THE ORIGINS, EVOLUTION, AND FUNCTIONS OF LINEAGE-SPECIFIC GENES IN DROSOPHILA

Josephine A Reinhardt

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biology

Chapel Hill
2012

Approved by,

Dr. Christina Burch

Dr. Corbin Jones

Dr. Mohamed Noor

Dr. Lillie Searles

Dr. Christopher Willett

© 2012
Josephine A Reinhardt
ALL RIGHTS RESERVED

ABSTRACT

JOSEPHINE A REINHARDT: The Origins, Evolution, and Functions of Lineage-Specific Genes in Drosophila
(Under the direction of Dr. Corbin D Jones)

To understand how species evolve and adapt to their environments, we must understand the nature of the genetic variation causing differences between and within species. Recent studies have identified entire genes that are unique to a single species (lineage-specific genes), but little is yet known about how these genes originate or function. Here I present the results of a number of studies of lineage-specific genes in a model organism, the fruit fly, *Drosophila melanogaster*. First (Chapter two), I show that even within species, genes can greatly expand or contract in size demonstrating that novel protein domains are segregating even within a species. Secondly (Chapter three), I show that two genes that appeared to be newly evolved and lineage specific are actually rapidly evolving, and surprisingly are essential. Finally, I find that a number of genes that arose recently from non-coding sequence (*de novo* genes) are diverse in their apparent mechanism of origin, but are surprisingly similar in their gene expression pattern and functions (Chapters four and five). Like the two rapid evolving genes, the *de novo* genes I studied appear to contribute to an essential function, as their loss causes lethality. This work represents the widest molecular screen for the function of lineage-specific genes yet attempted, and reveals surprising functional similarities between these novel genes despite their diverse evolutionary origins.

Dedicated to my husband Alex Reinhardt and my parents Carol and Lynn Ziegler for being completely and unfailingly certain that I would make it.

ACKNOWLEDGEMENTS

Thanks to my adviser Corbin for keeping me hopeful when things seem hopeless, for keeping me working when nothing is working. For pushing, prodding, cajoling, suggesting, reassuring and constantly reminding me why we love what we do.

Thanks to my labmates, department colleagues, & associates for Bflag, beers, cats, dogs, webcomics. wine, trash basketball and - best of all – science.

Thanks to my collaborator Grace Lee, for making this work possible and staying cheerful, optimistic, and focused through all the revisions, and my wonderful undergraduates Teni Coker, and Betty Wanjiru, for contributing their time, sweat, brainpower and possibly their eyesight into my flies.

Thanks to Ross and Charlotte Johnson, whose contributions to UNC's Royster society made my last year of graduate school possible, and to the entire Royster society for broadening my perspective just when it was becoming hopelessly narrow.

This research would not have been possible without funding from the National Science Foundation (NSF#MCB0920196)

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS AND SYMBOLS

ACP: Accessory gland protein

bp: Base pairs

CDS: Coding sequence

EST: Expressed sequence tag

GFP: Green fluorescent protein

ORF: Open reading frame

PTC: Premature termination codon.

SCL: Stop codon loss

SCP: Stop codon polymorphism

SNP: Single nucleotide polymorphism

UTR: Untranslated region

π: Pairwise sequence polymorphism

K: Pairwise sequence divergence

CHAPTER ONE: INTRODUCTION

A conventional view of phenotypic evolution holds that mutations of small effect accumulate during adaptation, gradually leading to substantial differences between the adapted population and its ancestor. Gene duplications provided the necessary material for the evolution of more profound genetic innovations (Ohno et al. 1968; Ohno 1970). As more molecular data have become available, however, some of these assumptions are being questioned. Genes and genomes were originally sequenced in order to facilitate functional genetic analyses in model organisms. Evolutionary geneticists quickly realized that the comparison of entire genomes and gene complements represented a new approach to understanding the genetics of adaptation and speciation. The first available molecular evolutionary data came from highly conserved proteins (e.g. hemoglobin). This led to the conclusion - consistent with Darwnian gradualism - that molecular evolution generally occurs through slight perturbations of existing proteins. Careful analyses of additional genomes has altered this perception - extreme changes in genes and genomes do occur, and surprisingly often. For example, sex chromosomes have originated numerous times in multiple lineages (Ellegren 2011), and new genes arise through mechanisms other than simple duplication (Long et al. 2003, and see Figure 5.1).

Perhaps the most surprising finding of all was that species contain genes that have no clear homologs, even in close relatives. These so-called orphan genes (Schmid and Aquadro 2001; Wilson et al. 2005), also sometimes called lineage-specific genes, could have originated through multiple means. Most trivially, genes may be misannotated, or may only appear novel because no sufficiently close relative has been sequenced. Indeed, many of the first orphans identified were subsequently discovered in newly sequenced genomes, and their “orphan” status revoked (Schmid and Aquadro 2001). Some suggest that all orphan genes may be rapid evolving, changing so fast that orthology becomes obfuscated even in close relatives (Schmid and Aquadro 2001; Cai et al. 2006).

In 2006, Levine and colleagues (2006) proposed that a set of lineage-specific genes in Drosophila had originated *de novo* from non-coding sequences. These genes were restricted to *Drosophila melanogaster* alone or only to *D. melanogaster* and its closest relatives, *D. simulans* and *D. sechellia*. In *D. yakuba* and all other sequenced species, the genes were apparently disabled (e.g. they included in frame stop codons or were deleted in their entirety). As more organisms were sequenced, so-called *de novo* genes were described in other taxa (reviewed in Kaessmann 2010 and Tautz and Domazet-Loso 2011), though the definition of *de novo* genes differed between studies.

In general, three computational methods have been employed to mine *de novo* genes from genomic data. The first was to use BLAST or other alignment tools to compare annotated genes in the focal species to all other genomes, and to identify genes whose sequences are completely novel - in other words, “true” orphans. This method was employed in flies (Levine et al. 2006; Begun et al. 2007b; Zhou et al. 2008; Li et al. 2010a, yeast Cai et al. 2008; Xiao et al. 2009), primates (Toll-Riera et al. 2009; Li et al. 2010a), and paramecium (Yang and Huang 2011). A second approach (Heinen et al. 2009) scanned EST databases for sequences that were transcribed in mice but lacked transcription data in human or rat, hence describing *de novo*-ness as lineage-specific transcription from well-conserved sequences. In the final method, (Knowles and McLysaght 2009; Wu et al. 2011) genes that had a human-specific stop codon loss/start codon gain in otherwise conserved putative non-coding sequences were identified - in other words, defining *de novo* genes as lineage-specific expansions of shorter open reading frames. The key difference between these methods is that the latter two require sequence conservation of the potential ORF, whereas conservation is specifically required *not* to exist for the first approach. Hence, the first method and the latter two are mutually exclusive and could in principle be described as different sets of genes. All methods identify transcripts or proteins that are by some description lineage-specific and evolutionarily novel.

De novo genes are intriguing because they are completely unique to one or few closely related species - providing the tantalizing possibility that they could be involved in lineage specific phenotypes. However, their novelty alone says nothing about whether they have or even could contribute to adaptation. It might be that these so-called genes are only accidentally expressed and are non-functional or even mal-adaptive - that is, dispensable evolutionary accidents. Functional studies of these and other types of new genes suggests otherwise. The yeast *de novo* gene, for example, was shown to be synthetic lethal and most likely involved in DNA repair and mating (Cai et al. 2008; Li et al. 2010b). Meanwhile, a surprising number of Drosophila novel genes of various types (including the *de novo* gene *CG31406*) were found to be essential in an RNAi screen in (Chen et al. 2010).

Expression data across many studies shows one striking pattern. Novel genes in several groups of mammals - regardless of their mode of origin or putative function - are often expressed in the testes. Levine and colleagues noted this pattern among *de novo* genes, noting that the testes may be a common birthplace for *de novo* genes (Levine et al. 2006). Further, mice carrying a knockout of *Poldi*, a mouse-specific *de novo* gene, showed lowered testes mass and sperm motility, implying that male expression may also contribute to male function. Kaessmann (2010) proposed a mechanism to explain this phenomenon. Expression of genes during pre-meiotic development in the testes that are usually not expressed could be the first step in establishing transcription of a novel gene: the so-called "out of the testes" model for novel gene evolution.

Two important questions about *de novo* genes remain unclear despite the progress that has been made across species - how do these genes arise, and why do they persist? The question of "why" is intimately connected to the function of *de novo* genes. As noted above, some *de novo* genes may be essential and/or important to fertility, but the functions these genes are performing remains vague. How these genes arose is also unclear. It has been speculated that *de novo* genes may begin as non-coding RNAs or other transcribed sequences ("transcription first" model). Alternatively, a nascent coding region may evolve and later acquire transcription (e.g. through the recruitment or evolution of a promoter - "CDS first" model). Most previous work on *de novo* genes failed to confirm whether orthologous non-coding sequences were expressed. Studies using focal species other than Drosophila relied exclusively on sometimes-sparse EST data and on Northern Blots using only probes from the focal species. Expression of a homologous sequence in other species does not necessarily rule out *de novo* gene evolution; it may simply imply that an RNA gene intermediate is common. However, if such expression seldom occurs before a stretch of contiguous coding sequence evolves, this would suggest that the advent of the CDS is a critical first step in the evolution of the *de novo* gene.

In order to answer these questions systematically, I have undertaken functional, molecular evolutionary, and population genetic studies of several *de novo* genes within *Drosophila melanogaster*. Chapter two describes a mechanism of how *de novo* genes might arise within populations through relatively small changes (loss of a stop codon). The third through fifth chapters are functional and molecular evolutionary studies of putative *de novo* genes. The genes described in chapter three - while extremely dissimilar in sequence to even very close relatives - are almost certainly simply rapidly evolving. Yet their rate of evolution is astonishingly fast - their protein sequences are as unique to *D. melanogaster* as any “true” orphan. The fourth chapter describes a set of genes that are in whole or part specific to *D. melanogaster* and its relatives. Surprisingly, RNAi knockdown of some of these genes led to total lethality, implying these genes are essential. The genes are expressed in the testes and also in developing larvae, lending credence to Kaessmann’s (2010) “out of the testes” hypothesis for the evolution of novel functions for new genes. Finally, chapter five describes two putative protein-coding genes that apparently have evolved from previously non-coding RNAs. One of these is apparently essential, though we hypothesize that the essential function is more likely due to the function of the older RNA gene than the new ORF. Together with chapter four, this result implies that the evolutionary trajectories of *de novo* genes are diverse, even while disrupting their function leads to surprisingly similar results.

We can only come to a complete understanding of how organisms adapt if we fully understand the nature of the genetic variation underlying adaptive phenotypic changes. Recent work in multiple species has made it clear that gradual models of molecular evolution do not describe the full spectrum of genomic variation that is possible during the adaptive process. My work further shows that even the newest genes are essential and thus contribute to the adaptation of organisms to their internal and external environments.

CHAPTER TWO: WIDESPREAD POLYMORPHISM IN THE POSITIONS OF STOP CODONS IN *DROSOPHILA MELANOGASTER*

Authors: Yuh Chwen G Lee and Josephine A Reinhardt

Published in Genome Biology and Evolution, accepted on November 8, 2011

ABSTRACT

The mechanisms underlying evolutionary changes in protein length are poorly understood. Protein domains are lost and gained between species, and must have arisen first as within species polymorphisms. Here we use *Drosophila melanogaster* population genomic data combined with between species divergence information to understand the evolutionary forces that generate and maintain polymorphisms causing changes in protein length in *Drosophila melanogaster*. Specifically, we looked for protein length variations resulting from premature termination codons and stop codon losses. We discovered that 438 genes contained polymorphisms resulting in truncation of the translated region (premature termination codons) and 119 genes contained polymorphisms predicted to lengthen the translated region (stop codon losses). Stop codon polymorphisms (especially premature termination codons) appear to be more deleterious than other polymorphisms, including protein amino acid changes. Genes harboring stop codon polymorphisms are in general less selectively constrained, more narrowly expressed, and enriched for dispensable biological functions. However, we also observed exceptional cases such as genes that have multiple independent stop codon polymorphisms, alleles that are shared between *D. melanogaster* and *D. simulans*, and high frequency alleles that cause extreme changes in gene length. Stop codon polymorphisms likely have an important role in the evolution of these genes.

INTRODUCTION

Genetic variation in natural populations has long been a source of interest to both population biologists and functional geneticists, and the genus *Drosophila* has been a system of choice for describing such variation (Timofeeff-Ressovsky 1927; Timofeeff-Ressovsky 1930; Dubinin 1937; Ives 1945; Spencer 1947). Natural variations allow one to infer patterns of gene flow, migration and selection. In addition, alleles discovered in natural populations have been used as tools to elucidate molecular mechanisms of specific phenotypes - e.g. meiotic mutants from natural populations Sandler et al. 1968. More recently, effort has focused on determining what specific genetic changes have led to adaptation between and within species. One hotly debated question is whether protein sequence, copy number, or gene regulation are more likely to be the genetic basis of adaptation (Prud'homme et al. 2007; Wray 2007; Emerson et al. 2008). However, the evolutionary role of genetic variants that lead to deviations from annotated gene models - such as the position of initiation codons, splicing junctions and stop codons - has received relatively little attention considering the potential impact of such variants on gene function. On the other hand, alleles containing premature termination codons (PTCs) are well characterized as the genetic cause of many human diseases including retinosis pigmentosa (Chang and Kan 1979; Rosenfeld et al. 1992) and beta-thalassemia, and so have been of particular interest to the genetics of human disease.

We expect the functional consequences of stop codon polymorphisms (SCPs) –in particular PTCs - on the affected gene to be at least as severe as those caused by nonsynonymous mutations and more severe than those caused by synonymous mutations. This is because transcripts of genes carrying PTCs are expected to undergo nonsense-mediated decay (Chang et al. 2007), which results in loss of gene expression and function. In humans, stop codons occurring more than 50 bases prior to the final exon-exon junction are silenced by nonsense-mediated decay (Nagy and Maquat 1998). This process occurs in all organisms in which it has been studied (Chang et al. 2007) but the trigger for nonsense-mediated decay is not as clear in other organisms as it is in humans (Gatfield et al. 2003; Behm-Ansmant et al. 2007). If transcripts harboring PTCs are not targeted by nonsense-mediated decay, they will likely still be deleterious because of the loss of 3' protein domains or dissociation from 3' untranslated region regulatory elements. The stop codon of a transcript may also be lost (stop codon loss, SCL), leading to either down-regulation of expression through the non-stop decay pathway (Vasudevan et al. 2002) or an expansion of the open reading frame. Non-stop decay results in post-transcriptional degradation of transcripts without an in-frame stop codon prior to the polyadenylation signal and is conserved throughout eukaryotes (Gatfield et al. 2003). SCLs that are not silenced could acquire novel downstream structural or regulatory sequence elements that might alter protein expression or function.

The length of the open reading frame of genes has clearly changed over evolutionary time (Yandell et al. 2006). It has been observed that divergence in the length of coding regions is disproportionately found at the beginnings and ends of genes (Bjorklund et al. 2005; Weiner et al. 2006), with the latter possibly caused by either loss or gain of the stop codon. The fact that we observe such changes between species implies that they must first arise as within species polymorphisms. But where do these SCPs first arise within populations and genomes, and what is their evolutionary fate after they arise? Previous work has documented the number and frequency of polymorphisms causing changes in the position of termination codons in humans. Yamaguchi-Kabata and colleagues (2008) used human dbSNP data (Sherry et al. 1999) and found 1,183 SNPs resulting in PTCs, 581 of which were predicted to trigger nonsense-mediated decay and were thus annotated as null alleles. They also observed 119 polymorphisms causing SCLs, which typically led to short expansions of the open reading frames. SCPs were found at a lower density (polymorphic site per mutable site) than nonsynonymous amino acid changes, implying that stop codon polymorphisms are more likely to be deleterious than changes to amino acid sequence. Another study (Yngvadottir et al. 2009) genotyped a subset of the SNPs reported above in order to measure allele frequency of these SNPs, and confirmed that PTCs were generally at low frequency and evenly distributed within the coding region of the proteins they were found in. Finally, the 1000 human genomes project has catalogued additional PTCs in the human population (Durbin et al. 2010).

The population genetics of null alleles has long been an area of interest in Drosophila (Voelker et al. 1980; Langley et al. 1981; Burkhart et al. 1984), and a number of individual stop codon polymorphisms have been described and characterized in detail (Begun and Lindfors 2005; Lazzaro 2005; Kelleher and Markow 2009). However, stop codon polymorphisms have not yet been described in Drosophila on a genome-wide scale. This analysis will provide a useful contrast to human data in an experimentally tractable organism, providing a unique opportunity to determine the functional importance and fitness impacts of SCPs observed from natural populations. With the advent of next generation sequencing technology, we are now able to describe thousands of natural variants in Drosophila simultaneously. Recently, 37 whole genomes from a population of *D. melanogaster* near Raleigh, North Carolina, USA (RAL) and seven genomes from a population in Malawi, Africa (MW) have been resequenced as part of the Drosophila Population Genomics Project (www.dpgp.org). Additionally, six genomes of *D. melanogaster's* close relative, *D. simulans* (Begun et al. 2007a), and ten other species of Drosophila (Clark et al. 2007) have been sequenced and annotated, providing a wealth of data for inferring the evolutionary history of within-species variation. In contrast to previous surveys of natural variants (e.g. Sandler et al. 1968), the described alleles from these 44 *D. melanogaster* genomes are preserved in living stocks, which can be rapidly leveraged towards functional work. This represents an unparalleled resource for answering questions about the origin, maintenance and functional impact of natural variants on a genomic scale.

Here, we describe one type of variation that was uncovered in the DPGP sequencing project: SNPs that cause changes in the position of the stop codon (SCPs). Our observations generally supported our *a priori* hypothesis that the sampled SCPs are, as a group, selected against, and generally more deleterious than other types of SNPs. However, we did find a number of alleles that were exceptions to this pattern, such as alleles that have been segregating since before the split between *D. melanogaster* and *D. simulans*, high frequency derived alleles and alleles at high frequency despite causing large changes in the original gene model. We also found 56 genes carrying more than two alleles with different stop codon positions. The evolution of these genes may be strongly affected by changes in gene model. Furthermore, the alleles described in this study are available in living stocks, providing an opportunity to directly measure the phenotypic consequences of stop codon variation.

MATERIAL AND METHODS

Characterizing stop codon polymorphisms within *D. melanogaster*

We used FlyBase version 5.16 for gene model annotations (Tweedie et al. 2009), giving a total of 14,072 annotated protein-coding genes. For each gene model, we searched through the 44 *D. melanogaster* genomes from DPGP (www.dpgp.org) and identified alleles with canonical (the same as the reference annotation) initiation codons and splice junctions but either premature termination codons (PTC, stop codon appearing before the canonical stop codon) or stop codon losses (SCL, loss of stop codon at the canonical stop codon position). For genes with more than one isoform, we determined which isoforms were affected. If the genomic position of the premature termination codon or lost stop codon was the same for multiple isoforms, we only considered the isoform with the longest coding region when calculating statistics on the changes in gene model. For alleles with a stop codon loss and an annotated 3' untranslated region, the "expanded region" for a given allele was defined as the downstream transcribed sequence until one of three features was encountered: 1. an in-frame stop codon, 2. an uncalled ("N") base that would have been an in-frame stop codon assuming the genome matched the reference genome, 3. the end of the known transcribed region. If no stop codon was encountered before an annotated polyadenylation site, the gene was labeled as a target of non-stop decay and was not included in the expansion length analysis. In total we predicted four alleles to undergo non-stop decay, and could not confirm non-stop decay status for another 90 alleles because there was no 3' untranslated region sequence data available or the region did not contain a polyadenylation site.

We estimated the density of SCPs following Yngvadottir *et al.* (2009)**.** Density represents the proportion of sites in which a SNP resulted in a PTC or SCL. The density of PTCs is the number of observed PTCs divided by 2,387,149 - the number of sites in the genome that can mutate directly to a stop codon (one mutable site for all codons in annotated genes that are one mutational step away from one of the three stop codons except TGG, which has two mutable sites). For SCLs, the density is the observed number of SCLs divided by the number of unique stop codons across all the isoforms of all genes (total 42,315 sites).

For each allele, we counted the number of lines agreeing or disagreeing with the reference annotations for North American (RAL) and African (MW) populations respectively. It is worth noting that while all the major chromosomes of the 37 North Carolina strains were sequenced, only some chromosomes were sequenced from each of the nine Malawi strains. This resulted in seven first (X), six second (2L and 2R) and five third (3L and 3R) chromosomes in the Malawi population. We then polarized each allele with respect to the major allele across both populations. We considered polarizing with respect to the ancestral state, but were concerned that this would bias against alleles of rapidly evolving genes. This is because the ancestral state of these genes are harder to determine due to the poor alignment between distantly-related species (*D. yakuba/D.*

erecta) with fast evolving sequences. Indeed, we found nearly half of our alleles dropped out of the analysis when we included *D. yakuba* and *D, erecta* as the outgroup lineage (see results). The reference genome carried a minor PTC allele in 8 cases and a minor SCL allele in 13 cases (designated "Reference Minor"). We contrasted the frequency of SCPs with that of nonsynonymous and synonymous SNPs from the DPGP dataset. Each DPGP assembly is missing data for some lines due to assembly problems and/or low sequencing quality. We controlled for the resulting variation in allelic coverage (the number of genomes at each site that have data) by removing sites with allelic coverage below 20 from the entire dataset. Allele frequency for each polymorphism was estimated as the proportion of the minor allele among all available alleles. We also used maximum likelihood methods to estimate the number of minor alleles if all alleles were available, assuming minor allele counts have a hypergeometric distribution. This method was only applied to Raleigh and all *D. melanogaster* samples. Our observations were consistent between the two methods and we only present the former. Among the annotated genes in version 5.16, there are 16 genes with premature termination codons in the annotated coding regions of the reference annotations and 12 genes that do not have a stop codon at the end of the reference annotated translated region. These 28 genes were excluded from our analysis (Table 2.1).

