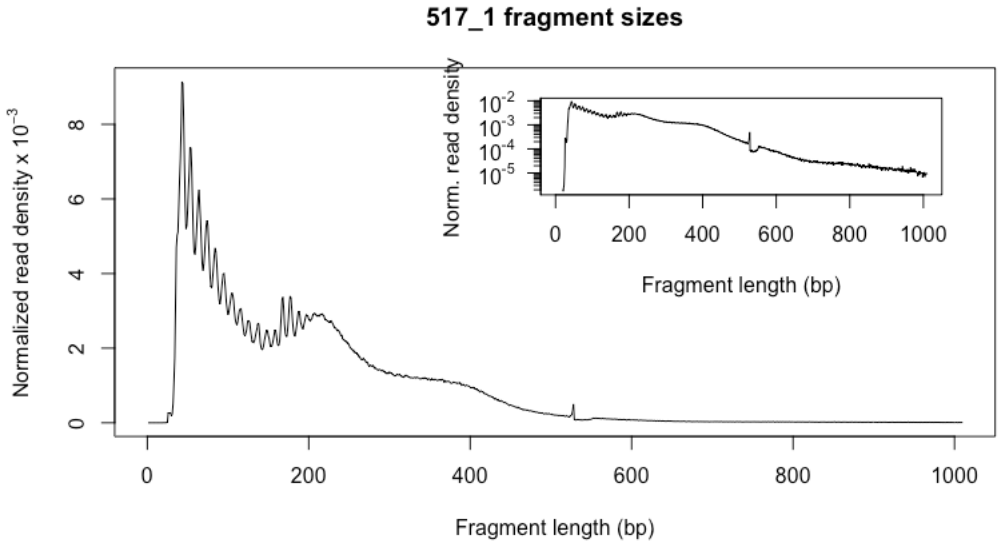


Figure S7: Representative distribution ATAC-seq fragment sizes following data processing. Short interval peaks indicate "pitch" of DNA, whereas peaks at ~200 bp and ~400bp correspond to the size of nucleosomes.