Error/sequence quality control

Release 1.0 of the DPGP data consists of fastq files, where the quality score of a base is derived from the quality of the Illumina reads covering that base and the quality of the consensus assembly at that base. The scoring system is based on Phred quality scores

Table 2.1 – Reference-specific excluded alleles

Gene-Isoform	Type	Position of fixed stop along annotated CDS (Flybase v5.19)	Codon fixed in the 50 genomes
CG11891-PD	ref-no stop	343	TAA
CG11891-PE	ref-no stop	343	TAA
CG1867-PA	ref-no stop	613	TAA
CG2698-PB	ref-premature	NA	AGG
CG5028-PC	ref-premature	NA	AGA
CG5192-PB	ref-no stop	706	TAG
CG6633-PB	ref-no stop	250	TAG
CG9611-PD	ref-premature	NA	AGA
CG10948-PB	ref-premature	NA	GGA
CG32042-PD	ref-premature	NA	GGA
CG32382-PB	ref-premature	NA	GAA
CG32383-PB	ref-premature	NA	GAA
CG42268-PG	ref-premature	NA	GAA
CG5747-PB	ref-premature	NA	AGA
CG9111-PA	ref-no stop	127	GTC
CG14047-PF	ref-no stop	4444	TAG
CG34143-PC	ref-no stop	1516	TAG
CG3757-PA	ref-no stop	148	TAA
CG40305-PB	ref-no stop	580	TGA
CG5227-PA	ref-no stop	664	TGA
CG5227-PB	ref-no stop	664	TGA
CG5227-PC	ref-no stop	664	TGA
CG5227-PD	ref-no stop	664	TGA
CG6121-PA	ref-premature	NA	AGA
CG13784-PB	ref-premature	NA	AGA
CG17377-PB	ref-premature	NA	GGA
CG31774-PA	ref-premature	NA	GGA
CG10245-PB	ref-no stop	250	TGA
CG1555-PA	ref-no stop	1519	TGT
CG16747-PA	ref-no stop	217	TAG
CG16747-PB	ref-no stop	199	TAG
CG16747-PC	ref-no stop	265	TAG
CG17632-PA	ref-no stop	1678	TGA
CG33964-PA	ref-no stop	322	TAA
CG9415-PC	ref-no stop	964	TGA

Note: ref-premature means that the 50 genomes lack a stop codon where the reference allele has a stop codon. Ref-no stop means that the reference lacks a stop codon where the 50 genomes have one.

where a score of Q50 indicates an estimated 1/100,000 error rate (Ewing et al. 1998). We used SNPs surrounding our stop codon polymorphisms to estimate the expected distribution of quality data for SNPs. The median of the distribution is Q50 with scores ranging from a minimum of Q30 to a maximum of Q74 (Figure 2.2, orange).

We calculated above that 2,387,149 bp could mutate to become a stop codon. Since we are calling bases independently across an average of 44 genomes, this is a total of approximately 103 million bases that could become stop codons. If all of these bases had the median Phred quality score of Q50 (1/100,000 error rate), then 1,026 bases would be incorrectly called as premature termination codons. We used an empirical distribution of quality scores for polymorphic bases in the 44 genomes to determine how many total false positive mutations would be expected. Given this distribution, we expect about 7,026 false positive premature termination codons would be called, and most of them (5,095) would have quality scores below Q40. Given that we actually observed 2,104 PTCs across the genomes (across all quality scores), the error rate is likely lower than calculated above. We found that premature termination codon SNPs had lower quality scores than other SNPs, with an apparent excess of SNPs with quality scores less than 40 (Figure 2.2A). However, the experimental distribution of quality scores for observed SCPs is not consistent with the expectation if most of the SCPs called were errors (Figure 2.2B), implying that most SCPs we have observed are not errors.

We went on to sequence 73 alleles chosen randomly from the different quality score classes, and found that alleles with a quality score below 40 were usually false positives (3/12 alleles were validated). Conversely, alleles with a quality score of 40 or greater were validated the majority of the time (29/46), and alleles with quality above 60

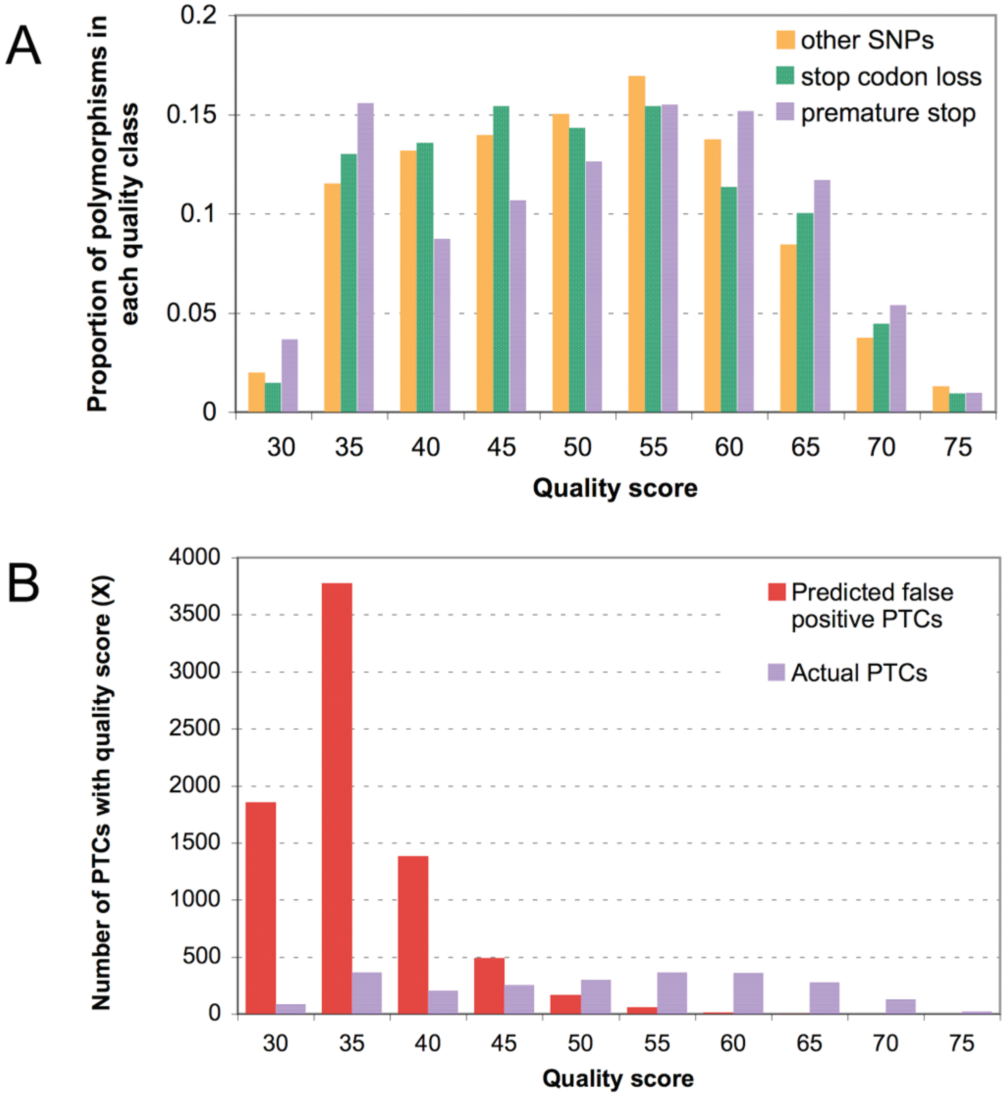


Figure 2.1 - Sequencing error is not driving the appearance of SCPs.
(A) The proportion of different types of polymorphisms from the DPGP 50 genomes project (www.dpgp.org) falling into each quality class. The median quality score was 50. PTCs (violet) are more likely than SCLs (green) or other SNPs (orange) to have low quality scores. (B) The predicted number of false positive premature termination codons in each quality class based on the empirical distribution of quality scores (red) compared with the actual number of called premature termination codons in each quality class (violet).

were validated in 12/15 cases. We decided to pursue the remaining analyses having discarded all alleles with a quality score below 40, as these alleles were mostly erroneous. Although some of the remaining alleles are likely false positives, we are confident that the majority of them are true positive. We also did all analyses with a more conservative dataset in which polymorphisms called in only a single line were removed. As each SNP is called using an independent sequencing dataset in each line, it is extremely unlikely that the same error would be found in more than a single line provided that error is random. Further, we found there was no difference in quality scores between nonsingleton and singleton alleles, implying no obvious bias in base calling across genomes. We used the same quality score cutoff (Q40) for other types of SNPs (nonsynonymous, synonymous, non-coding) that were used to contrast with SCPs.

Phylogenetic analyses

For each SCP, we asked whether the sequenced genome of other Drosophila species in the *melanogaster* subgroup shared the major or minor allele from the *D. melanogaster* population. This allowed us to infer which of the extant *D. melanogaster* alleles are newly derived *versus* shared with *D. simulans*. We used the multi-species whole-genome alignment (*D. melanogaster, D. simulans, D. yakuba* and *D. erecta*) created for the DPGP genome project (Langley CH, *personal communication*). In order to infer the age of the origin for an SCP, we required that a base that is polymorphic in *D. melanogaster* has data available (not an "N" or deletion) in at least one *D. simulans* genome and either the *D. yakuba* or *D. erecta* genomes in the multi-species alignment. For analyses using an outgroup, we used the *D. yakuba* data if available, and if it was not available, we used the *D. erecta* data. Alleles without sufficient data were designated as "missing data." If fixed, the *D. simulans* allele was required to be the same as the *D. yakuba/D. erecta* allele. Nineteen alleles that violated this rule were given the designation of "ambiguous history." The age of the remaining alleles were annotated (with the assumption of only a single mutation leading to the allele) as having arisen in the ancestor of *D. simulans* and *D. melanogaster*, in the ancestor of the two *D. melanogaster* populations, or in one of the two *D. melanogaster* populations. We also asked whether the major allele currently found in *D. melanogaster* was the ancestral allele or the derived allele by comparing it to *D. yakuba/D. erecta*. Finally, we asked whether any genes were segregating with two or more alleles in one or both of the *D. melanogaster* populations and/or the *D. simulans* populations. A caveat of using *D. yakuba* and *D. erecta* as an outgroup is that fast evolving genes have a higher chance of being misaligned in the multi-species genomic alignment and thus removed from the analysis. We performed above analyses and tests without using the *yakuba-erecta* outgroup to polarize the changes - that is we only asked if an allele was specific to one or both *D. melanogaster* populations, or if it was shared with *D. simulans*. Our observations were insensitive to whether or not we use the *D. yakuba/D. erecta* clade to polarize the direction of the changes. To contrast the age of SCPs to other SNPs, we used the same criteria to classify nonsynonymous and synonymous polymorphism into different age classes.

When comparing the number of SCPs unique to either MW or RAL populations, we made the following correction. It has been demonstrated that the expected number of polymorphic sites observed from a sample of size n is proportional to $\sum_{i=1}^{n-1} 1/i$ (Watterson

1975). Accordingly, in the *Chi-square test* table, we applied this correction to the sample size of each population.

Population genetics analyses of genes with stop codon polymorphisms

The GC content of each gene was estimated as the proportion of GC bases in the coding regions annotated in the reference *D. melanogaster* genome. The Codon bias index Fop (the percentage of preferred codons) was estimated with CodonW (Peden 1999). We used PAML (version 4, Yang 1997) to estimate the lineage-specific substitution rate on the *D. melanogaster* and *D. simulans* lineage, using *D. yakuba* as the outgroup. For each gene, we used the two *D. melanogaster* and two *D. simulans* alleles with the highest allelic coverage per bp (e.g. the proportion of bases that are not missing data) together with the *D. yakuba* allele, to estimate *dN/dS* on the *D. melanogaster* branch. This prevents within-species polymorphism from inflating the estimate of the 9substitution rate. *dN/dS* estimates tend to have larger variance when there is not enough information. We thus removed estimates for genes that have fewer than 100 sites (nonsynonymous plus synonymous sites) included in the PAML analysis or whose *dS* estimates are below 0.001. To account for the variation in allelic coverage in the *D. melanogaster* genomes, Tajima's *D* (Tajima 1989) was calculated as the sum of Tajima's *D* for each allelic coverage class normalized by the square-root of the number of allelic classes. We also calculated Tajima'*s D* and estimated *dN/dS* for protein-coding genes without SCPs using the DPGP polymorphism data and multispecies alignment. All statistical analyses were done using R version 2.8 (RDevelopmentCoreTeam 2010).

Gene expression analysis

In order to determine whether 1) Genes having SCPs were likely to be expressed more or less broadly than other genes and 2) whether genes having SCPs were enriched in certain tissues, we used multiple tissue microarray data from FlyAtlas (Chintapalli et al. 2007). We downloaded the raw data from the FlyAtlas gene expression database then categorized each gene as 1) protein-coding, 2) protein-coding and harboring an SCL or PTC, and 3) non-protein-coding. For the rest of the analysis, we excluded non-protein-coding genes.

To test for broadness of expression, we used the FlyAtlas "present" call data. The FlyAtlas data consists of four duplicate arrays for each tissue type tested. Each gene is called as either present or not on each array. Therefore, a given gene can have a "present" call score from 0/4 to 4/4. We declared a gene as expressed if it was called as present in 3/4 or 4/4 of the arrays in at least one of the probes for that gene. We first asked how many tissues a gene was called as present in and calculated the means and variances for PTCs, SCLs, and all protein-coding genes. Next, we used a contingency test (*Chi-square test*) to ask whether the most broadly expressed category (i.e. present call in all tissues) or the least expressed category (i.e. present call in zero tissues) were enriched among PTCs and SCLs compared to the remaining protein-coding genes.

To determine if any single tissue was enriched among SCPs, we used the raw expression data from FlyAtlas to determine what the most highly expressed tissue was for each gene. We then asked whether there was an excess or paucity of SCPs expressed at their highest level in any given tissue compared to the total genes annotated as being expressed in at least one tissue (*Fisher's Exact Test*). We then corrected for the number of tissue types tested (10) using the *Bonferroni* adjustment (Abdi 2007).

GO analysis

We used the online GO functional annotation tool DAVID to determine if the genes were enriched for any biological, cellular, or molecular functional terms (Huang da et al. 2009). We used the FATGO annotation categories, which give extra weight to GO terms that are more specific (for example, less weight is given to broad GO terms such as "cellular component" and more weight is given to specific terms such as "vesicle"). DAVID uses a modified *Fisher's exact test* called the EASE score to test for enrichment (Dennis et al. 2003). We separately uploaded the list of genes found to contain PTCs and SCLs to DAVID's servers, bulk-downloaded the resulting enriched GO categories and then ranked the results by *p-value* to obtain a list of the top enriched categories for each gene list.

Annotation with InterProScan

We used the program InterProScan (Quevillon et al. 2005) to annotate domains in coding regions lost due to PTCs or gained due to SCLs. InterProScan cannot annotate domains in peptides shorter than 20 amino acids. Accordingly, for both SCLs and PTCs, we excluded truncated or expanded sequences from PTCs and SCLs that were shorter than 60 bp. PTCs can lead to a truncated protein or silencing of the gene by nonsense-mediated decay. Studies using *D. melanogaster Adh* transgenes suggested that the decay process was triggered if there was more than 400 bp between the stop codon and the polyadenylation site (Behm-Ansmant et al. 2007). Because the average size of a 3' untranslated region is 200 bp in *D. melanogaster* (Retelska et al. 2006), we looked for domains in CDS truncations that were 200 bp or shorter, as these are predicted to avoid nonsense-mediated decay. For SCLs, we only used alleles from genes that were not predicted to trigger non-stop decay as described above. Extracted sequences were translated and sent in bulk to the InterProScan server (www.ebi.ac.uk/Tools/InterProScan/).

RESULTS AND DISCUSSION

Hundreds of stop codon polymorphisms are present in *D. melanogaster*

We searched through the 44 *D. melanogaster* genomes generated by DPGP (www.dpgp.org) for alleles with PTCs or SCLs compared to the annotations of the *D. melanogaster* reference genome (version 5.16). To confirm that observed SNPs are not sequencing or assembly errors, we used direct sequencing and found that polymorphisms with an assembly quality score Q40 or greater were correct 60% of the time, while alleles with quality below Q40 were errors 80% of the time. This false discovery rate implies the quality of a 50 genomes Q40 SNP is actually much higher than Phred Q40 (see methods and Figure 2.1 for details). As a result, we used the DPGP genome assembly with a cutoff of quality score Q40 (bases with quality score lower than Q40 are treated as missing data). We then polarized the direction of the change for each allele with respect to the major allele in the population, which was not always the same as the reference allele. We considered polarizing by ancestry, but were unable to determine the ancestry of nearly half of the alleles (237 PTCs and 59 SCLs). This was often because of high levels of divergence between *D. melanogaster* and the *D. yakuba/D. erecta* clade.

In *D. melanogaster*, we observed 438 genes harboring 498 PTC alleles and 119 genes harboring 124 SCL alleles (Table 2.2). After quality screening and polarization, there were a total of 1,667 occurrences of all minor alleles across all genomes - this gives roughly 37.9 SCPs per genome analyzed. Although we are confident that most of the observed SCPs are not sequencing or assembly errors, we created a more conservative dataset by removing alleles that were present only once across the two *D. melanogaster* populations (Table 2.2, nonsingletons). We performed our analyses using both datasets but present only results using all alleles unless the observations are different between the two datasets.

We expect our dataset to be biased towards polymorphisms with minor fitness effects because the sequenced DPGP genomes were prepared as inbred strains (RAL populations) or strains with targeted pairs of homozygous chromosomes (MW populations). In both cases, polymorphisms that are strongly deleterious in nature were likely removed from the strains prior to sequencing.

Stop codon polymorphisms are as a group selected against

Due to the potential impact of SCPs on the function and expression of the genes harboring them, we expected *a priori* that most SCPs should be selected against more strongly than other types of variation. Four aspects of the data supported our hypothesis. First, the density (the number of polymorphic sites per mutable site across the genome) of SNPs resulting in PTCs and SCLs are 0.00021 and 0.0029 respectively, both of which are

Table 2.2 – Stop codon polymorphisms in *Drosophila melanogaster*

		All alleles			Nonsingletons only		
		Malawi	Raleigh	Total	Malawi	Raleigh	Total
PTC	Genes	146	353	438	88	147	157
	Alleles	153	395	498	91	158	170
SCL	Genes	62	93	119	21	59	65
	Alleles	65	97	124	22	63	68

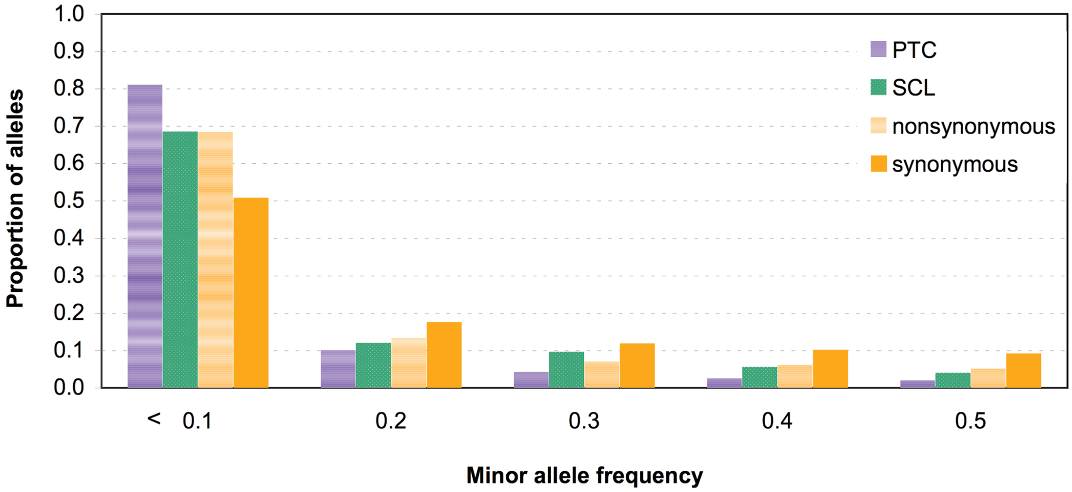


Figure 2.2: The allele frequency spectra for SCPs are skewed towards rare alleles.

PTC (violet) and SCL (green) polymorphisms are both enriched for rare alleles. SCPs are more likely to be at low frequency than synonymous SNPs (dark orange). In addition, PTCs - but not SCLs - are more skewed than highly constrained nonsynonymous SNPs (light orange).

small when compared to a density of 0.0089 for nonsynonymous sites and 0.090 for synonymous sites. Secondly, we found that the allele frequency distributions of SCPs of both types are skewed towards rare variants when compared with synonymous polymorphisms (Figure 2.2, *Chi-square test, p* < 10^{-16} (PTC), p = 0.03 (SCL)). The allele frequency distribution of PTCs is also more skewed than that observed for highly constrained nonsynonymous polymorphisms (Figure 2.2, *Chi-square test, p* < 10^{-11}) while the distribution for SCLs was neither more nor less skewed than nonsynonymous polymorphisms (Figure 2.2). We noted that several SCPs only affect some of the many isoforms of the genes they reside in. The fitness consequences of such polymorphisms are likely to be less extreme, and may be less likely to be selected against. To test this hypothesis, we removed any SCPs that affected less than 50% of a gene's isoforms (18.9% of PTCs and 23.4% of SCLs), repeated the comparison and observed an even stronger enrichment of rare alleles for PTCs but no change in the result for SCLs.

Thirdly, we found that both PTCs and SCLs were enriched for alleles that cause less extreme changes of the coding region length (Figure 2.3), suggesting that extreme alleles are more strongly selected against and less likely to be sampled. Assuming mutations occur randomly, the positions of PTCs within coding regions should be uniformly distributed. However, we estimated the empirical distribution of codons that are one mutational step away from a stop codon (one-step codons), and found that such codons are *not* uniformly distributed within coding sequences, having a slightly higher proportion of one-step codons in the 3' regions of genes (Figure 2.3A blue dots). We found that the distribution of PTCs was significantly different from both the uniform distribution and the one-step codon distribution (*Chi-square test,* both *p* < 10^{-4}), with an excess of PTCs at the start and end of coding regions (Figure 2.3A). Our observation that an excess of PTCs are found near the start of coding sequences is intriguing. As expected, most of the alleles with highly truncated coding sequences are segregating at low frequency in the population with a few interesting exceptions (see below). The number of base pairs added after an SCL is expected to follow a *Poisson* process with parameter λ as the mean length, provided that stop codons are randomly distributed in the 3' untranslated region. However, the absence of a stop codon can lead to gene silencing by non-stop decay (Vasudevan et al. 2002), which is triggered when there is no stop codon prior to the polyadenylation site. In order to measure the effect of length expansion on allele frequency, we considered only those alleles that are not predicted to undergo non-stop decay (see methods). Among these, the mean number of codons added was 5.5, with the longest and shortest expansions being 1 and 49 codons respectively. We found that the distribution of length change was significantly different from the *Poisson* process expectation, with an excess of both small and large length changes (*Kolmogorov-Smirnov test, p* $< 10^{-7}$, Figure 2.3B). The excess of small changes may be due to selection against extreme changes in gene length whereas the excess of longer changes is intriguing and could be due to nonrandom distribution of stop codons in some 3' untranslated regions. Given these observations, we expected to see a negative correlation between the size of change caused by the SCP and its frequency in the population. However, we did not see this correlation for either PTCs or SCLs (*Spearman's rank ρ, p* all > 0.05). Restricting to alleles influencing more than half of a gene's isoforms yielded a similar insignificant result. It is possible that the realized allele frequency is more affected by the function of the gene in question than by the extremity of the allele.

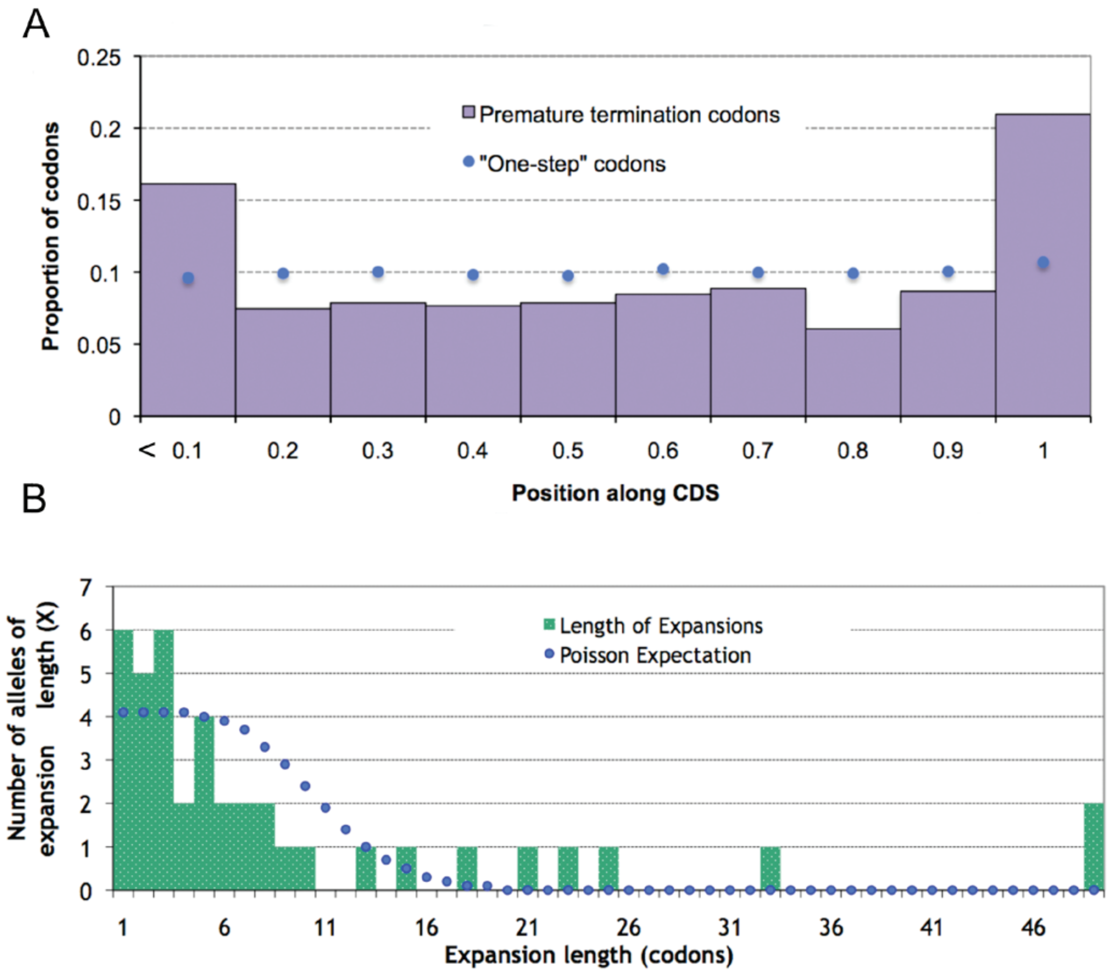


Figure 2.3 - Extremely long truncations and expansions of genes are more rare than expected.
The change in gene length was predicted for PTCs (A) and SCLs (B). (A) PTCs appearing earlier in the coding regions are expected to have a more extreme effect on gene function than those appearing near the ends of genes. There was a significant excess of short truncations compared to the distribution of one-step codons (blue dots). (B) For alleles with SCLs whose gene model has an annotated 3′ untranslated region, the number of codons added is shown (green bars) along with the expectation if the length expansions due to SCLs followed a *Poisson* process (blue dots). The distribution was significantly different than the *Poisson* process expectation, with an excess of both long and short alleles.

Finally, we observed that there is a deficiency of genes harboring PTCs on the X chromosome compared with autosomes, which likely results from stronger purifying selection against deleterious recessive alleles on the X chromosome in males (1.56% for X chromosomes and 3.29% of autosomes, *Fisher's exact test, p* $< 10^{-3}$, Figure 2.4). This pattern was only marginally significant for SCLs (0.37% of X-linked and 0.91% of autosomal genes, *Fisher's exact test, p* = 0.02). We repeated the analysis with SCPs that affected more than 50% of a gene's isoforms and found a significant paucity of SCLs on the X (0.18% of X-linked and 0.71% of autosomal genes, *Fisher's exact test*, *p* = 0.007), and an even stronger X deficiency for PTCs (1.5% on the X and 3.4% on the autosomes, *Fisher's exact test*, *p* $< 10^{-4}$). This pattern was not significant when we only considered nonsingletons because such alleles are likely to be less deleterious due to their high population frequency.

PTCs that trigger nonsense-mediated decay and SCLs that trigger non-stop decay are both expected to have greatly reduced expression and to be functional null alleles. Alleles that escape these surveillance processes are likely to have impaired gene function. *A priori* then, these polymorphisms should have equal likelihood to be deleterious. However, this conclusion must be taken with caution, as it is not known how universal these processes are. We observed that the change in length of protein sequence is more dramatic in alleles harboring PTCs than SCLs (Figure 2.3), and that most SCLs are not predicted to trigger non-stop decay. This is due to the fact that length expansion resulting from SCLs is constrained by the length of the 3' untranslated region, which is in general shorter than the coding regions where PTCs could happen. Therefore, it seems likely that

PTCs would be more deleterious than SCLs. Our observations are consistent with this scenario. We found that PTCs are present at a lower density (0.00021 for PTCs and 0.0029 for SCLs), their frequency spectrum is more skewed towards rare variants (*Chi-square* test, $p = 0.01$) and a smaller proportion of PTCs are observed on the X chromosome compared with SCLs, though the difference was not statistically significant (8.43% for PTCs and 9.68% for SCLs, *Fisher's exact test*, $p > 0.05$).

Our observations supported our hypothesis that SCPs of both types as a group should be selected against. However, we are unable to distinguish deleterious SCPs, which are maintained at mutation-selection balance from weakly deleterious SCPs, whose frequencies are also affected by genetic drift. To determine whether selection might be acting on the observed polymorphisms, we predicted the number of polymorphisms we would expect to sample under a variety of selective scenarios (Table 2.3). By assuming additive dominance and that in *D. melanogaster*, Ne ~ 10^{-6} (Kreitman 1983; Charlesworth 2009) and m ~ 10^{-9} (Keightley et al. 2009), f_{eq} for nearly neutral alleles should be m / Ne or ~ 10^{-3}. Meanwhile, f_{eq} for recessive deleterious alleles under mutation-selection balance should be at least an order of magnitude lower (for example, when s = 10^{-5}, mu/h*s ~ 10^{-4}). If we assume that every possible PTC in the genome (2,387,149 possible sites) is nearly neutral, we would expect to sample 102,808 alleles across 44 genomes (Table 2.3). We only observed 498, implying selection is acting robustly to remove the vast majority of possible PTC alleles from the population. Hence, a significant number of the sampled alleles are likely under purifying selection (s >10^{-5}). On the other hand, alleles under strong selection are unlikely to be sampled more than

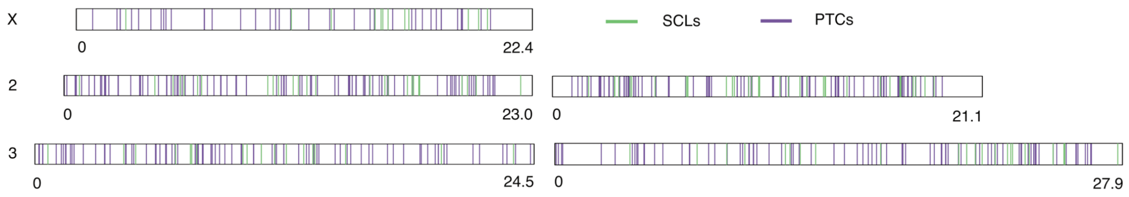


Figure 2.4 - Fewer SCPs are found on the X chromosome.
The genomic position of each PTC (violet) and SCL (green) are shown to scale on the chromosomes on which they are found (the length in Mb of each chromosome is shown). The X chromosome is underrepresented for PTCs and SCLs compared to the expectation given the number of genes on each chromosome.

Table 2.3 – Effect of purifying selection on sampling probability

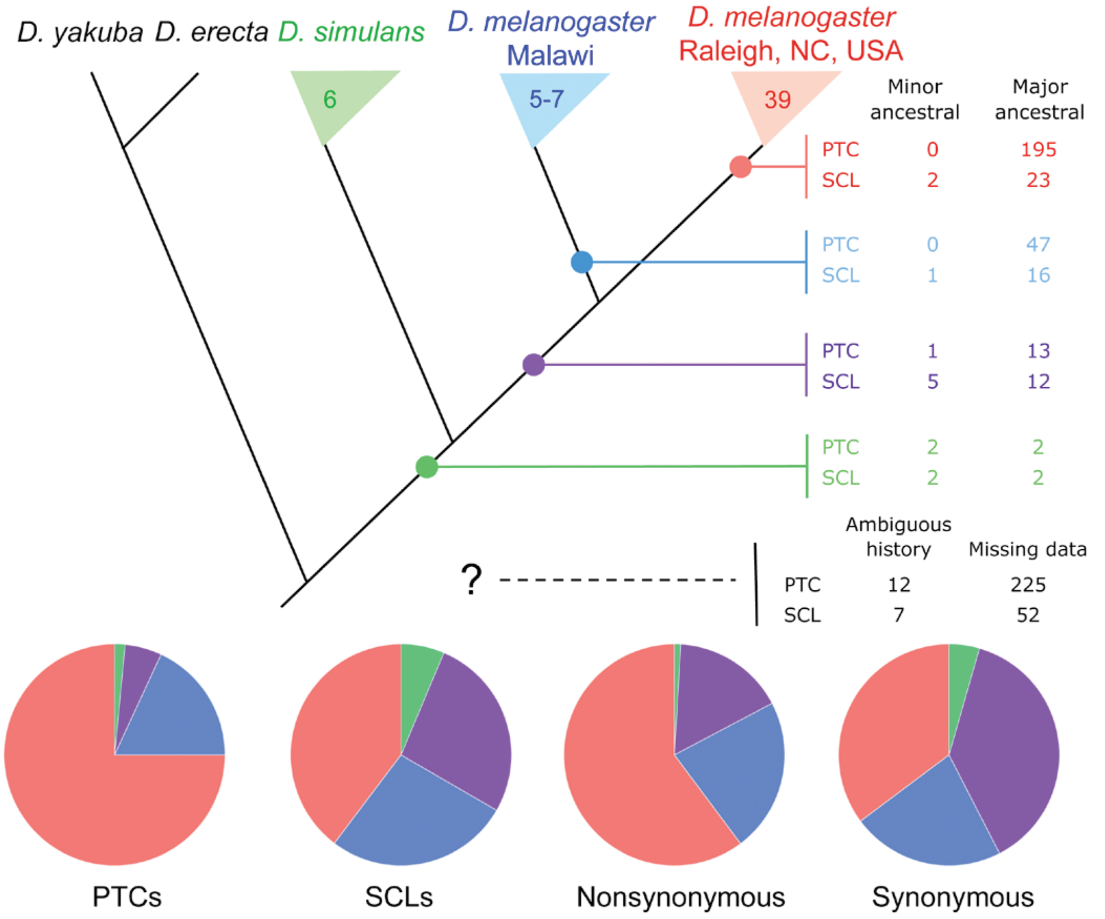
	f_{eq}	probability of sampling zero times	probability of sampling once	probability of sampling more than once	number sampled zero times	number sampled once	number sampled more than once
s~0 (nearly neutral)	$\sim 10^{-3}$	0.95693	4.21E-02	9.20E-04	2284341	100612	2196
s=0.00001	$\sim 10^{-4}$	0.99561	4.38E-03	9.43E-06	2376668	10458	22.5
s=0.0001	$\sim 10^{-5}$	0.99956	4.4E-04	9.46E-08	2386099	1050	0.2
s=0.001	$\sim 10^{-6}$	0.99996	4.4E-05	9.46E-10	2387044	105	~0
s=0.01	$\sim 10^{-7}$	0.999996	4.4E-06	9.46E-12	2387138	11	~0
s=0.1	$\sim 10^{-8}$	0.9999996	4.4E-07	9.62E-14	2387148	1	~0

* Probabilities are assuming binomial sampling across 44 genomes
** Number of observations is assuming 2,387,149 sites were sampled at the calculated probabilities

once. Therefore, we predict that the alleles we sampled more than once are neutral or nearly neutral.

Most SCPs are newly derived on the *D. melanogaster* lineage

Given that SCPs are selected against as a group, we predicted that most SCPs should be newly derived. To infer whether stop codon polymorphisms are recently derived on the *D. melanogaster* lineage and to identify alleles with interesting evolutionary histories, we determined whether any of the six *D. simulans* genomes or the *D. yakuba* and *D. erecta* reference genomes shared each SCP with the *D. melanogaster* populations (Figure 2.5). 315 SCPs within *D. melanogaster* were fixed in *D. simulans* for the allele in the *D. yakuba/D. erecta* outgroup, suggesting a recent origin of these SCPs on the *D. melanogaster* lineage. Conversely, we found 13 alleles that are polymorphic in both *D. simulans* and *D. melanogaster*, although only eight of these (4 PTCs and 4 SCLs) have data available in the outgroup. These alleles likely have been segregating since before the species diverged approximately 5.4MYA (Tamura et al. 2004) and are of substantial interest. We also observed that *D. melanogaster* population frequencies of SCL and PTC alleles that are shared between the two species are significantly higher than those that are specific to *D. melanogaster,* suggesting these shared polymorphisms may have been present for long periods of time (*Chi-square test, p* $< 10^{-5}$ for PTCs and 0.03 for SCLs). However, we cannot exclude the alternative possibility that our observations were the result of independent mutations arising in the two lineages.



	Minor ancestral	Major ancestral
PTC	0	195
SCL	2	23
PTC	0	47
SCL	1	16
PTC	1	13
SCL	5	12
PTC	2	2
SCL	2	2

	Ambiguous history	Missing data
PTC	12	225
SCL	7	52

Figure 2.5 - PTCs are more derived than nonsynonymous polymorphisms We classified each PTC and SCL allele as recently derived in the Raleigh, NC population (red) or the Malawi population (blue); shared by the two *D. melanogaster* populations (violet); or shared with *D. simulans* (green). The number of alleles sampled is shown for each branch (branch lengths are not to scale). The outgroup alleles (*D. yakuba* and *D. erecta*) allowed us to determine whether the current *D. melanogaster* major or minor allele was likely ancestral (side panel minor/major). Almost half of the alleles could not be categorized due to a lack of sequencing/alignment data from one or more species ("Missing data"), or because it was unclear whether the minor or major allele was ancestral ("Ambiguous history"). Pie charts show the proportion of alleles in each described age category for PTCs, SCLs, nonsynonymous SNPs, and synonymous SNPs.

We next asked whether the age distribution of PTCs or SCLs differed from nonsynonymous or synonymous polymorphisms using *Chi-square* tests (Figure 2.5). We found that PTCs had an excess of Raleigh- and Malawi-specific alleles compared to either nonsynonymous polymorphisms ($p < 10^{-6}$) or synonymous polymorphisms ($p < 10^{-16}$). The age distribution of SCLs was not different from the observations for synonymous polymorphisms ($p > 0.05$), but was different from either PTCs ($p < 10^{-8}$) or nonsynonymous polymorphisms ($p < 10^{-5}$), having an excess of alleles shared between the two *D. melanogaster* populations and with *D. simulans*. These results show that PTCs are even more likely to be new mutations than nonsynonymous polymorphisms, while SCLs show a very different pattern, with a similar age distribution as synonymous polymorphisms. This corroborates the pattern in our data that suggested PTCs are more strongly selected against than SCLs.

Among the *D. melanogaster*-specific alleles, 31 alleles are polymorphic in both *D. melanogaster* populations, 64 are segregating in only the Malawi population and 220 are segregating only in the Raleigh population (Figure 2.5, inset table). Previous research suggests that the Malawi population has higher overall polymorphism than non-African populations (Begun and Aquadro 1993; Haddrill et al. 2005; Hutter et al. 2007). However, after correcting for the effect of sample size (see methods), we found no significant excess of alleles in the Malawi population (*Fisher's exact test, p* > 0.05). This may be explained by the recent demographic history of non-African *D. melanogaster* populations (Stephan and Li 2007), which could result in less effective selection against deleterious SCPs. Finally, we found nine alleles where the major allele in the population is derived with respect to the inferred ancestral state. These alleles have recently increased in frequency and are good candidates to be targets of recent positive selection (see below).

Mutation contributes to the appearance of new SCPs

Our observations supported the hypothesis that most SCPs are likely to be either deleterious or weakly deleterious. Population frequencies of SCPs should thus be determined by the intensity of selection removing SCP alleles and the rate of new mutations increasing their frequency. Accordingly, we expected that genes with larger mutational targets and/or weaker selective constraint would be more likely to harbor SCPs. The mutational targets of PTCs are any codons that can mutate directly to a stop codon (one-step codons). Hence, genes containing a larger number of one-step codons should be more likely to harbor PTCs. We would expect the pattern to be even stronger when considering the proportion of codons that are one mutational step away from two stop codons (two-fold one-step codons, TAC, TAT, TCA, TTA, TGG). Indeed, although we did not find an excess of one-step codons in genes carrying PTCs, these genes had a significantly larger number of two-fold one-step codons (32.1 *versus* 28.2 *Mann-Whitney U test, p* = 0.001). Additionally, the three stop codons of Drosophila are AT rich and we observed higher AT content among PTC genes than other genes (49.5% *versus* 46.4%, *Mann-Whitney U test*, $p < 10^{-16}$). However, it is worth noting that most of the unpreferred codons in *D. melanogaster* are also AT-rich (Akashi 1994). Highly expressed, slowly evolving genes have stronger codon bias and higher GC content (Duret and Mouchiroud 1999; Marais et al. 2004; Subramanian and Kumar 2004; Lemos et al. 2005; Larracuente et al. 2008) and we also found that genes carrying PTCs have weaker codon bias than other genes (Fop 0.44 *versus* 0.51, *Mann-Whitney U test*, $p < 10^{-14}$). Accordingly, it is difficult to tease apart whether mutation or indirect selective forces are the underlying cause of our observation that PTCs have larger numbers of one-step codons.

The mutational target of SCLs is the original stop codon. We would predict that TGA and TAG stop codons should more likely to be lost than TAA codons because two possible mutations from the TAA retain a stop codon whereas only one mutation from TGA or TAG is silent. Supporting this idea, we observed that TAA stop codons are more likely to harbor silent polymorphisms (minor allele has an alternative stop codon at the same position) than TAG or TGA stop codons (*Fisher's exact test,* $p = 0.02$). However, we did not see the predicted paucity of TAA non-silent changes among SCLs compared to TGA and TAG changes (*Fisher's exact test,* $p > 0.05$). Therefore, we cannot conclude that mutational bias has a strong role in the origin of new SCLs. Yet, this was a weak test as TAA codons have only a marginally lower chance of being lost than TAG or TGA codons (2 of 9 mutations are silent rather than 1 of 9).

Genes harboring SCPs exhibit lower evolutionary constraint than other genes

Given our *a priori* expectations of fitness impacts of SCPs, the intensity of selection against a SCP depends both on how severely the SCP allele affects gene function and how essential the affected gene is. We can test the hypothesis that genes harboring SCPs are less evolutionarily constrained by comparing the *dN/dS* estimates between genes with and without SCPs. High *dN/dS* estimates can be interpreted as either elevated rates of adaptive evolution (positive selection) or as weaker selective constraint (reduced purifying selection). Here, we used the *dN/dS* ratio as a proxy for selective constraint, as most of the genes have a ratio well below one and so are not likely to be under positive selection. Our results are consistent whether or not we include genes showing evidence of adaptive protein evolution (*dN/dS* > 1). We found that genes harboring SCPs have significantly higher *dN/dS* ratios than other genes (*Mann–Whitney U*, $p < 10^{-6}$ for both PTCs and SCLs, Figure 2.6A). We also used Tajima's *D* to address this question. Tajima's *D* summarizes the frequency spectrum of the within population polymorphism, and strong purifying or directional selection usually leads to highly negative Tajima's *D* estimates**.** We found no significant difference in Tajima's *D* between genes with and without SCPs (Figure 2.6B).