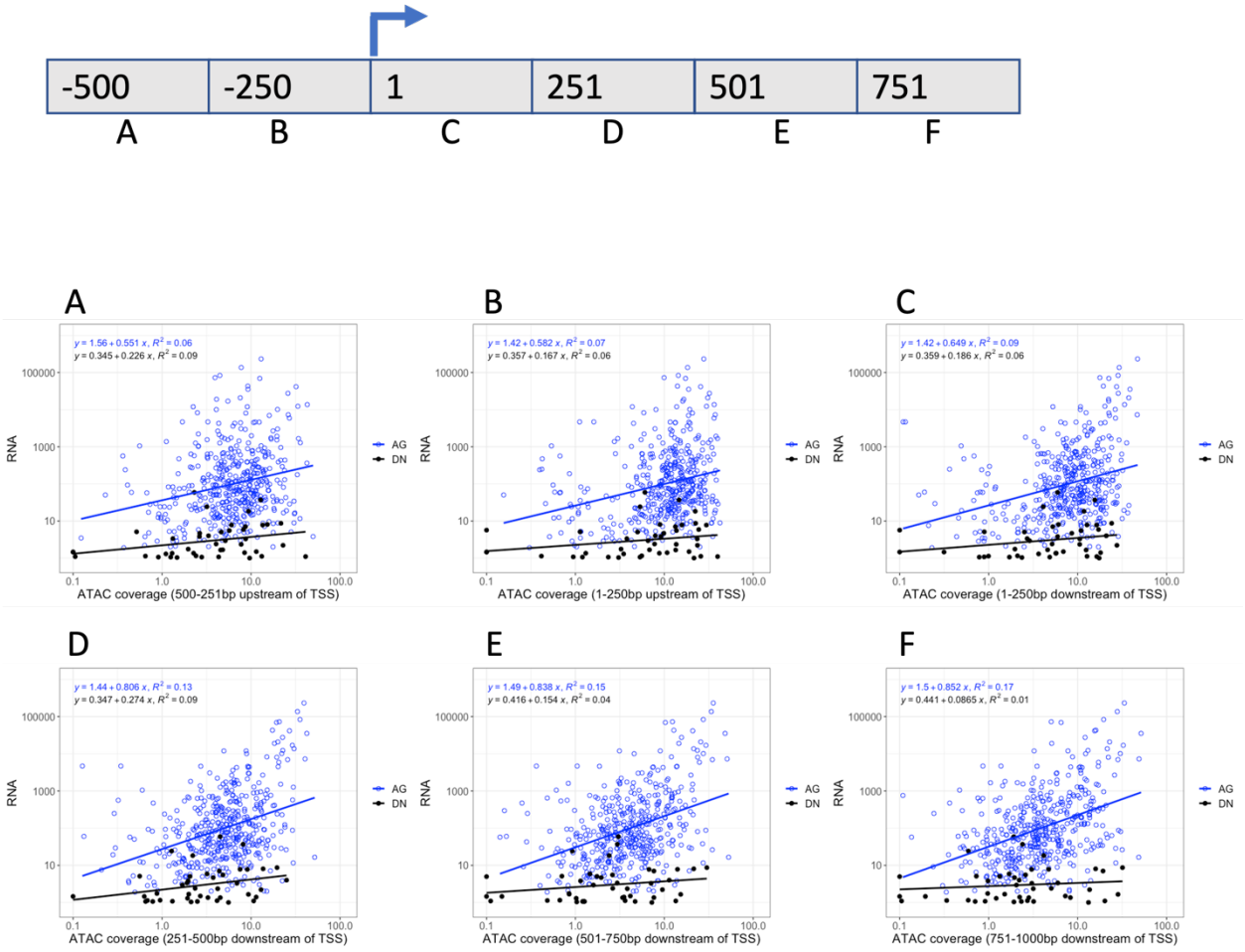


Figure S8: Relationship between ATAC-seq coverage and RNA expression, shown in non-overlapping window in region near TSS. Expression for highest-expressed line, per gene, is shown, with minimum expression >1 tpm. Relationship between ATAC and RNA seq has greater slope for AG specific genes than expressed *de novo* genes, though most AG-specific genes exhibit much higher expression. The highest correlation coefficients are seen for AG-specific genes downstream of the TSS, generally where highly-expressed genes are undergoing active transcription.

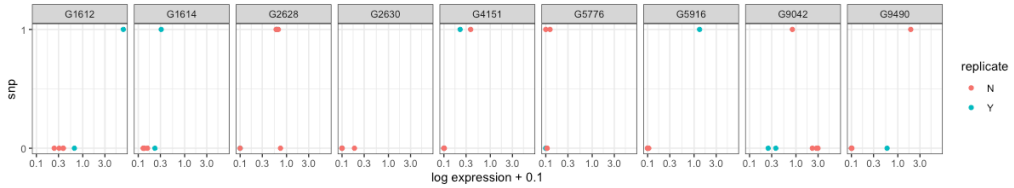


Figure S9: Associations between most-significant SNP per gene and expression dataset from Zhang 2022. "Replicate" indicates whether genotype was the same as one used in our study (2 total) or a unique DGRP line that we did not use (4 total). These data suggest the same rank-order genotype expression relationship for 5/9 genes (G1612, G1614, G4151, G5916, and G9490). No SNPs for G2630 were present in this dataset. We found 3/9 genes that did not support the importance of the most significant SNPs using this additional dataset (G2628, G5776, and G9042).

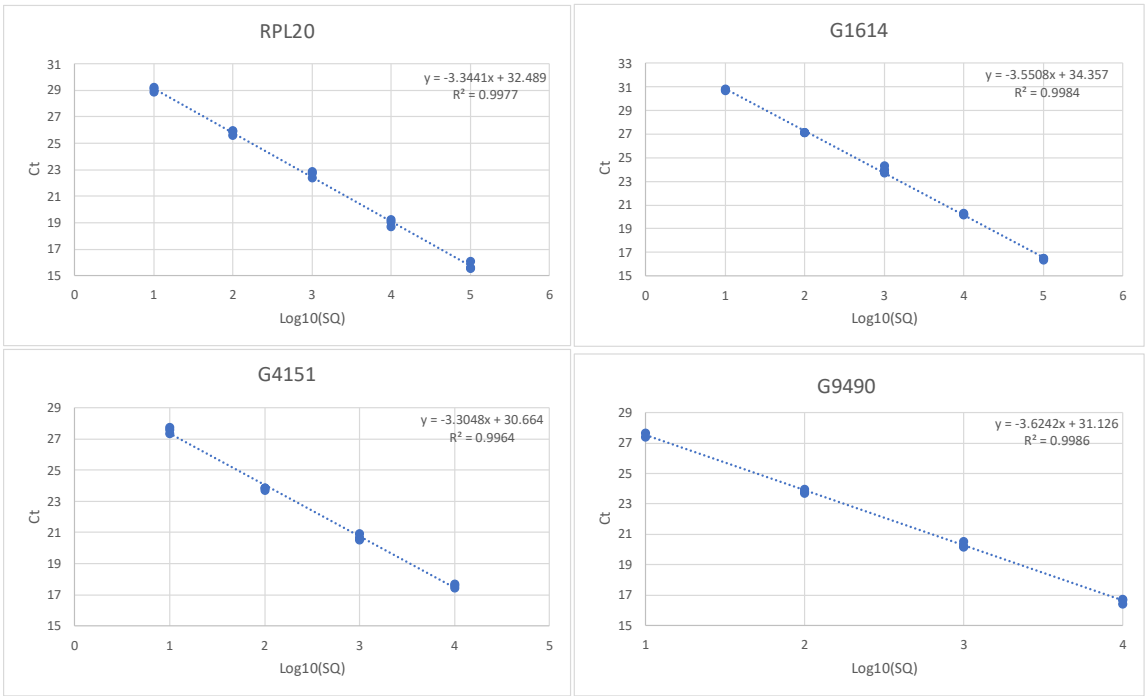


Figure S10: Standard curves for qRT-PCR primers. Calculated efficiencies were 0.95297945 (RPL20), 0.91289063 (G1614), 0.96412516 (G4151), 1.0071978, and 0.88904844 (G9490).