We might predict certain groups of SCPs are particularly likely to be under weak constraint or even affected by positive selection. For example, alleles which have increased in frequency recently could be under positive selection, or could have drifted to fixation as nearly-neutral alleles. We found that the nine genes harboring alleles in which the major allele is derived relative to the ancestral state had less negative (closer to zero) Tajima's *D* statistics and larger (but still on average < 1) *dN/dS* estimates than other genes or genes harboring other SCPs (*Mann-Whitney U tests,* $p < 0.05$ all tests). The genes carrying the 13 SCPs segregating in both *D. melanogaster* and *D. simulans* also showed a less negative Tajima's *D* than other genes and than other SCP genes (*Mann-Whitney U test,* $p < 0.05$ for both tests) and a larger, though insignificant, *dN/dS* ratio. Together, these observations are consistent with the hypothesis that genes carrying these subsets of SCPs are under weaker selective constraint than other genes. It is worth noting our overall observation that SCP genes have higher *dN/dS* than other genes was not

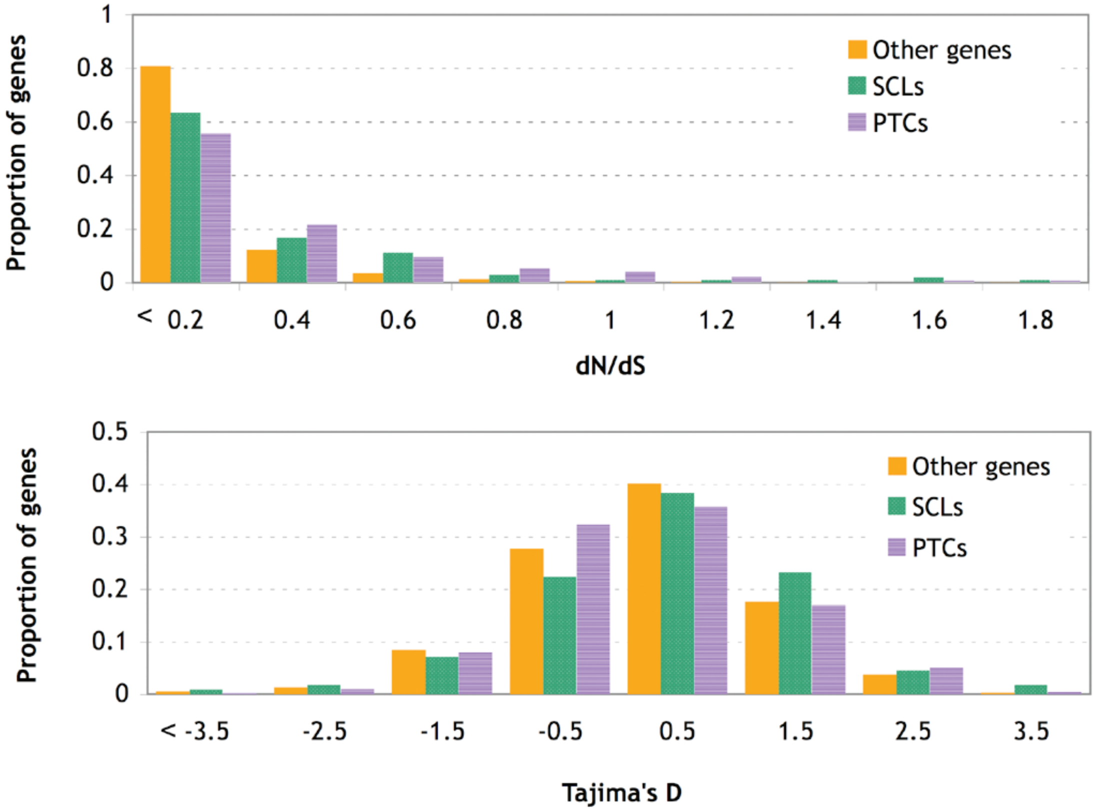


Figure 2.6 - The *dN/dS* ratios for genes harboring SCPs are higher than typical genes.
dN/dS (A) and *Tajima's D* (B) were calculated across the coding sequence for genes harboring stop codon losses (green), genes harboring premature stop codons (violet), and all other genes (orange). PTCs and SCLs both had a significantly elevated *dN/dS* ratio compared to all genes. There was no statistical difference in *Tajima's D* between the different gene types.

driven by these special groups, as removal of these genes still yielded significant differences (*Mann–Whitney U*, $p < 10^{-4}$ for both PTCs and SCLs).

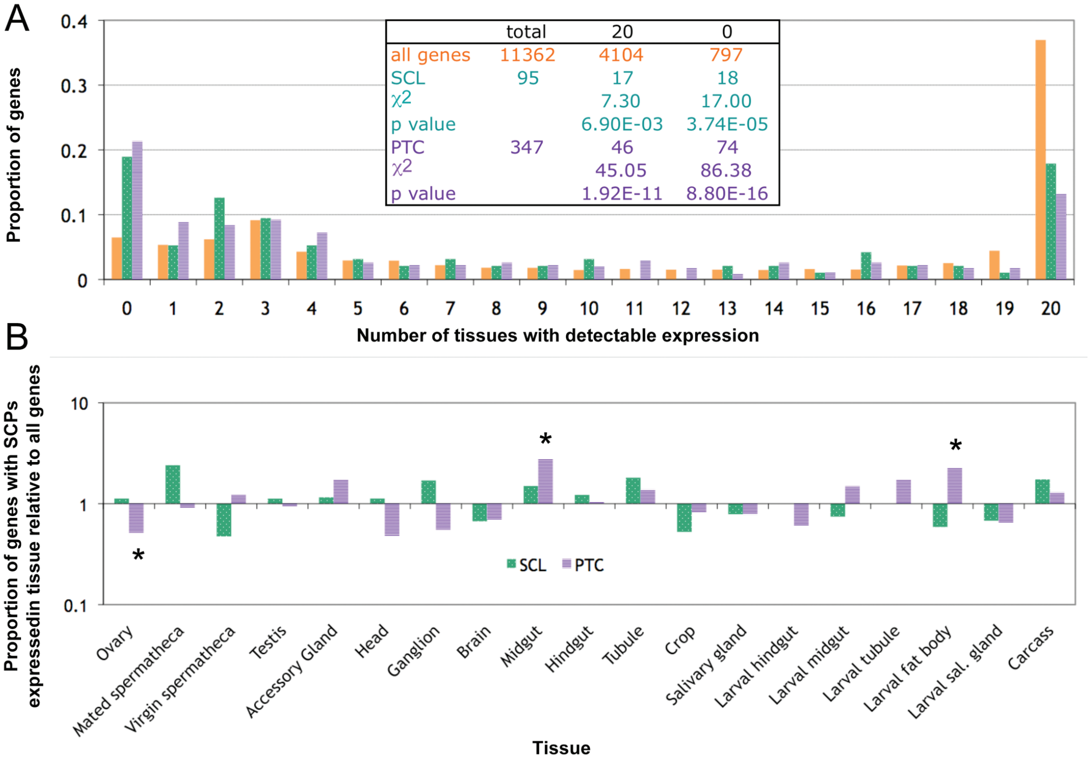
Genes harboring SCPs are more narrowly expressed than other genes

We found above that genes harboring SCPs are likely to be under weak functional constraint. The expression pattern of a gene is one of the most important indicators of gene function. It has been shown that genes expressed broadly and at a high level are more likely to be under strong selective constraint whereas narrowly and weakly expressed genes are more likely to evolve with less selective constraint (Subramanian and Kumar 2004; Larracuente et al. 2008) or frequent directional selection (Begun and Lindfors 2005; Schully and Hellberg 2006). Given our observation that SCLs have elevated *dN/dS* ratios, we expected to see an excess of narrowly expressed genes. We used microarray expression data from FlyAtlas (Chintapalli et al. 2007), to determine whether genes harboring SCPs had different expression patterns than other genes. We asked whether genes harboring SCPs were more likely than other genes to have no detectable expression, and whether they were less likely to be expressed broadly. Consistent with our predictions, PTCs and SCLs were both significantly more likely than other protein-coding genes to be expressed in none of the tissues tested, and significantly less likely to be expressed in all twenty tissues (Figure 2.7A, inset *Chi-square test, p* < 0.05).

We also asked whether there was an enrichment of genes expressed in particular tissues among either PTCs or SCLs. For each gene, we asked what the most highly expressed tissue was and determined whether each tissue was enriched or depleted among either type of SCP compared with all genes (Figure 2.7B). PTCs were more likely than expected to occur in genes expressed at their highest level in the larval fat body (*Fisher's Exact test, p* = 0.011) and the adult midgut (*Fisher's Exact test, p* = 0.004) and they are less likely to occur than expected in genes expressed at their highest level in the ovary (*Fisher's Exact test, p* = 0.011). SCLs had no significant enrichment or depletion in any tissue. Genes expressed in the ovary include maternally deposited developmental genes, many of which are essential. This may explain why few ovary-specific genes carry PTCs. Conversely, the larval fat body is a common place for immunity genes to be expressed. Several SCPs in immunity genes have previously been observed (Jiggins and Kim 2005; Lazzaro 2005). Further, immunity genes are known to show unusually rapid copy-number evolution (Sackton et al. 2007), including changes in copy number due to duplication, deletion, and pseudogenization. As acquisition of stop codons can lead to pseudogenization, we may be witnessing the early stages of copy number evolution.

Gene ontology analysis shows SCPs are enriched for chemoreceptors

We can also infer levels of functional constraint using the functional annotation of a gene. Loss-of-function alleles in genes with dispensable functions are less likely to be strongly selected against than similar alleles in essential genes. We used the Gene Ontology annotation tool DAVID (see methods) to determine whether genes with SCPs were enriched for specific functions. We found that genes with PTCs and SCLs are equally likely to be associated with at least one GO category as other genes (all genes 63.5%, PTCs 66.7%, SCLs 67.2%, *Chi-square tests*, *p* > 0.05), indicating that genes with SCPs are not strongly biased towards unannotated genes. We found that PTCs were



	total	20	0
all genes	11362	4104	797
SCL	95	17	18
$\chi 2$		7.30	17.00
p value		6.90E-03	3.74E-05
PTC	347	46	74
$\chi 2$		45.05	86.38
p value		1.92E-11	8.80E-16

Figure 2.7 - Stop codon polymorphisms are expressed in fewer tissues than other protein coding genes.

A) All protein-coding genes (orange) were far more likely to be expressed in all twenty tissues tested than genes harboring either SCLs (green) or PTCs (violet), and far less likely than either group of SCPs to be expressed in none of the tissues tested (inset *Chi-square test*, *p* < 0.01 for all cases). B) Gene harboring PTCs were more likely than other genes to be expressed at their highest level in the larval fat body and midgut but less likely in the ovary. None of the assayed tissues are significantly enriched for genes harboring SCLs compared to other genes.

Table 2.4 – Enriched GO terms among genes with stop codon polymorphisms

Allele type	GO term	Description	# of genes	P-value
PTC	GO:0006508	Proteolysis	37	3.7E-06
PTC	GO:0007606	Sensory perception of chemical stimulus	17	6.1E-06
PTC	GO:0008233	Peptidase activity	37	4.1E-05
PTC	GO:0070011	L-amino acid peptidase activity	35	6.9E-05
PTC	GO:0044421	Extracellular region component	14	8.0E-05
PTC	GO:0050909	Sensory perception of taste	9	8.9E-05
PTC	GO:0005576	Extracellular region	28	9.9E-05
PTC	GO:0008527	Taste receptor activity	9	1.2E-04
PTC	GO:0007600	Sensory perception	18	1.4E-04
PTC	GO:0007186	G-protein coupled receptor signaling	19	2.1E-04
SCL	GO:0007186	G-protein coupled receptor signaling	8	2.6E-03
SCL	GO:0005615	Extracellular space	4	1.5E-02
SCL	GO:0050890	Cognition	7	1.8E-02
SCL	GO:0007600	Sensory perception	6	2.5E-02
SCL	GO:0033043	Regulation of organelle organization	4	2.8E-02
SCL	GO:0004965	GABA-B receptor activity	2	3.2E-02
SCL	GO:0007166	Cell surface receptor signal transduction	9	3.3E-02
SCL	GO:0016021	Integral to membrane	15	3.4E-02
SCL	GO:0051493	Regulation of cytoskeleton organization	3	3.6E-02
SCL	GO:0031224	Intrinsic to membrane	15	3.9E-02

enriched for GO terms associated with proteolytic activity and that both PTCs and SCLs were enriched for GO terms associated with the sensation of chemical stimuli and the plasma membrane (Table 2.4). However, non-singleton PTCs and SCLs did not show enrichment for these chemoreceptory GO terms. For both PTCs and SCLs, the enrichment of chemical sensation and plasma membrane GO terms appears to be driven by the fact that many gustatory receptors (GRs), odorant receptors (ORs), and other chemoreceptors (IRs) harbor SCPs. Most chemoreceptors are dispensable (that is, null mutations do not cause lethality or sterility) and both GRs and ORs are known to evolve rapidly between species (Matsuo et al. 2007; McBride 2007; McBride et al. 2007; Dworkin and Jones 2009). Therefore, it is unsurprising we find SCP alleles present in these genes in *D. melanogaster*.

Genes with unusual evolutionary histories

The pattern of variation observed among SCPs is consistent with our expectations if SCPs are as a group selected against. However, we were also interested in investigating genes that may not be following this overall pattern. First, we noted that 56 genes harbored more than two SCPs in *D. melanogaster.* These genes may be evolving under weak selective constraint, but could also be selected for multiple variants (diversifying or balancing selection). Named genes in this group included *Acp26Aa*, which is one of the most rapidly evolving genes in the *D. melanogaster* genome (Schully and Hellberg 2006; Wong et al. 2006), *att-ORFB*, two gustatory receptors (*Gr59f* and *Gr36a*), and one predicted chemosensory protein (*CheA86a*). *Acps* were observed to undergo rapid loss-and-gain in the *melanogaster* species subgroup (Begun et al. 2006) and length variations of *Acp26Aa* in this species subgroup has been described (Aguadé 1998). Several loss-of-function and PTC alleles of other *Acps* were also documented in a survey of natural variation (Begun and Lindfors 2005). Consistent with this, we observed that *Acp26Aa* harbors one SCL allele that expands the open reading frame by one codon and one PTC allele that shortens it by seven codons, along with the major allele that matches the *D. melanogaster* reference annotation. It is possible that rapid diversifying selection of *Acp26Aa* includes the acquisition of SCPs among other types of polymorphism. The observation of *att-ORFB,* one of the several transcripts from a bicistronic mRNA expressed in adult testes (Madigan et al. 1996), is unsurprising given that many genes related to male reproduction are rapidly evolving in Drosophila (Zhang et al. 2004, Richards et al. 2005, Schully and Hellberg 2006). Similarly, chemoreceptors are also known to rapidly evolve between species (Matsuo et al. 2007; McBride 2007; McBride et al. 2007; Dworkin and Jones 2009). Genes that carry many SCPs warrant further study due to the possibility that diversifying selection may drive these genes to carry many alleles. On the other hand, not all of these genes have positive evidence for protein-coding ability, raising the possibility that their open reading frames are less constrained because they are mRNA-like non-coding RNA genes that are misannotated as protein-coding genes (Rymarquis et al. 2008). Such genes would be expected to tolerate SCPs because they are not translated.

Another interesting group are nine genes (one PTC and eight SCLs) whose major alleles are derived relative to the ancestral state, probably resulting from recent, rapid increases in allele frequencies. However, only small protein length differences were generated by these SCP alleles. Most of these are unnamed genes and none have known functions. Two interesting cases are *CG15531*, a predicted *stearoyl-CoA 9-desaturase,* and *att-ORFB,* a testis-expressed gene that also harbors multiple SCP alleles (see above).

We noted that SCPs are enriched with alleles causing small as well as large protein length changes (see above). Among the PTC alleles causing extreme changes (truncation of more than half of the coding sequence), ten alleles have population frequency above 25% and three named genes (*gfA, Flo-2* and *dpr2*) are among this list. However, PTCs in these named genes influenced only a few isoforms out of the many isoforms of the genes, suggesting their influence on *D. melanogaster* fitness may be less severe than predicted by change of coding region alone. All the SCL alleles with extreme number of codons added or predicted under non-stop decay have low population frequency.

Finally, we observed 13 *D. melanogaster* SCP alleles that are also segregating in *D. simulans* and four of them (PTCs) are in named genes (*Sucb, dpr2*, *Vha100-1* and *Fak56D*). While large truncations of protein sequences (from 16% to 97% of coding sequences) were caused by PTCs in these named genes, only one isoform was affected. These results are generally consistent with our finding that purifying selection is removing mutations in essential genes or essential parts of genes and mutations causing extreme changes in protein length. These alleles usually affect only unnamed genes or a few isoforms of named genes. Thus, the unusual alleles we found may be explained by the overall pattern we have observed –SCPs are as a group selected against and affect weakly constrained genes. However, we also found an enrichment of genes previously known to be rapidly evolving within Drosophila or to harbor nonsense alleles (e.g.

chemoreceptors and male-specific genes), indicating that SCPs might be important to the evolution of these genes.

Stop codon polymorphisms lead to the loss and gain of protein regions

Previous studies have shown that domains of proteins can be lost and gained through evolutionary time, and that these mutations are biased towards the 5' and 3' ends of proteins (Bjorklund et al. 2005; Weiner et al. 2006). Although we observed that there is a bias towards SCPs causing small changes in protein length, we wanted to know whether protein sequence features might be added or lost in the SCP alleles. We used the annotation tool InterProScan (Quevillon et al. 2005) to determine if truncated parts of PTC alleles that are not targeted by nonsense-mediated decay or expanded parts of SCP alleles that are not targeted by non-stop decay contained any known sequence features or domains.

We found that 23 of 71 alleles causing truncations that are expected to escape nonsense-mediated decay had lost characterized sequence features including signal peptides, protein-binding domains, DNA-binding domains, and catalytic domains. However, one caveat is that the exact trigger for nonsense-mediated decay is not well understood on a genome-wide scale (Behm-Ansmant et al. 2007). Exactly which PTC alleles will lead to domain loss and which will lead to silencing will vary depending on how much the mechanism of nonsense-mediated decay differs between genes, which has not yet been established in Drosophila.

Among SCLs expected to avoid non-stop decay (the same set as used for gene expansion analysis above), we found one gene with an SCL allele resulting in the acquisition of an apparently novel sequence features. The SCL allele of *muscleblind,* which codes for a zinc-finger protein with roles in apoptosis, muscle development (Begemann et al. 1997), and sexual behavior (Juni and Yamamoto 2009), acquired a 20 amino acid signal peptide. This allele has a population frequency of 0.19, which is among the highest frequency SCLs. The idea that a protein might expand into its 3' untranslated region and acquire a novel peptide is intriguing, and certainly warrants further functional study.

CONCLUSIONS

Natural mutations causing null alleles of genes have long been of interest to geneticists. Many of the first disease causing alleles characterized in humans carried premature termination codons (Chang and Kan 1979; Rosenfeld et al. 1992), and null alleles of allozymes in Drosophila were some of the earliest natural variants to be characterized (Voelker et al. 1980; Langley et al. 1981; Burkhart et al. 1984). Until recently, it has been unclear how common null alleles caused by variation in the position of stop codons are, as study has been restricted primarily to alleles defined by lack of function. Further, we do not understand how stop codon variants first arise within populations, leading to changes in gene models over evolutionary time.

Here, we used newly available *D. melanogaster* genomes from North American and African populations, and performed a genome-wide survey for alleles causing changes in the position of the stop codon. We found several hundred such polymorphisms segregating in the *D. melanogaster* genome, and these alleles are a mixture of deleterious and slightly deleterious mutations. SCPs had more extreme allele frequency spectra than other types of polymorphisms, were enriched for small changes in protein length, and were found less often on the X chromosome, indicating purifying selection is acting to reduce the frequency of such polymorphisms. An appreciable number of SCPs in more than one genome were also observed, suggesting some of the observed SCPs are subject to both the effects of selection and genetic drift. We also found evidence that both mutational pressure and selective constraint are important in determining the likelihood a gene harbors SCPs. We described several exceptional SCPs, which include alleles that are shared between *D. melanogaster* and *D. simulans*, alleles with high population frequency despite causing dramatically altered protein lengths, and alleles that arose and quickly became the major allele in *D. melanogaster*. Additionally, there are 56 genes that carry more than two alleles with different stop codon positions in *D. melanogaster*. These include rapidly evolving genes such chemoreceptors and male-expressed genes. Finally, one SCL gene, *muscleblind*, appears to have gained 3' sequence with similarity to a signal peptide. This implies the possibility for genes to gain domains as well as lose them.

Parallel resequencing projects have uncovered stop codon polymorphisms in humans (Yamaguchi-Kabata et al. 2008; Yngvadottir et al. 2009; Durbin 2010), providing an opportunity to contrast findings across species. The human and Drosophila data differ in some important ways - human genomes were sequenced in a heterozygous state whereas the DPGP project sequenced homozygous flies. It is therefore expected that the human data would contain more alleles - especially deleterious alleles - than does the fly data. Indeed, the reported density of PTCs and number of observed PTCs per genome in humans is much higher than in Drosophila (PTC density: 0.00021 (fly) *versus* 0.00085 (human) (Yamaguchi-Kabata et al. 2008); PTC per individual: 37.9 (fly) *versus* 80-100 (human) (Durbin 2010). Further, it was reported that PTCs are distributed evenly across the coding regions in humans (Yngvadottir et al. 2009), which is in contrast to our observation that PTC alleles are enriched for those causing small changes. These difference may also be explained by the much smaller effective population size of human compared to Drosophila*,* which results in less effective selection. Further, 59% of human nonsense alleles were found to be present in the homozygous state in some individuals (Yngvadottir et al. 2009). If this frequency were similar in Drosophila, a sampling of alleles in the heterozygous state should uncover many more SCPs than we were able to find in this study. Yet, we must be cautious when comparing these datasets because there may be different (and unknown) biases resulting from the fundamental differences in sequencing technology and SNPs-calling methods between the human and *Drosophila* data. Finally, while both the Drosophila and human data showed that selection is acting to reduce population frequency of nonsense SNPs as a whole, some SCPs violating this pattern were identified. Yngvadottir *et al.* (2009) reported *MAGEE2,* which appeared to have increased in frequency in Asian human populations despite causing a 77% truncation of the open reading frame. Likewise, we identified several SCPs that have increased in frequency ("Ancestor minor" alleles), and several genes that carry many SCPs. Intriguingly, Gene-Ontology enrichment analysis in both species found chemosensory receptors are enriched with nonsense SNPs, consistent with the idea that dispensable, rapidly evolving genes are more likely to harbor strong-effect mutations.

In sum, our study provides the first comprehensive description of the variation in stop codon position in Drosophila, and we show that polymorphisms changing the position of the stop codon were as a group selected against.   However, a number of genes that broke this pattern in various ways were identified and warrant further analysis. Because the study system was Drosophila, this analysis also provides a list of *D. melanogaster* and *D. simulans* stocks harboring a variety of natural nonsense polymorphisms, which can be readily applied to studies of the functional consequences of these natural variants.

CHAPTER THREE: TWO EXTREMELY RAPIDLY EVOLVING GENES CONTRIBUTE TO MALE FITNESS IN DROSOPHILA

Authors: Josephine A Reinhardt and Corbin D Jones

Originally submitted to Molecular Biology and Evolution (MBE) on January 18, 2012

ABSTRACT

Purifying selection often results in conservation of gene sequence and function. The most functionally conserved genes are also thought to be among the most biologically essential. These observations have led to the use of sequence conservation as a proxy for biological conservation. Here we describe two genes that are exceptions to this pattern. We show that lack of sequence conservation among orthologs of *CG15460* and *CG15323* -- herein named *jean-baptiste (jb)* and *karr* respectively -- does not predict lack of functional conservation. These two *Drosophila melanogaster* genes are among the most rapidly evolving protein-coding genes in this species, being nearly as diverged from their *D. yakuba* orthologs as random sequences. *jb* and *karr* are both expressed at an elevated level in larval males and adult testes, but they are not accessory gland proteins and their loss does not affect male fertility. Instead, we found that knockdown of these genes in *D. melanogaster* via RNA interference causes male-specific viability defects. These viability effects occur prior to the third instar for *jb* and during late pupation for *karr.* We show that sequences syntenic to *jb* and similar to *karr* are also expressed testes-

specifically in *D. yakuba, D. erecta*, *D. simulans* and *D. sechellia.* These genes maintain similar expression patterns and gene structure across species despite very low levels of sequence conservation. While standard tests for non-neutral evolution could not reject neutrality, other data hint at a role for natural selection. Together these data provide a clear case where a lack of sequence conservation does not imply a lack of conservation of expression or essential function.

INTRODUCTION

A cornerstone of molecular evolution is that sequence conservation and functional conservation go hand-in-hand. This makes sense as a protein's function is related to its amino acid sequence. Similarly, functional conservation is commonly considered an indicator of how biologically or evolutionarily essential a gene is. These principles are so universally accepted that it is common practice to use molecular evolutionary conservation to identify the most functionally important parts of proteins (Friedman et al. 2009; Temple et al. 2010; Marks et al. 2011). Following similar logic, "ultraconserved" elements have been identified across numerous taxa and at various evolutionary distances (Bejerano et al. 2004). These ultraconserved sequences are under strong purifying selection (Katzman et al. 2007), and as a result it is assumed that they would be required for life. Surprisingly, mice carrying knockouts for four ultraconserved elements showed no measurable defects (Ahituv et al. 2007), suggesting that ultraconserved elements may not always (or even usually) be as essential as expected. This fact hints that the relationship between sequence conservation, functional conservation and biological importance may not be as robust as commonly assumed.

At the other end of the spectrum, DNA and protein sequences can change rapidly for a variety of reasons. Often the most rapidly evolving sequences do not have conserved function and are evolving under relaxed purifying selection. For example, pseudogenes show high rates of sequence evolution and are assumed to be nonfunctional (Li et al. 1981; Daines et al. 2011). Natural selection can also drive rapid sequence divergence. Van Valen (1973) theorized that organisms and their genes may both be forced to evolve rapidly to meet the demands of a changing environment. Empirical data support this hypothesis. Many genes vital to immunity (Sackton et al. 2007; Obbard et al. 2009) and sexual function (Turner and Hoekstra 2006) evolve at elevated rates and show molecular signatures of positive selection.

In Drosophila, male-biased genes evolve particularly rapidly, often as a result of positive selection. Genes specific to male tissues are more likely to be orphans (have no known orthologs) and have higher rates of molecular evolution than genes expressed in other tissues or only in females (Haerty et al. 2007). The male accessory gland proteins (*Acps*) in Drosophila are a classic case of sexual conflict driving rapid molecular evolution. *Acps* are expressed in the male, are transferred to females during sex, and perform functions that benefit males -- sometimes at the expense of females (Chapman et al. 2001; Chapman et al. 2003; McGraw et al. 2004; Adams and Wolfner 2007; Avila and Wolfner 2009). Overall, *Acps* are among the most rapidly evolving genes in Drosophila though they perform functions vital to fitness.

Some *Acps* are so diverged that identifying orthologs in closely related species is difficult (Wagstaff and Begun 2005a, b, 2007). This finding raises the possibility that some functional genes in Drosophila are evolving even more rapidly than these *Acps* - perhaps so quickly that orthologs have not been identified in even the closest relatives. But what would such genes do, and can function be maintained in the face of rapid evolutionary change?

Here, we identify two genes in *Drosophila melanogaster* that are evolving so rapidly that they initially appeared to be lineage-specific orphans. These genes are have testes-biased expression and are important to male viability. We identified putative orthologs in *D. yakuba* and *D. erecta* and showed that their expression level and pattern was conserved despite low levels of both amino acid and nucleotide sequence conservation. Finally, while molecular evidence is inconclusive about the role of positive selection on the evolution of these genes, they are probably the two most rapidly evolving genes yet characterized in Drosophila. Because these genes are so rapidly changing but have conserved expression patterns, we propose to name *CG15460 jean-baptiste (jb) CG15323 karr* in homage to Jean-Baptiste Alphonse Karr, the author of the phrase "the more things change, the more they stay the same."

MATERIAL AND METHODS

Screen for candidate genes

To find extremely rapidly evolving genes in *D. melanogaster*, we searched for genes that appeared to be lineage-specific (following Levine et al. 2006). Briefly, genes in *D. melanogaster* were compared by local BLAST to *D. yakuba, D. erecta,* and *D. annanassae*. Genes with an *e*-value > 0.000001 in all three species and good EST support in *D. melanogaster* were considered candidate *D. melanogaster*-subgroup specific genes ("orphans"). We aligned candidates to all insect genomes using FlyBase's BLAST (Tweedie et al. 2009) and removed genes that had been retained in *D. melanogaster* and other more diverged species. We also performed BLAST against NCBI's nr database and removed candidates that were or contained known transposable elements, microbial genes, or other genome annotations.

We searched for the remaining candidates in other species (*D. yakuba, D. simulans, D. sechellia* and *D. erecta*) using UCSC's whole genome chained BLASTZ alignments, which are more sensitive to highly diverged hits than BLAST or BLAT (Chiaromonte et al. 2002). We then used the UCSC and Flybase genome browsers to ask whether the *D. yakuba, D. erecta, D. simulans,* and *D. sechellia* chained BLASTZ alignments covered annotated genes. We retained candidate genes that matched at least one annotated gene in all four species.

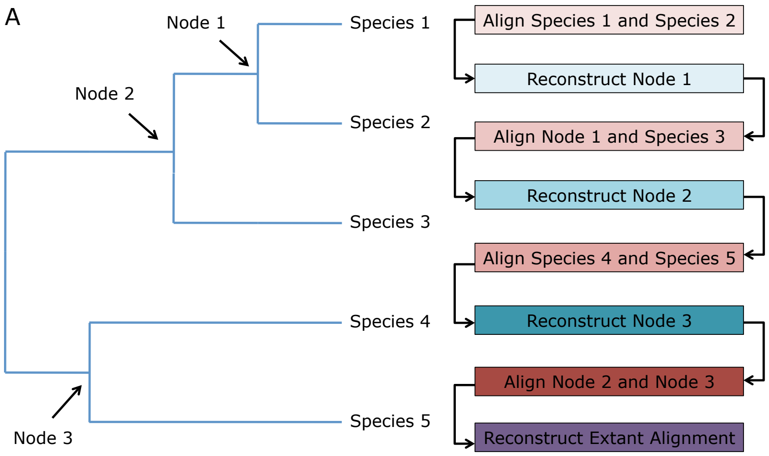
Molecular evolutionary analyses

We aligned the extended gene region (5-10kb surrounding the gene) of each candidate and its putative orthologs (see Table 3.1) to one another using MAUVE (Darling et al. 2004; Darling et al. 2010) to determine if the putative orthologs were colinear to the *D. melanogaster* gene. When colinearity existed, we performed a progressiveMAUVE multiple alignment assuming colinearity (progressiveMauve --collinear --seed-family --disable-backbone) and input the alignment into PAML's baseml (Yang 2007). We estimated the per base pair rate of substitution along the gene region. We counted the number of fixed differences between *D. melanogaster* and *D. simulans* in 500 bp windows along the alignment, then aligned the 39 *Drosophila melanogaster*

Raleigh genomes (www.dpgp.org) to these regions and calculated polymorphism (π) in each window. We also calculated Tajima's *D* (Tajima 1989) and Fu and Li's *D* and *F* Fu and Li 1993 for 500 base pair windows across the region using DNAsp v5 (Librado and Rozas 2009).

The high level of divergence between sequences made automated alignment of extant genes difficult. We reconstructed the ancestral sequences for each node using PAML's codeml (Figure 3.1) and used the reconstructed nodes to facilitate alignment. The most closely related extant genes were aligned pairwise by translated clustalW (Thompson et al. 2002), and then remapping to the coding sequences. We used codeml to reconstruct the most likely ancestral state from each pair of sequences. The internal nodes were aligned to one another or to related extant sequences as appropriate (Figure 3.1). This process was repeated until the common ancestral sequences for the *D.* yakuba/*D. erecta* orthologs were aligned to the common ancestral sequences in the *D. melanogaster* species subgroup. The extant sequences were then aligned to one another using these guide alignments.

Next, we used PAML's codeml to compare several models of codon evolution (e.g. branch-selection, site-selection, neutral). We used log-ratio tests to determine if any models were significantly better than the neutral model. We used the alignment of *D. melanogaster* and *D. simulans* along with the 39 DPGP Raleigh lines (www.dpgp.org) to estimate the number of silent and non-silent fixed differences and polymorphisms within the protein-coding regions. We compared these values using the McDonald-Kreitman test (McDonald and Kreitman 1991).



B *jean-baptiste* (*CG15460*) – default CLUSTALw alignment

```
Dmel_CG15460         MSQRNFKMPNPNNESHSYFLRNIPEELQPQFTRGFNQWMGNQS--TMTKGLPSNTVNKSA
Dsim_GD15539_co      --------------------------MQSQVPR----GMGNQS--TQTKTSPTTTANKST
Dsec_GM22677_co      -------MSNPNNETDSYFLRNFPHLVQPQIPRGFNQGMSNQS--TQTKTSPCKTANKST
Dyak_GE15353_co      MPNQSKASSPGIKETPSCSHHEFLMQEQPQQPRGSKVASTQTQDWPSDMVDQLTGAQDSP
Dere_GG17996_co      --MPNKNKSKAGPAGKKEALSSSPHEFRTPELRNKKANMGTQS--APKLVDQSTQTPDSP
                                                :.   *         .  .       . . .*.

Dmel_CG15460         QTASVNDGNLQASVIAMLAGMDSILDMEQPNRSPSREEHERLNELLF-------------
Dsim_GD15539_co      QTEPGNDGNVQASVIAMLTGLDSILDMKP-SRSPSPEEHQRLNELLS-------------
Dsec_GM22677_co      QTEPVNDDNVQASVIAVLAGLDSILDLQLRSRSPSPEEHQRLNEMLTG------------
Dyak_GE15353_co      SDMVGQLPGAQDSPSNMVDKLPGAQDSPSDMVDKLPGAQDSPSNMVNQSTQTDLSKLFMS
Dere_GG17996_co      SPALGFDRELQVRVTRMMGIMNAIMLRKVQSVLDR--NPANLDS----------------
                     .         *     ::  : .                   ..

Dmel_CG15460         ---SSNLL--MLDVKKRTFPVEDPTGYLTSVSEQNPDGNPLAKRLKLERPQ---------
Dsim_GD15539_co      ---NCFLLREVLESRKRTLPEEELTGLLTTILEQSPDGKPLAKRQKPERQEEESSGQPHQ
Dsec_GM22677_co      --DNCDLLREVLESRKRALPVDDLTGLLTTILERNPDGKPLPKRQKLECPQ---------
Dyak_GE15353_co      GEDDINLQLGTLGLRKRVAPEEDPTGDLTAVPEQNRNGNPLAKRQKVEGTQYR-------
Dere_GG17996_co      ----ISLLQSVLDDLT-----GNLTAVAELLPEQNPDGDPPAKRIKLERNQ---------
                           *    *   .      : *.    : *:. :*.* .** * *  :
```

C *jean-baptiste* (*CG15460)* – alignment using ancestral reconstruction

```
Dmel_CG15460         MSQRNFKMPNPNNESHSYFLRNIPEELQPQ--FTRGFNQW-----MGNQSTMTKGLP---
Dsim_GD15539_co      --------------------------MQSQ--VPRG---------MGNQSTQTKTSP---
Dsec_GM22677_co      -------MSNPNNETDSYFLRNFPHLVQPQ--IPRGFNQG-----MSNQSTQTKTSP---
Dyak_GE15353_co      --------------------------MPNQ--SKASSPGIKETPSCSHHEFLMQEQPQQP
Dere_GG17996_co      --------------------------MPNKNKSKAGPAGKKEALSSSPHEFRTPELR---
                                               :  :     .          . :.

Dmel_CG15460         --SNTVNKSAQTA-------SVNDGNLQASVIAMLAGMDS----ILDMEQP------NRS
Dsim_GD15539_co      --TTTANKSTQTE-------PGNDGNVQASVIAMLTGLDS----ILDMKP-------SRS
Dsec_GM22677_co      --CKTANKSTQTE-------PVNDDNVQASVIAVLAGLDS----ILDLQLR------SRS
Dyak_GE15353_co      RGSKVASTQTQDWPSDMVDQLTGAQDSPSDMVGQLPGAQDSPSNMVDKLP-GAQDSPSDM
Dere_GG17996_co      --NKKANMGTQSAP-KLVDQSTQTPDSPSPALGFDRELQVRVTRMMGIMN-------AIM
                        . ..  :*              :  :  :.     :     ::.

Dmel_CG15460         PSREE--HE------RLNELLF--SSNLLMRTLLDVKK---------------PVEDPTG
Dsim_GD15539_co      PSPEE--HQ------RLNELLSGDNCFLLVRTLLESRK---------------PEEELTG
Dsec_GM22677_co      PSPEE--HQ------RLNEMLTGDNCDLLVRALLESRK---------------PVDDLTG
Dyak_GE15353_co      VDKLPGAQDSPSNMVNQSTQTD--LSKLFM---SGEDDGTLGLRKRVAPEEDPINLQLTG
Dere_GG17996_co      LRKVQSVLD------RNPANLD--SISLLQ---SVLDD---------------LTGNLTA
                             :      .           *:        .                  : *.

Dmel_CG15460         YLTSVSEQNPDGNPLAKRLKLERPQG
Dsim_GD15539_co      LLTTILEQSPDGKPLAKRQKPER-QE
Dsec_GM22677_co      LLTTILERNPDGKPLPKRQKLECPQG
Dyak_GE15353_co      DLTAVPEQNRNGNPLAKRQKVEGTQY
Dere_GG17996_co      VAELLPEQNPDGDPPAKRIKLERNQV
                         : *:. :*.* .** * *  *
```

Figure 3.1 - Using ancestral sequence reconstruction to guide alignment We aligned the amino acid sequences of the most closely related species to one another, then used PAML (codeml) to reconstruct the ancestral nucleotide sequence for each node (Methods). We continued this process until Nodes 2 and 3 could be aligned to one another. Finally, we remapped the extant sequences onto this alignment (A). Results of this procedure (C) are shown compared to a ClustalW alignment (B).

Sequence similarity of *D. melanogaster* orthologs and rapidly evolving genes

We used EMBOSS' water pairwise alignment program (Rice et al. 2000) to determine the sequence similarity of all *D. melanogaster* genes to their orthologs in *D. yakuba* and *D. simulans.* We pulled the best hit from BLAT and found the percent identity and proportion of the *D. melangoaster* sequence that aligned to the ortholog (proportion matching). We plotted these values using R (RDevelopmentCoreTeam 2009), and compared the percent identity and proportion matching to 1) the rapidly evolving genes we identified and 2) 100 randomly generated 500 base pair sequence pairs.

Tissue collection and dissection

Male reproductive tracts were dissected on ice from whole flies (*D. yakuba, D. simulans, and D. melanogaster*) in PBS. Male reproductive tracts and carcasses were each pooled and then flash frozen in liquid nitrogen. Whole females and males of each species were collected and flash-frozen. *D. melanogaster* and *D. yakuba* male reproductive tracts were further dissected into accessory glands and testes in PBS and flash frozen. *D. melanogaster* third instar larvae were sexed by identification of genital discs following *Drosophila protocols* (Blair 2000), then flash-frozen. Testes were also dissected from males carrying a null mutation at the gene *tombola* (*tomb*^{GS12862}, Jiang et al. 2007, stock generously supplied by Dr. Helen White-Cooper), and sons of females mutant for the *tudor* gene (Bloomington stock #1786, Boswell and Mahowald 1985).

Gene expression analyses

We mined expression information from online databases - FlyAtlas (Chintapalli et al. 2007), modENCODE RNAseq data (Graveley et al. 2011), Baylor RNAseq data (Daines et al. 2011), and FlyTED: Testes expression database (Zhao et al. 2010). We then extracted RNA from at least two biological replicates of each dissected tissue using TRIZOL reagent (Invitrogen, Grand Island, NY #15596-026), and made cDNA using M-MLV reverse transcriptase (Invitrogen, Grand Island, NY #28025013). We performed relative qRT-PCR quantification using gene-specific primers and a control primer that worked across all species (*Actin5c*). All qRT-PCR was performed using two technical replicates. 5' and 3' RACE were performed following manufacturer's instructions on *D. melanogaster, D. yakuba,* and *D. simulans* testes RNA using the FirstChoice RLM-RACE kit from Ambion (Grand Island, NY #AM1700) and nested gene-specific primers.

RNAi knockdown

Virgin *Actin*-GAL4 females (P[Act5C-GAL4]25FO1, Bloomington #4414) were collected and crossed to lines carrying UAS-RNAi constructs for *CG15323 (karr)*, and *CG15460 (jb)* (www.VDRC.org #35689 and #43403, Dietzl et al. 2007). *CyO* (control) and straight winged (RNAi) progeny of both sexes were counted and collected. We confirmed RNAi knockdown using the same qRT-PCR methods as described above but using *gpdh* instead of *Actin* as the control gene.

Viability assays

To estimate effects on adult viability, we simply counted the number of control (*CyO*) and RNAi (straight-winged) progeny eclosing from each RNAi cross (described

Table 3.1: putative orthologs of *CG15323* (*karr*) and *CG15460* (*jb*)

	Gene name	Colinearity	%Matching	Length matched
Dmel/CG15323 (karr)	*Dmel/CG42580*	NA	59.9%	0.902
	Dmel/CG34332	NA	67.8%	0.944
	Dsim/GD15552	5' (exon 2)	49.0%	0.941
	Dsec/GM17452	5'	53.7%	0.941
	Dsim/GD17478	3'	47.4%	0.948
	Dsim/GD17479	3’	53.3%	0.948
	Dsim/GD15554	3'	53.3%	0.948
	Dsec/GM22694	3'	67.8%	0.948
	Dsec/GM13264	3'	67.8%	0.948
	Dsim/GD15543	no	63.0%	0.912
	Dsim/GD17496	no	63.4%	0.902
	Dsec/GM23042	no	60.5%	0.807
	Dsec/GM23010	no	71.0%	0.948
	Dyak/GE17891	no	50.2%	0.951
	Dere/GG19692	no	46.4%	0.944
Dmel/CG15460 (jb)	*Dsim_GD15539*	yes	77.2%	0.800
	Dsec_GM22677	yes	80.2%	0.922
	Dyak_GE15353	yes	53.3%	0.924
	Dere_GG17996	yes	53.1%	0.968
	Dsec_GM23024	no	78.0%	0.924
	Dyak_GE15357	no	46.1%	0.962
	Dyak_GE17873	no	48.7%	0.966
	Dere_GG19284	no	50.7%	0.880
	Dere_GG18002	no	50.7%	0.968
	Dere_GG17998	no	54.6%	0.966
	Dere_GG19283	no	54.7%	0.958
	Dere_GG19282	no	52.4%	0.968
Random sequences	random 500bp	NA	44.9%	0.890

above). To determine the stage at which lethality was occurring, we crossed the same RNAi lines to a stock with the same *Actin*-GAL4 and *CD8*::UAS-GFP on the same chromosome (kindly donated by S. Chen). RNAi or control status can be ascertained at any stage (RNAi larvae/pupae/adults will express GFP). We collected larvae from the cross during the late third instar ("wandering")/prepupal stage, and sorted by GFP expression and sex (Blair 2000). We then allowed each type to continue development and counted the number that survived, or that died prior to pupation or prior to eclosion.

Fertility assays

We used a sperm exhaustion assay to estimate the effect of RNAi knockdown of *CG15460 (jb)* and *CG15323 (karr)* on male fertility. In this assay (modified from Sun et al. 2004), single males are challenged with two virgin females per day across a five-day period. Males with defects in sperm production should produce fewer offspring per female over the assay period. We used a linear model ($mean\_offspring = genotype + day + genotype \times day + \varepsilon$) to determine if there were significant effects of genotype (indicating a general fertility defect), or a genotype by day interaction effect (indicating a defect in sperm production).

RESULTS

CG15460 (jb) and *CG15323 (karr)* are among the most rapidly evolving genes in *Drosophila melanogaster*

We identified two genes in *D. melanogaster* that have evolved so rapidly that their *D. yakuba* orthologs had not previously been identified. Following Levine et al.