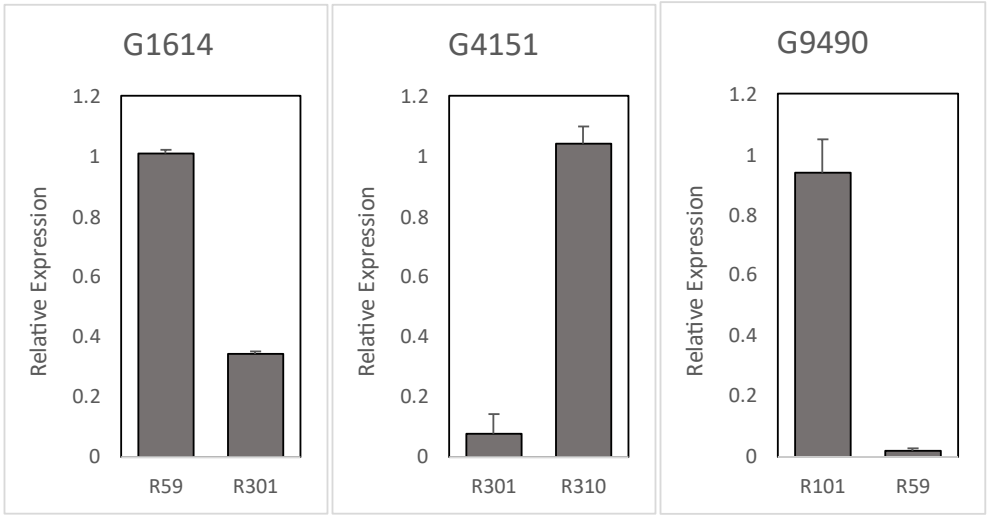
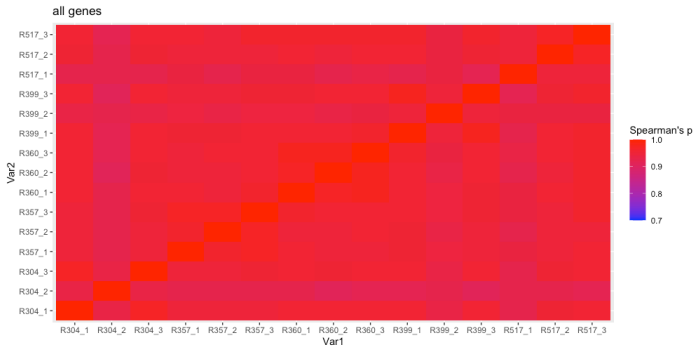
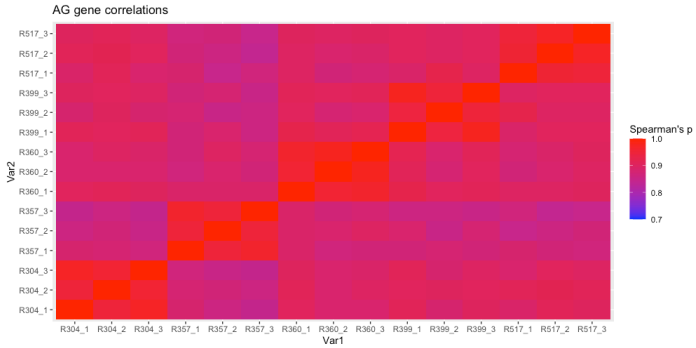


Figure S11: qRT-PCR validation of association tests. Results were significant for all three *de novo* genes tested (G1614, G4151, and G9490; student's t test $p<0.01$).

A



B



C

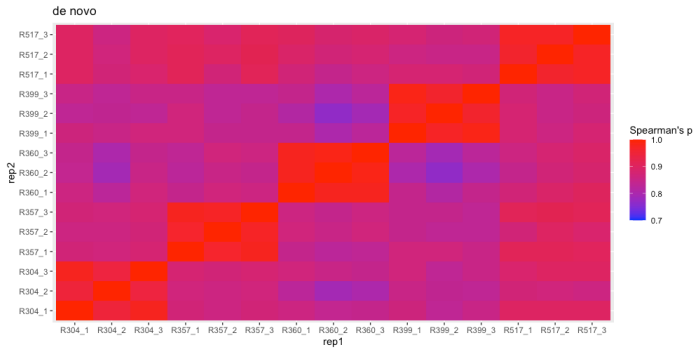


Figure S12: Pairwise correlation matrices of ATAC-seq enrichment in peaks between replicates. (A) All promoter ATAC-peaks. (B) AG-specific gene promoter peaks. (C) Promoter peaks for *de novo* genes. Replicates do not cluster in all peaks, but in looking at *de novo* gene peaks alone there is strong delineation across genotypes.