(2006), we compared genes in *D. melanogaster* by local alignment (BLAST) to the *D. yakuba, D. erecta,* and *D. annanassae* genomes (Clark et al. 2007). Genes matching poorly to all three species but with EST support in *D. melanogaster* became candidate *D. melanogaster*-subgroup specific genes. We aligned these to all insect genomes and removed genes that had been retained in any other species. This eliminated genes that were selectively lost in the *D. yakuba, D. erecta,* and *D. annanassae* genomes. To distinguish rapid evolvers from true lineage specific genes, we searched the BLASTZ alignments from UCSC and retained genes that overlapped at least one *D. yakuba* and *D. erecta* gene. This search yielded *CG15460* and *CG15323* hereafter referred to as *jean-baptiste (jb)* and *karr* respectively.

jb and *karr* aligned to annotated genes in all five sequenced species in the *D. melanogaster* subgroup, but could not be found in distantly-related species. Some of the other candidates are colinear to non-coding sequences in *D. yakuba* and *D. erecta* - these other genes likely evolved *de novo* from the non-coding sequences (Levine et al. 2006) or may be misannotated as non-coding regions in these other species. *karr (CG15323)* was originally reported as a *de novo* gene, but the BLASTZ alignment showed weak similarity to the *D. yakuba* gene *GE17891* and the *D. erecta* gene *GG19692*; see Table 3.1. The *jb* CDS aligned to multiple genes in *D. sechellia*, *D. erecta* and *D. yakuba*. One of these copies flanks the colinear *jb* ortholog in each species, suggesting that this gene is a tandem duplicate and one copy was lost in the *D. melanogaster* lineage. Additionally, *D. erecta* and *D. yakuba* also have a few distributed copies of *jb* (Table 3.1). *karr* has potential paralogs within *D. melanogaster* and matches to multiple genes in *D. simulans* and *D. sechellia*, but only matches one gene in *D. yakuba* and *D. erecta.* Though the *D.*

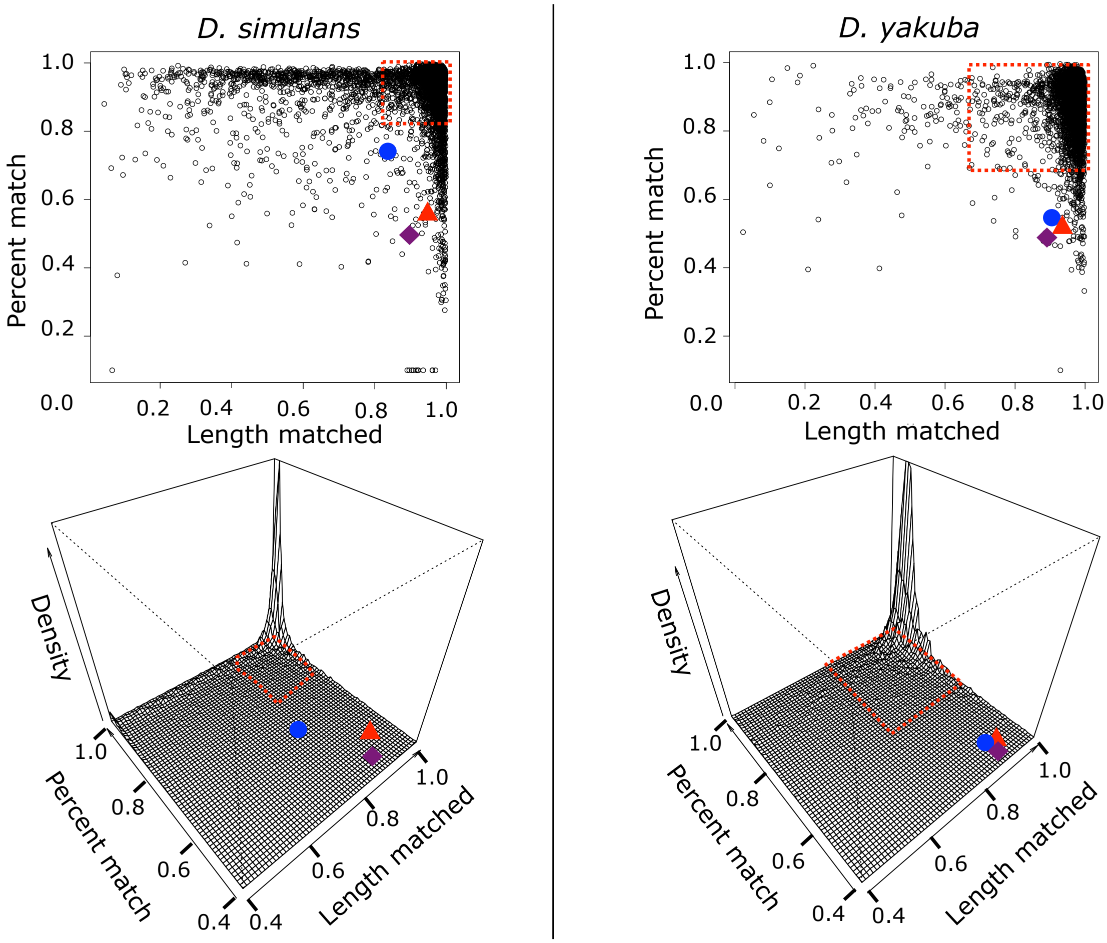


Figure 3.2 - *jb* and *karr* are among the most diverged genes in *D. melanogaster*

We aligned the nucleotide sequence from the CDS of every gene in *D. melanogaster* to its annotated orthologs in *D. simulans* and *D. yakuba* using EMBOSS' water aligner (black dots). We also aligned *jb* (blue) and *karr* (red) to their putative orthologs from *D. simulans* and *D. yakuba*. The red dashed box shows where 90% of known protein-coding genes lie. Both *jb* and *karr* fall outside this box in each species. Finally, we generated 100 pairs of random 500bp nucleotide sequences and align each pair of sequences to each other to estimate the average similarity of random sequences. The average sequence conservation and length matched across the 100 replicates is in purple. Both genes are nearly as dissimilar to their *D. yakuba* orthologs as the average pair of randomly generated nucleotide sequences.

yakuba and *D. erecta* copies are not colinear to the copies in *D. melanogaster*, they are colinear to one another (see Table 3.1).

***jb* and *karr* and their putative orthologs are among the least similar ortholog pairs in Drosophila**

The CDSs of *jb* and *karr* and their *D. simulans* and *D. yakuba* orthologs have among the lowest sequence similarity of any orthologous pairs in Drosophila (Table 3.1, Figure 3.2). We also generated and aligned (EMBOSS) 100 pairs of randomly generated DNA sequences to determine the lowest expected similarity scores using this method. *jb* and *karr* are among the top 10% most diverged orthologous pairs in both *D. simulans* and *D. yakuba* (Figure 3.2, *jb* is blue and *karr* is red) and similarity to the *D. yakuba* orthologs is nearly as weak as similarity between random sequences (purple dots). It is therefore unsurprising that these genes were not annotated as orthologs in this species. However, in contrast to some other highly diverged genes, both *karr* and *jb* align along most of their length and appear to have conserved intron/exon boundaries and splice forms (see below).

***jb* and *karr* are strongly expressed in male tissues**

The high level of sequence divergence between these genes and their putative orthologs makes confirmation of true orthology difficult. Similar expression patterns would support orthology and would suggest that these divergent orthologs perform similar functions. Data from FlyAtlas (Chintapalli et al. 2007) and RNA-seq (Daines et al. 2011; Graveley et al. 2011) show that expression in *D. melanogaster* adults is highest

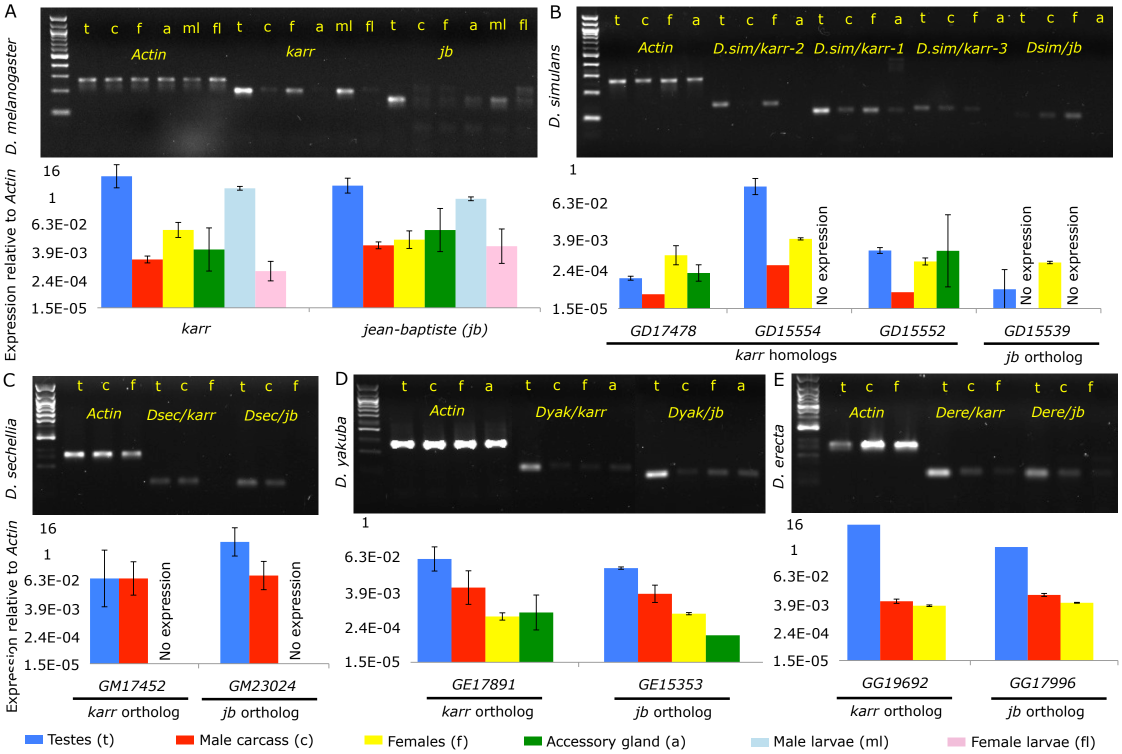


Figure 3.3 - Expression of *karr* and *jb* are male biased and this pattern is conserved across five species

RT-PCR (gels) and qRT-PCR (bar graphs) measurements of gene expression are shown for *D. melanogaster* (A), *D. simulans* (B), *D sechellia* (C), *D. yakuba* (D) and *D. erecta* (E). In each species, expression of putative *jb* and *karr* orthologs was compared between the testes, the remaining male tissues ("carcass"), and whole females. In *D. melanogaster, D. yakuba,* and *D. simulans*, male accessory glands were also assayed. Multiple orthologs of *CG15323* exist in *D. simulans* and *D. sechellia* and expression was measured for the three "colinear" copies in *D. simulans*. Only *GD15554* (*Dsim/karr-1*) and *GM17452* (*Dsec/karr-1*) in *D. sechellia* showed the characteristic expression pattern seen in the other species. Expression of *jb* and *karr* was also measured in male and female *D. melanogaster* larvae.

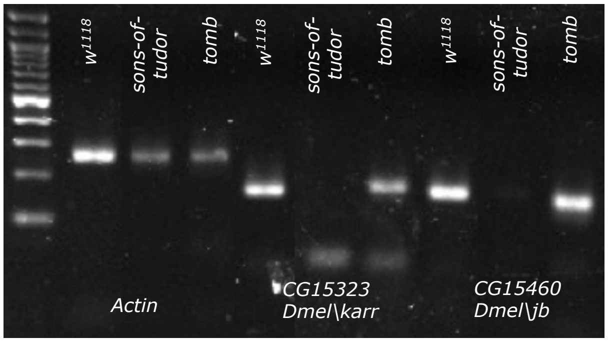


Figure 3.4 - Expression of *karr* and *jb* is dependent on the germline but not on the meiotic arrest gene *tombola*

We measured expression of *jb* and *karr* in testes from w^{1118} males, *sons-of-tudor* males, and *tombola* males using RT-PCR. Expression of both genes was reduced in the *sons-of-tudor* males but not in the *tombola* males, indicating that a germline is required for expression of *karr* and *jb*, but that neither gene is dependent on *tombola*.

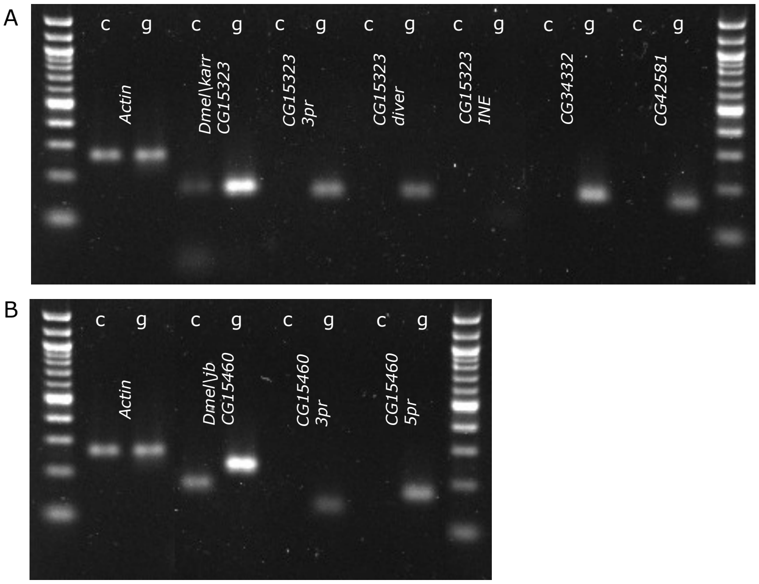


Figure 3.5 - No expression of *karr* and *jb* flanking regions or transposable elements

RT-PCR was used to contrast expression of *karr* (A) and *jb* (B) with neighboring non-coding sequences including the transposable elements *diver* and *INE*. PCR was also performed using genomic DNA as a template to confirm primer specificity. While the genes themselves could be amplified from cDNA, the neighboring non-coding sequences could be amplified from genomic DNA but not cDNA indicating the flanking sequences are not expressed.

in male tissues, and can be detected from the third larval instar through adulthood (flybase.org/cgi-bin/gbrowse). We confirmed these patterns by measuring expression of *jb* and *karr* in the testes, accessory glands, the remaining male carcass, and whole females. Both genes showed peak expression in the testes (Figure 3.3a). Expression was weak (*jb*) or undetectable (*karr*) in the accessory glands, demonstrating that *karr* and *jb* are not accessory gland proteins (ACPs). We confirmed that expression of both genes is reliant on the germline by measuring expression in testes from mutant flies lacking a male germline (Boswell and Mahowald 1985 *sons-of-tudor*, Figure 3.4). Expression was greatly reduced. Many genes expressed in male meiotic cells are under the control of so-called meiotic arrest genes (e.g *tombola* Jiang et al. 2007), but both *karr* and *jb* were expressed at normal levels in *tomb*^{GS12862} (*tombola* null) testes (Figure 3.4). This implies both genes function in parallel to or independently of the meiotic arrest pathway.

Next, we compared expression of the presumed orthologs in adult male testes, male carcass, and female *D. simulans, D.sechellia, D. yakuba, and D. erecta*. We also measured expression in accessory glands from *D. simulans* and *D. yakuba*. The orthologs of both genes showed peak expression in the testes of *D. sechellia, D. yakuba, and D. erecta*. *D. simulans* was more complicated, because we measured expression of three of the duplicate copies of *karr*. *GD15554* (*Dsim/karr-1*) shows a nearly identical expression pattern to *D. melanogaster*, but the other two copies (*Dsim/karr-2* and *Dsim/karr-3*) have weak expression in all tissues. We next verified that expression of orthologs was not due to nonspecific "background" transcription. First, we used RT-PCR to confirm there was no expression of sequences directly up- or down-stream of the annotated mRNA in the testes (Figure 3.5). We eliminated the possibility that transposable elements in proximity of *karr* could be driving expression by confirming that flanking transposons were not expressed (Figure 3.5). Additionally, matching the pattern observed in the *D. simulans* paralogs, neither of the *D. melanogaster* "paralogs" of *CG15323* were expressed in the testes (Figure 3.5A). Finally, we used 5' and 3' RLM-RACE to verify the expression and sequence of the mature mRNA in *D. yakuba* (Supplemental data). We confirmed the annotated CDS for *GE17891* (*Dyak/karr*) using both 5' and 3' RACE, and found additional 5' and 3' sequence, presumably representing unannotated 3' and 5' UTRs. We only were able to sequence a fragment of the 5' RACE product for *GE15353* (*Dyak/jb*), but this matched 55 base pairs just 5' of the annotated CDS. The RACE results indicate that stable mRNA are produced from the putative orthologs of *jb* and *karr*. These data imply that despite extremely rapid rates of protein divergence between species, these genes have retained the same gene structure and pattern of strong expression in the male germline.

RNAi silencing of these rapidly evolving genes is semi-lethal in male *Drosophila melanogaster*

We used RNA interference to knock down expression of *karr* and *jb* in *D. melanogaster*. We drove the expression of UAS-RNAi constructs for each gene by crossing RNAi stocks to a ubiquitous GAL4 driver (*Actin*-GAL4) and confirmed by qRT-PCR that expression of each gene was successfully knocked down (data not shown). We found a significant reduction in the number of RNAi male offspring compared to the other offspring classes (*Fisher's exact test,* for *karr* $P = 0.022$; for *jb* $P = 0.028$, Table 3.2). This result was unexpected as expression appeared to be strongest in the male

Table 3.2 – RNAi of either *jb* or *karr* is semi-lethal in males

	Control		RNAi		Fisher's exact test
	Males	Females	Males	Females	*P*-value
Jb (CG15460)	511	558	393	503	0.028
Karr (CG15323)	476	532	439	593	0.022

Table 3.3 - *karr* and jb are important to larval and pupal development respectively

		UAS:RNAi	UAS:RNAi	UAS:RNAi	UAS:RNAi	Fisher’
		/*CyO*	/*CyO*	/*Actin*GAL4;	/*Actin*GAL4;	exact
				CD8:UAS-GFP	CD8:UAS- GFP	test
	Surviving to	Males	Females	Males	Females	*P*-valu
Jb	3^{rd} instar	106	105	122	134	0.642
(CG15460)	Eclosion	73	75	30	62	0.011
Karr	3^{rd} instar	147	108	164	183	0.013
(CG15323)	Eclosion	114	88	122	148	0.020

reproductive tract in adults. However, RNAseq data showed that both genes were expressed during larval development as well as in adults. As larvae were of mixed sex in the RNAseq experiment, we measured expression of both genes in third instar larvae after sorting by sex and found higher expression in males (Figure 3.3a). Lethality may be occurring during development or metamorphosis phenotype. To determine the stage of lethality, we crossed RNAi stocks to an *Actin*-GAL4 driver stock that also contained UAS-GFP, allowing identification of RNAi offspring of any stage by GFP expression. We sorted late third instar "wandering" larvae by both sex and GFP expression, then allowed these larvae to continue development, and scored the number of each genotype surviving to pupation and eclosion. We reconfirmed that there was a significant reduction in the number of successfully eclosed male RNAi offspring when compared to controls for both genes (Table 3.3). However, the stage of lethality differed between the two genes. For *jb*, a comparable number of all offspring types survived to the third larval instar, but a large proportion of the RNAi male pupae failed to eclose (25% eclosion rate versus 69% for controls). Observationally, *jb-RNAi* pupae arrested at the pharate stage, appearing fully developed inside the pupae with discernable eyes, wings, and legs. For *karr*, a smaller proportion of RNAi male offspring reached the third larval instar, but eclosion rates were similar across all groups. We conclude therefore that *karr* is important for development during either embryonic or early larval stages whereas *jb* acts during pupation.

We tested if RNAi flies had fertility defects, as would be expected given the strong expression in the testes and germline dependence of *jb* and *karr*. We set up a series of single-fly matings using RNAi and control males for both genes as well as a more

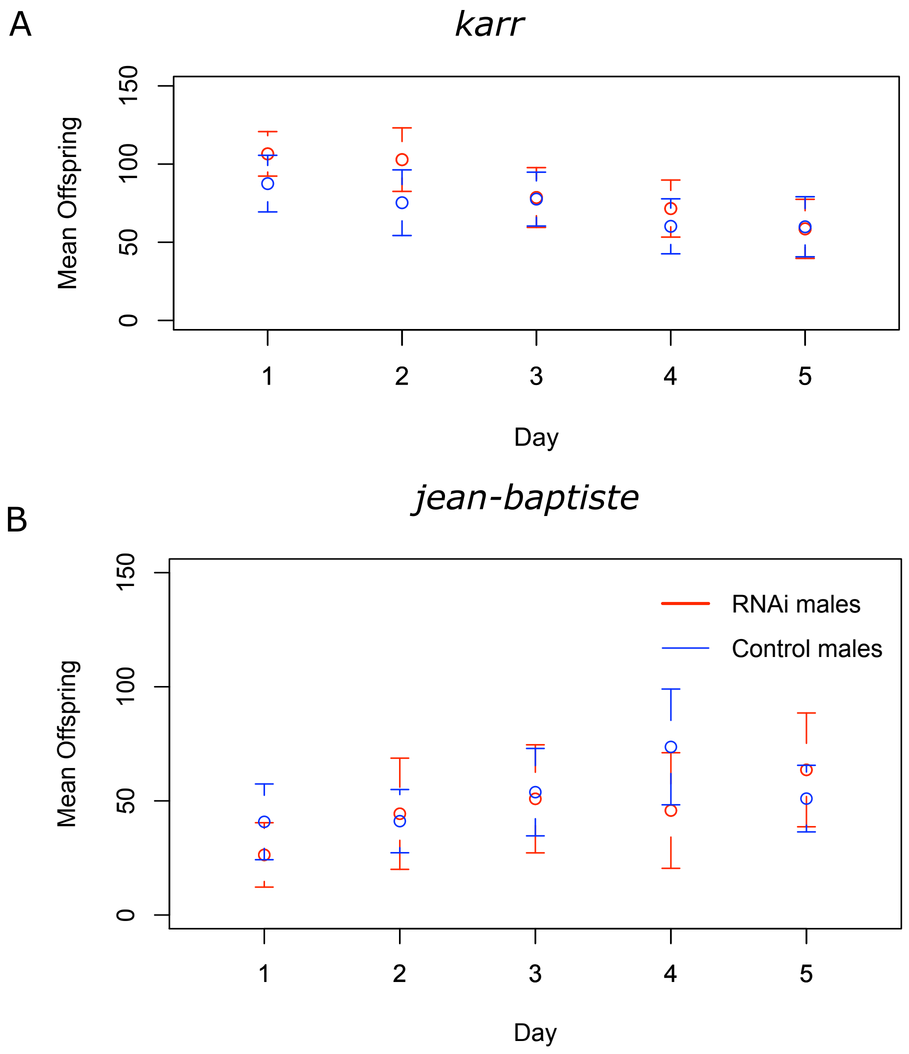


Figure 3.6 - Fertility was not affected by RNAi of *jb* or *karr*

karr (A) and *jb* (B) RNAi males did not have significantly reduced fertility when compared to their control siblings in a sperm exhaustion fertility assay (Methods). We measured the effect of genotype and the interaction between genotype and day on offspring produced per male and found no significant effect for either gene ($P > 0.05$ for all tests).

intensive fertility assay - sperm exhaustion (Sun et al. 2004). We found no difference between control and RNAi males in the number of offspring produced by either assay (Figure 3.6). Thus, despite being strongly testes expressed, these genes are not essential to male fertility.

jb **but not** ***karr*** **is colinear across the five Drosophila species in which it is found**

ProgressiveMAUVE (Darling et al. 2010) alignments of the 10kbp surrounding each putative ortholog from FlyBase in all five species showed that for *jb* there was a single, colinear region across all five species that included a gene with similar orientation and structure (Figure 3.7). The neighboring genes were present and highly conserved (although as previously mentioned, there was a tandem duplicate of *jb* in *D. sechellia,* and *D. yakuba* that was not present in *D. melanogaster*). However, the colinear orthologs to *jb* showed the weakest sequence similarity across the entire region. *karr,* on the other hand*,* was more complicated. A single ortholog is identifiable in *D. erecta* and *D. yakuba*, but in both *D. simulans* and *D. sechellia* multiple regions aligned suggesting recent gene duplication (Table 3.1). None of these genes are colinear to the *D. melanogaster* copy.

jb **is evolving at an elevated rate compared to flanking sequences and other rapidly evolving genes**

Because *jb* was colinear across all five species, we could reconstruct the evolutionary history of the gene region and the evolution of the protein. We hypothesized that the high level of divergence of *jb* was due to positive selection rather than simple neutral drift. Genes under positive selection are predicted to show high

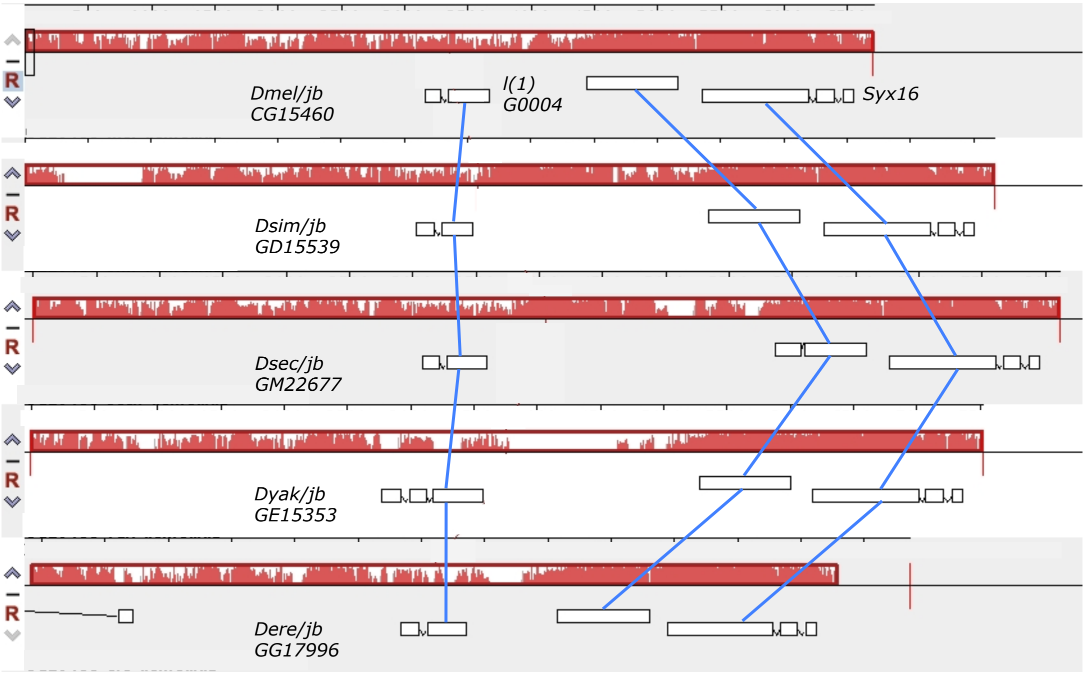


Figure 3.7 - *jb* is colinear to and shares a conserved gene structure with orthologs from four other *Drosophila* species

We used progressiveMAUVE to align the extended gene regions of *jb* and each of its four putative orthologs. We found that despite weak sequence conservation over the gene regions (red lines), the genes were colinear (blue lines), maintained their orientation relative to conserved flanking genes, and in all but one case have identical gene structure (the *D. yakuba* ortholog has an additional exon).

levels of divergence (especially nonsynonymous divergence) and low levels of polymorphism compared to sequences evolving neutrally or under purifying selection. We tested this concept using baseml (Yang 2007) to estimate the number of nucleotide substitutions occurring along all branches in 500 bp windows across the MAUVE multiple alignment. We estimated polymorphism in the same windows using population genomics data from DPGP (www.dpgp.org). As a positive control, we performed the same analysis on *ovulin* (*Acp26Aa*), a male-specific protein-coding gene known to have diverged under positive selection in the *D. melanogaster* subgroup (Aguadé 1998; Tsaur et al. 1998; Wong et al. 2006; Wong et al. 2010). The highest substitution rates in these gene regions (Figure 3.8, blue bars) were over the windows including the genes *jb* (Figure 3.8, top) and *ovulin* (Figure 3.8, bottom), suggesting that both genes are evolving more rapidly than their immediate genomic background. Conversely, polymorphism (π) was low over the windows containing *jb* and *ovulin* (Figure 3.8, red dots). We attempted to detect recent positive selection using Tajima's *D* and Fu and Li's *D* and *F*, but did not see significant enrichment in the windows overlapping *jb*. We hypothesize this is due to insufficient power because of how few polymorphic sites were present. Overall, *jb* has a similar pattern of regional nucleotide divergence and polymorphism as a gene known to be evolving under positive selection. We tested for positive selection acting on the *jb* protein in the lineage leading to *D. melanogaster* using the McDonald and Kreitman test (McDonald and Kreitman 1991) and polymorphism data from DPGP. We found that *jb* had high numbers of nonsynonymous differences between species, but few polymorphic sites (10 sites, Table 3.4). Thus, the absence of a signature of positive selection (P = 0.440) may again reflect weak power. In fact, the low level of polymorphism suggests

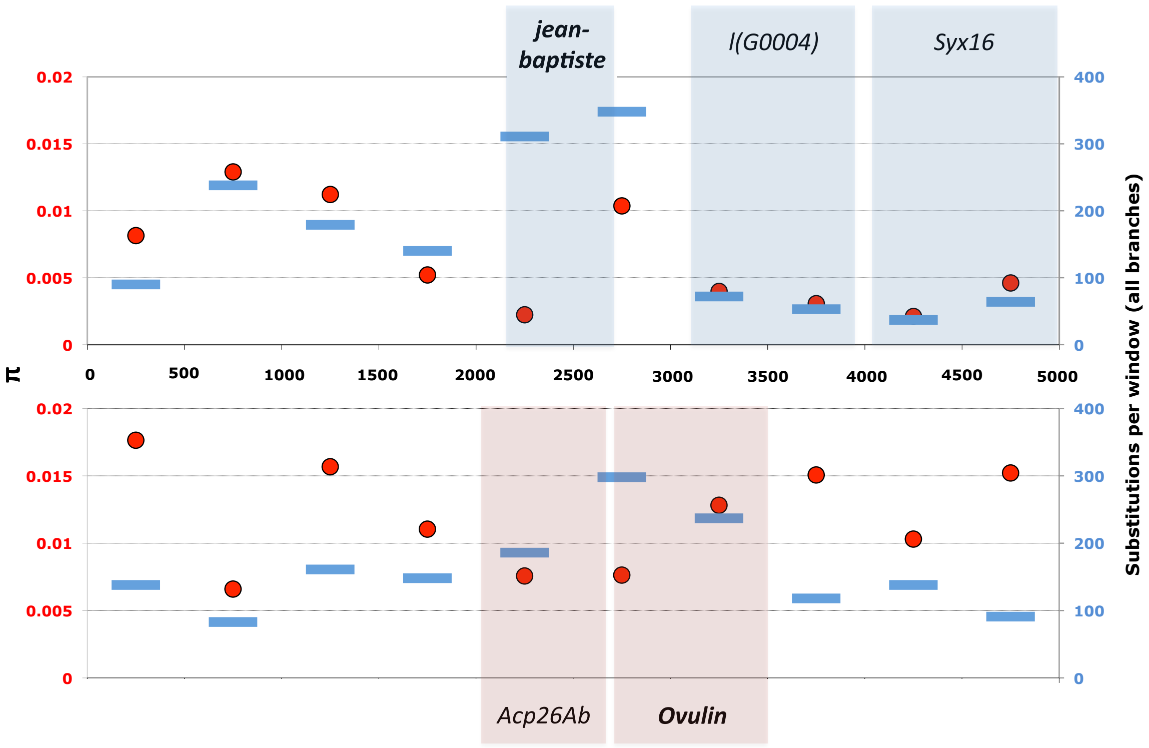


Figure 3.8 - *jb* has high levels of divergence but low levels of polymorphism relative to flanking sequence

We used PAML (baseml) to estimate the number of substitutions (blue bars) that have occurred along all branches in 500bp windows in the *jb* expanded gene region (top panel) and the *ovulin* gene region (bottom panel), a rapidly evolving male expressed gene known to have undergone positive selection. We also measured π (red dots) in the same windows using 39 Raleigh lines from the *Drosophila* 50 genomes data (www.dpgp.org). Gene models are shown above and below each panel.

Table 3.4 - Results of McDonald-Kreitman test on *jean-baptiste*

	Fixed differences		Polymorphisms		Fisher's exact test
	R	S	R	S	*P*-value
Jb (CG15460)	61	15	7	3	0.43
*Acp26Aa**	75	21	22	16	0.031

* from Tsaur, Ting, and Wu 1998, MBE

that the gene could have undergone one or more selective sweeps, and the high rate of substitution could have been caused by multiple sweeps.

As we were unable to distinguish whether evolution of *jb* is being driven by positive selection using polymorphism-based approaches, we next compared models of codon substitution in the *jb* protein across five species. If *jb* is evolving under positive selection, we expect to observe an elevated rate of nonsynonymous codon substitutions. Particular codons should be substituted at a level above the background of the gene (indicating positive selection acting repeatedly at these sites) or nonsynonymous substitutions should occur at an elevated rate along one specific lineage (indicating positive selection along that lineage). We contrasted site and branch models assuming selection to models assuming neutrality (codeml, Yang 2007), and saw no improvement using the selection models. Hence we were again unable reject the null hypothesis that *jb* is evolving under neutral drift alone. However, the overall rate of both synonymous and nonsynonymous protein codon substitution was rapid along all lineages and almost double that of the rapidly evolving gene *ovulin* (Figure 3.9).

The genomic dynamics of *karr* may be linked to the action of transposable elements

We observed that *karr* expanded its copy number in the three species of the *D. melanogaster* species subgroup through a number of large segmental duplications and rearrangements as well as dispersed duplication (Figure 3.10). In contrast to *jb*, the location of all *D. melanogaster*, *D. simulans,* and *D. sechellia* copies of *karr* differ from that of the homologs in the *D. yakuba/D. erecta* clade. We noted that all three potential paralogs in *D. melanogaster* had annotated transposable elements nearby (*diver* and *INE*).

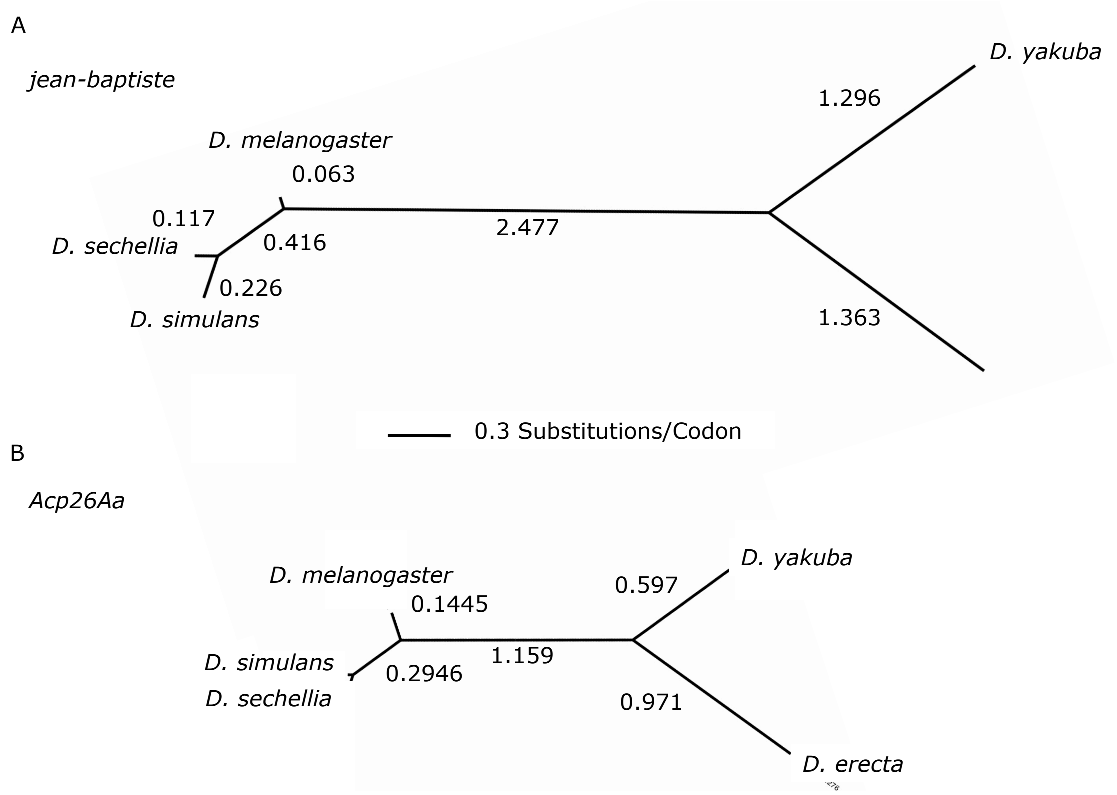


Figure 3.9 - *Jean-baptiste* protein is evolving at twice the rate of *ovulin* We used PAML (codeml) to estimate the rate of codon substitution between *jb* (A) and its putative orthologs, in comparison to the rapidly evolving gene *ovulin* (B) and found that the former had roughly double the rate of substitution along all branches. Branch lengths are to scale.

We searched the colinear gene regions in the five species for potential TEs, and found homology to *INE* and *diver* elements near every ortholog in *D. simulans* and *D. sechellia,* but no evidence for either TE in *D. yakuba* or *D. erecta* -- two species in which *karr* is single copy and colinear. This suggests that transposable elements were introduced into the common ancestor of *D. melanogaster*, *D. simulans* and *D. sechellia*, and the action of these elements may have induced a series of duplications and translocations of the *karr* gene region (Figure 3.10).

Of the two putative paralogs of *karr* in *D. melanogaster*, and the three colinear *D. simulans* homologs, qRT-PCR showed that only one gene from each species is strongly expressed in the testes (Figure 3.3a and b, Figure 3.6, RNAseq data shows that the paralogs of *karr* are also expressed in males, albeit weakly). This strong testes expression Of the two putative paralogs of *karr* in *D. melanogaster*, and the three colinear *D. simulans* homologs, qRT-PCR showed that only one gene from each species is strongly pattern is apparently ancestral, as it is shared by the *D. yakuba* and *D. erecta* orthologs (Figure 3.3d and e). Transposable elements - particularly active ones - often include regulatory machinery that can induce expression of neighboring genes suggesting that the association with transposable elements might be driving the expression of *karr* and its putative orthologs. We measured expression of the *diver* and *INE* elements near to *Dmel/karr*, but found no expression of *diver*, *INE* or other flanking sequences in the testes (Figure 3.5). We surmise that some of the putative orthologs of *karr* may have been duplicated and carried across the genome by transposable elements, but their expression patterns have been altered by their new genomic positions.

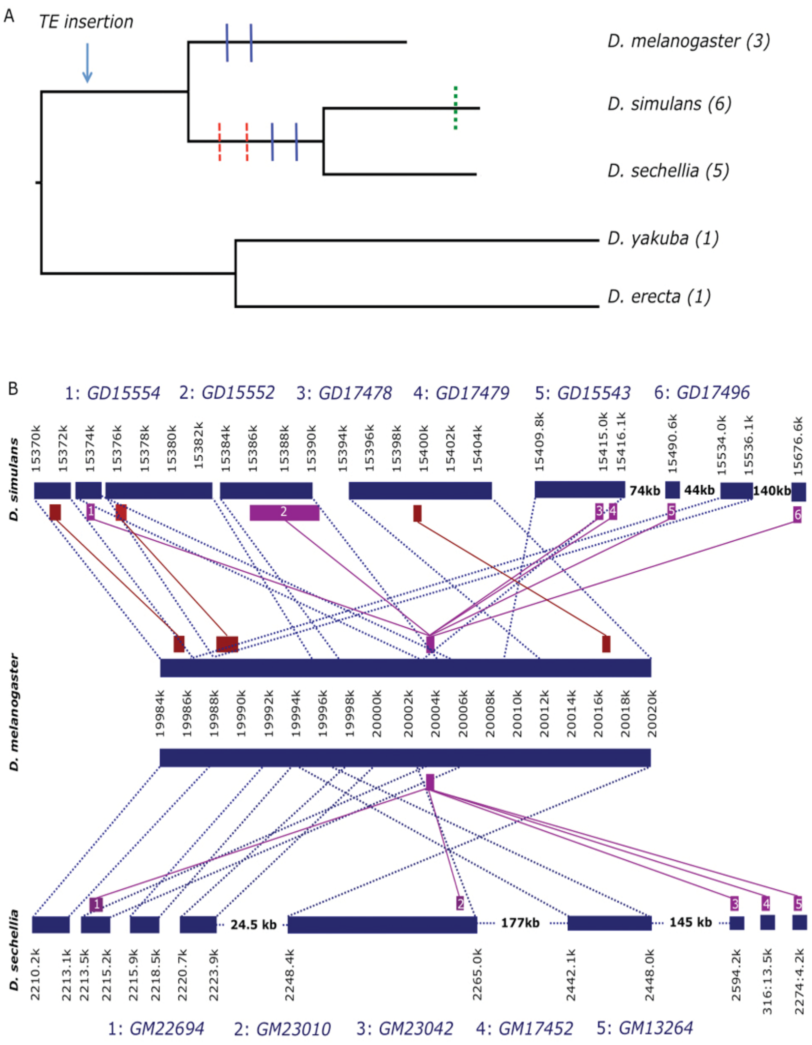


Figure 3.10 - Multiple copies of *karr* exist in the *D. melanogaster* species subgroup and appear to be TE associated. Panel (A) shows *karr* has multiple putative orthologs in each of *D. melanogaster, D. simulans* and *D. sechellia* (parenthesis show number of duplicates), and each is associated with one or more transposable elements (*diver* and *INE*). *D. yakuba* and *D. erecta* each have only a single copy and no evidence of the associated TEs. Solid lines indicate inferred large scale rearrangements, dashed lines gene translocations, and the dotted line a tandem duplication. Panel (B) shows that the region of the X chromosome containing *karr* has been duplicated, rearranged, and transposed multiple times in *D. melanogaster's* sister species *D. simulans* (top) and *D. sechellia* (bottom). The ends of colinear regions are shown as dotted lines, genes are shown as small blocks, and orthologs are connected by solid lines. *karr* orthologs are numbered and labeled. When distances between aligned regions are not to scale, the distance is as shown. In *D. simulans*, all copies are found on a 300kb region of the X chromosome. In *D. sechellia*, two of the copies are found on small, unordered scaffolds and the remainder are X-linked.

DISCUSSION

Functionally important genes are often evolutionarily constrained because amino acid sequence must be preserved to maintain a protein's catalytic or structural role. Here, we describe two genes that are startling exceptions to this pattern. *karr* and *jb* are among the most rapidly evolving putative protein-coding genes in Drosophila, yet knockdown of these genes in *D. melanogaster* via RNA interference causes male-specific developmental defects leading to semi-lethality. These genes are expressed strongly in male larvae and adult testes, and expression is reduced in the absence of a male germline. Yet despite their functional role, the rate of sequence divergence in these genes is so great that alignment of orthologs is difficult. Nevertheless, we found sequences syntenic to the *D. melanogaster* CDS in *D. yakuba* and *D. erecta.* These orthologs showed the same intron/exon structure and expression pattern as observed in *D. melanogaster.* Thus, despite low sequence conservation, these genes unexpectedly appear both structurally/functionally conserved and biologically essential.

We must, however, reconcile a number of seemingly contradictory findings. These genes are extremely rapidly evolving, and they are expressed at their highest level in the testes, yet their loss causes defects during male development. Our finding that these two genes are both testes biased and rapidly evolving is consistent with previous work in Drosophila*.* Studies of male-specific genes and traits have focused on the evolution and role of sperm and seminal proteins, mating behavior, and genital morphology. As a whole, this work has suggested that the genes responsible for these traits may evolve quickly within and among species due to sexual selection and sexual conflict. There is, however, little evidence from these earlier studies that rapidly evolving male-biased genes are essential for male viability.

How can we explain our observation that the knockdown of testes-biased genes causes defects during development? While nearly 20% of annotated genes show male-biased expression (Graveley et al. 2011), genes expressed in male germline stem cells prior to meiosis are typically expressed in at least one other cell type (White-Cooper and Bausek 2010). Therefore, elevated expression in the testes may not always indicate a gene's primary function is testes specific. Rather, genes may be expressed at a high level due to general transcriptional "permissiveness" in the testes (Kleene 2001, 2005). Kaessmann (2010) has proposed that the testes are something of an "evolutionary playground," where novel genes may become expressed for the first time, and later co-opted to function in other tissues. The fact that we could detect some expression in other tissues suggests this model may explain the evolution of *jb* and *karr*.

We next must explain what forces could have led to the extremely rapid sequence evolution of genes that strongly affect male fitness. Most essential genes evolve slowly under purifying selection. The extensive protein-coding divergence of *jb* indicates that purifying selection was not the primary evolutionary force acting across these species. Surprisingly, we were unable to reject simple neutral sequence evolution of *jb* using standard tests of molecular evolution. Natural selection may still be playing a role in *jb* evolution - levels of polymorphism are strikingly low in spite of an overall rate of divergence far above background levels. This pattern is suggestive of recurrent selective sweeps altering the amino acid sequence and stripping polymorphism from this biologically important gene despite our failure to statistically reject the null hypothesis of neutrality. Our work compliments recent studies showing that new genes can strongly affect fitness (Chen et al. 2010; Ding et al. 2010). So far, however, no complete molecular explanation has been found for how or why such genes have become essential.

We speculate that *jb* and *karr* have changed rapidly in response to some extrinsic or intrinsic factor - perhaps these genes recently integrated into an essential pathway or are functionally or structurally linked to a rapidly coevolving gene. A survey of the online interactions database (Murali et al. 2011) shows both genes have a number of potential interactors, but none of these show a rate of sequence evolution comparable to *jb* or *karr.* This observation suggests that an intrinsic interaction with a rapidly evolving gene is not driving the evolution of either gene (although much more data is needed to truly test this hypothesis). Perhaps *jb* or *karr* have instead evolved extensively in order to interact with previously existing pathways as has been speculated for newly evolved genes (Ding et al. 2010).

This pair of exceptionally fast evolving genes highlights a challenge facing the study of genes that are lineage-specific in Drosophila and other species (Levine et al. 2006; Cai et al. 2008; Knowles and McLysaght 2009; Toll-Riera et al. 2009; Chen et al. 2010; Li et al. 2010a). It is difficult to distinguish whether lineage-specificity is due to multiple losses, rapid sequence evolution, or true *de novo* evolution. Genes that appear to be entirely "new" may simply be so diverged that sequence similarity is difficult to detect. In fact, *karr* was first identified as a *de novo* gene (Levine et al. 2006), based on the fact that it could not be found within the colinear region in *D. yakuba* or *D. erecta.* We found *D. yakuba* and *D. erecta* genes with weak homology to *karr*, that share its expression pattern but reside at another genomic locus - apparently having translocated in the *D. melanogaster* lineage after the split of the *D. yakuba/D. melanogaster* ancestor. If genes can evolve at such a rate that they cannot be identified between closely related species, we must be cautious in interpreting a simple lack of sequence similarity as true lineage specificity.

Sequence conservation is often used as a hallmark of functional conservation and an indicator of evolutionary importance. While this trend often holds genome-wide, the exceptions to this pattern - such as *jb* and *karr* - provide a window into how evolutionary novelty becomes incorporated into the essential biological processes of an organism. Our work is the converse of functional studies in mice showing that ultraconserved sequences are apparently *not* essential (Ahituv et al. 2007). Similarly, Chen et al (2010) recently found that dozens of young genes have become essential in the last 10 million years. The next critical question to answer is why these rapidly evolving essential genes exist, why they evolve so quickly, and how these genes retain their essential function in the face of this exceptional rate of molecular evolution.

CHAPTER FOUR: *DE NOVO* GENES IN DROSOPHILA ARE OFTEN ESSENTIAL AND EVOLVE TO BECOME MORE COMPLEX OVER TIME

Authors: Josephine A Reinhardt, David J Begun and Corbin D Jones

ABSTRACT

We carefully analyzed the gene structure and transcriptional evolution of four *de novo* genes in the *D. melanogaster* subgenus. We found that when they could be identified, related sequences were typically transcribed, and open reading frames - though sometimes highly diverged - were often present even when no gene was annotated. Further, the structural complexity of *de novo* genes appears to increase over evolutionary time, with the "younger" *de novo* genes (*CG31909* and *CG33235*) having the least complex structure. We found evidence that RNAi of three of these genes (*CG33235, CG31406,* and *CG34434*) caused lethality prior to eclosion. RNAi with a second RNAi library caused semi-lethality and fertility defects in one gene (*CG34434*). *De novo* genes in *D. melanogaste*r are apparently sometimes essential, despite the fact that they are not shared with other similar species.

INTRODUCTION

The gene complement of every organism is unique - some of this variation is due to genes that arose very recently. While most of these new genes arise after duplication from existing genes (Ohno et al. 1968; Ohno 1970), genes may also evolve from previously non-coding sequences, making them entirely lineage-specific (so-called *de novo* evolution, Long et al. 2003). *De novo* genes were once thought to be vanishingly rare, but recent work (Zhou et al. 2008; Toll-Riera et al. 2009; Tautz and Domazet-Loso 2011) suggests that these brand-new genes make up a significant proportion of recently evolved genes (up to 11%).

When they arise, *de novo* genes tend to be short, simply structured genes (Levine et al. 2006) presumably due to the fact that longer open reading frames (ORFs) would be unlikely to evolve *de novo*. Considering that most genes have multiple exons, one hypothesis is that *de novo* genes start with simple structures and would over time to be more similar to a "typical" gene. However, due to their extremely young age, it is difficult to determine the exact steps in the evolutionary history of *de novo* genes.

Likewise, one might expect that most *de novo* genes initially have minimal functional importance to the organism but gradually integrate themselves into the molecular pathways of their host organism. Thus there is no *a priori* reason to expect a single function to be shared by all *de novo* genes. While the functions of *de novo* genes are mostly unknown, *de novo* genes in Drosophila share testes-biased expression (Levine et al. 2006; Begun et al. 2007b; Zhou et al. 2008) and a *de novo* gene in mouse testes was found to affect male fertility when knocked out (Heinen et al. 2009). This and other work led to the suggestion that the testes might be a place where new genes arise often due to general transcriptional permissiveness (Kaessmann 2010). Meanwhile, an RNAi screen of a number of novel genes showed that one *de novo* gene (*CG31406*) is essential: its loss leads to developmental arrest during the late pupal stage (Chen et al. 2010).

Here, we investigate the evolutionary history of four previously published *de novo* genes (Levine et al. 2006; Zhou et al. 2008). We find that these *de novo* genes represent a variety of evolutionary "stages." Some *de novo* genes appear to be genuinely specific only to *D. melanogaster, D. simulans,* and *D. sechellia*. Others have a deeper evolutionary history, with evidence of transcription in. *D. yakuba* and *D. erecta* - and in one case even *D. annanassae*. Thus this suite of genes appears to capture the evolutionary progression of new genes from their initial formation through the acquisition of increasing structural and functional complexity. We confirm that the *D. melanogaster* expression pattern (testes-biased expression) is conserved across all species tested, supporting the hypothesis that permissive transcription in the testes might contribute to the origin of *de novo* genes. Finally, we show that in *D. melanogaster*, strong RNAi of the three oldest genes studied led to developmental arrest during late pupation, and RNAi of one of these genes (*CG34434*) gene with another (presumably weaker) RNAi line led to male fertility defects and weakened survival. This implies that despite their dissimilarity to other known proteins, *de novo* genes have quickly evolved to be involved in the most basic functions of life - survival and reproduction.

MATERIAL AND METHODS

Molecular evolutionary annotation

Using data from Levine et al (2006) and Zhou et al (2008), we chose a number of published *de novo* genes to further characterize. These genes have no significant hits by BLAST (e = 10^-6) to genes outside of *D. yakuba/D. erecta*. We also mined the NCBI trace archive to rule out the possibility that assembly error led to the misannotation of these genes as *de novo* and found no evidence these genes existed among the traces in species outside of what was previously reported. We searched UCSC's whole genome chained BLASTZ alignments, which are more sensitive to highly diverged hits than BLAST or BLAT (Chiaromonte et al. 2002) in order to find colinear genomic regions. We then used the UCSC and Flybase genome browsers to ask whether the *D. annanassae, D. yakuba, D. erecta, D. simulans,* and *D. sechellia* chained alignments covered annotated genes in whole or in part, despite not matching by BLAST/BLAT. Genes that were found to be colinear to annotated genes with similar structure in all five species were excluded as putative homologous rapidly evolving loci and reported previously (Chapter three). In cases where gene structures were radically different, but there was overlap with an annotated gene, we used RT-PCR to verify (or exclude) the annotated gene models (see below). In the case of *CG34434*, we found that the annotation of the putative *D. yakuba* ortholog incorrectly connected the putative ortholog of *CG34434* with a neighboring gene, and that the *D. simulans* gene had a second, unannotated exon similar to the second exon of the *D. sechellia* ortholog. Finally, the *D. sechellia* ortholog had an incorrect splicing pattern leading to a frame shifted second exon. All other gene models were found to be as-annotated across the five species tested.

Molecular evolutionary and population genetic analyses

We aligned the gene region (5-10kb surrounding the gene) of each candidate to other species using Flybase BLAST and extracted colinear gene regions for six species most closely related to *D. melanogaster* (*D. simulans, D. sechellia, D. yakuba, D. erecta,* and *D. annanassae*) from UCSC BLASTZ alignments. We used these alignments to determine the extent of gene model evolution in the transcribed region of the gene. For each block of aligned sequence (5' UTR, each coding exon, 3'UTR), we aligned the *D. simulans, D. yakuba,* and *D. annanassae* sequences to the *D. melanogaster* sequence using the pairwise alignment algorithm water (Rice et al. 2000). We calculated 1) the proportion of the *D. melanogaster* sequence that was aligned 2) the sequence similarity of aligned bases and 3) the difference in length between species for that block.

We also aligned the colinear extended gene regions (5-15kb surrounding the gene) to the colinear extended gene regions in *D. simulans* using the progressiveMAUVE (Darling et al. 2004; Darling et al. 2010) multiple alignment assuming colinearity (progressiveMauve --collinear --seed-family --disable-backbone). We then aligned 39 *Drosophila melanogaster* Raleigh genomes and 6-9 African genomes (www.dpgp.org) to these regions and calculated polymorphism (π) and divergence (K) in each window. We also calculated Tajima's *D* (Tajima 1989) and Fu and Li's *D* and *F* (Fu and Li 1993) for 500 base pair windows across the region using Variscan (Hutter et al. 2006).

Finally, we aligned each gene with its *D. simulans* ortholog (or in the case of *CG33235*, its inferred ortholog), to determine the number of synonymous and nonsynonymous substitutions and we used the DPGP polymorphism data described above to determine the number of synonymous and nonsynonymous polymorphisms. We calculated the Neutrality Index (Rand and Kann 1996) and the direction of selection (Stoletzki and Eyre-Walker 2011), and performed the Macdonald-Kreitman test (McDonald and Kreitman 1991).

Tissue collection and dissection

Male reproductive tracts were dissected on ice from whole flies (*D. yakuba, D. simulans, and D. melanogaster*) in PBS. Male reproductive tracts and carcasses were each pooled separately and then flash frozen in liquid nitrogen. Whole females and males of each species were collected and flash-frozen. *D. melanogaster* and *D. yakuba* male reproductive tracts were further dissected into accessory glands and testes in PBS and flash frozen. *D. melanogaster* third instar larvae were sexed by identification of genital discs following *Drosophila protocols* (Blair 2000), then flash-frozen. Testes were also dissected from males carrying a null mutation at the gene *tombola* (*tomb*^{GS12862}, Jiang et al. 2007, stock generously supplied by Dr. Helen White-Cooper), and sons of females mutant for the *tudor* gene, which lack a germline (Bloomington stock #1786, Boswell and Mahowald 1985).

Gene expression analyses

We extracted RNA from at two biological replicates of each dissected tissue using TRIZOL reagent (Invitrogen, Grand Island, NY #15596-026), and made cDNA using M-MLV reverse transcriptase (Invitrogen, Grand Island, NY #28025013) and oligo dT primer. We performed relative qRT-PCR quantification using gene-specific primers and a control primer that worked across all species (*Actin5c*). All qRT-PCR was averaged across two technical replicates. 5’ and 3’ RACE were performed following manufacturer’s instructions on *D. melanogaster, D. yakuba,* and *D. simulans* testes RNA using the FirstChoice RLM-RACE kit from Ambion (Grand Island, NY #AM1700) and nested gene-specific primers.

Additional RNAseq and expression data were mined from online databases - FlyAtlas (Chintapalli et al. 2007), modENCODE RNAseq data (Graveley et al. 2011), Baylor RNAseq data (Daines et al. 2011), and FlyTED: Testes expression database (Zhao et al. 2010).

RNAi knockdown

Virgin *Actin*-GAL4 females (Bloomington #4414) were collected and crossed to lines carrying UAS-RNAi constructs for *CG33235, CG31909, CG31406, CG34434* and *Gr22c* - a control (stocks used: 19355, 23550, 39194, 41772, 102263, 104704, 105072, and 110307). Each of these genes except *CG31909* had at least one of the original P-element (GD) and newer *phiC31* (KK) RNAi lines available (www.VDRC.org, Dietzl et al. 2007). The KK lines inserted into a known genomic location and therefore may have more consistent knockdown effects. *CyO* (control) and straight winged (RNAi) progeny of both sexes were counted and collected. We confirmed RNAi knockdown using the same qRT-PCR methods as described. In the case of lines that caused complete lethality prior to the adult stage, we collected larvae in the wandering stage and compared expression of the target gene.

Viability assays

To estimate effects on adult viability, we counted the number of control (*CyO*) and RNAi (straight-winged) progeny eclosing from each RNAi cross (described above). To determine the stage at which lethality was occurring, we crossed the same RNAi lines to a stock with the same *Actin*-GAL4 and *CD8*::UAS-GFP on the same chromosome (kindly donated by S. Chen). RNAi or control status can be ascertained at any stage (RNAi larvae/pupae/adults will express GFP). We collected larvae from the cross during the late third instar/prepupal stage, and sorted by GFP expression and sex (Blair 2000). We then allowed each type to continue development and counted the number that survived, that died prior to pupation, or that died prior to eclosion.

Fertility assays

We used a sperm exhaustion assay to estimate the effect of RNAi knockdown of on male fertility. In this assay (modified from Sun et al. 2004), single males are challenged with two virgin females per day across a five-day period. Males with defects in sperm production should produce fewer offspring per female over the assay period as their sperm stores become depleted. We used a linear model with R (v.2.13.1, R Foundation for Statistical Computing, Vienna Austria) and the lme4 package (Bates and Maechler 2009) to determine if there were significant effects of genotype (indicating a general fertility defect), or a genotype by day interaction effect (indicating a defect in sperm production). For one RNAi line (*CG31406* GD, VDRC#39194), we used single matings instead of sperm exhaustion - single males were paired with single females in individual vials for 24 hours. Then the males were removed and females allowed to oviposit for five days. Offspring were counted and compared between control and RNAi males.

RESULTS AND DISCUSSION

Lineage specific genes in *D. melanogaster* share testes-biased expression

The four genes considered have all been reported previously as *de novo* evolved genes, in some cases by multiple groups (Levine et al. 2006; Zhou et al. 2008; Chen et al. 2010). The CDS of these genes cannot be found using simple BLAST prior to *D. anannasae*, consistent with earlier results. Previously (Levine et al. 2006), expression of *CG31909* and *CG31406* was reported to be testes-specific. RNA-seq data from MODencode (Tweedie et al. 2009; Daines et al. 2011) shows all four genes under consideration share peak expression from late 2nd to 3rd instar larvae and in male adults of all ages. *CG34434* and *CG31406* show additional but much weaker expression in earlier stages (embryonic and L1). In adults, males of all ages show peak expression, while expression is absent in females. FlyATLAS (Chintapalli et al. 2007) confirms the characteristic (Blair 2000; Levine et al. 2006) pattern of elevated expression in the testes of all four genes. In the case of all genes except *CG33235*, expression was also high in the larval fat body.

To test whether expression was specific to males in earlier stages, we sorted larvae by sex (Blair 2000) and used RT-PCR to compare expression between males and females. Expression of all four genes was higher in male than female larvae (Figure 4.1, pink and light blue) implying the male bias in expression is stable through development. Male-biased expression may be due to differential expression in tissues shared by males and females (e.g. the brain) or to expression in male specific tissues (e.g. testes, accessory glands). We compared expression (Figure 4.1) in females (yellow), testes (blue), male accessory glands (accessory gland, mauve), and the carcasses of dissected males (carcass, green). We found that all four genes were expressed at a higher level in the testes than in other tissues. Additionally, we found that expression of these genes was germline-dependent, as expression was reduced in *sons-of-tudor* males, which lack a germline (Boswell and Mahowald 1985). Following on this result, we determined if these genes were involved in meiosis by measuring expression in a *tombola* mutant background. *Tombola (tomb)* is a transcription factor known to be responsible for activating expression of genes important during meiosis in Drosophila (Jiang et al. 2007). We found that expression of *CG31406* was reduced in the *tomb* mutant background (red), implying expression of this gene is partially dependent on an intact meiotic arrest pathway (Figure 4.1).

These four genes represent discrete stages of the evolution of *de novo* genes.

These four genes have all been reported previously as being specific to *D. melanogaster* species subgroup (shared only with *D. melanogaster*), but we find upon closer analysis that fragments of three of the four genes are present in *D. yakuba* or *D. erecta*, and therefore at least partially existed prior to the split of *D. simulans* from *D. melanogaster*. However, none of the genes existed prior to *D. annanassae*. Among these three genes we can observe the gradual acquisition of *de novo* exons and other transcribed regions across the phylogeny (Figure 4.2), supporting the hypothesis that *de novo* genes first arise with simple structures and evolve to become more complex over time. Below, we describe the patterns of molecular and structural evolution within the subgroup for each gene using current annotations (Figure 4.2). We also measure expression of each gene across the five species of the *D. melanogaster* subgroup to determine when transcription began - marking the origin of these transcripts. When

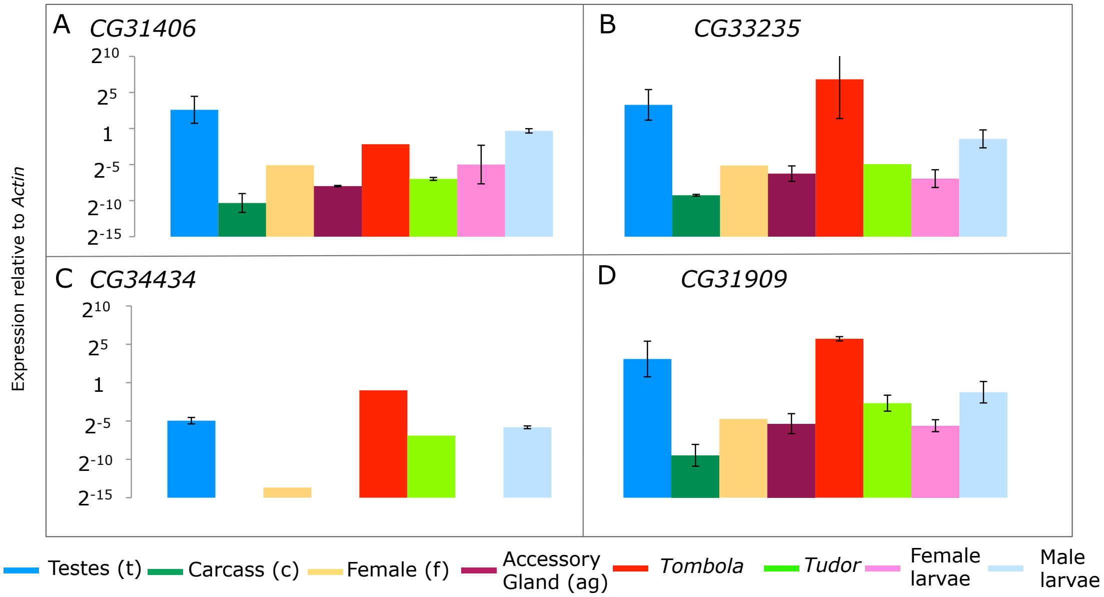


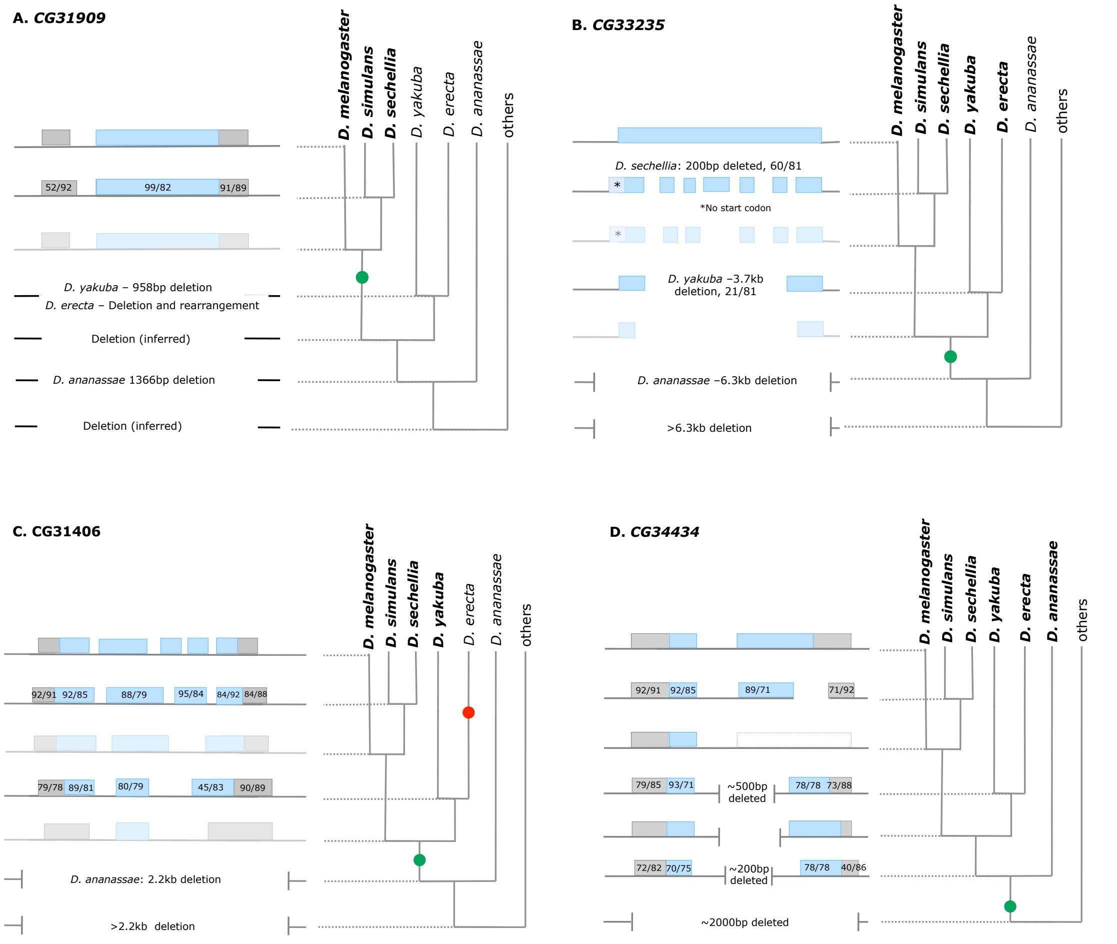
Figure 4.1 - *De novo* genes exhibit male-biased and germline-dependent expression

We compared the expression of four *D. melanogaster de novo* genes (A. *CG31406*, B. *CG33235*, C. *CG34434*, D. *CG31909*) in a variety of tissues dissected from *D. melanogaster* using qRT-PCR. Expression is relative to the reference gene *Actin*. Expression of all four genes was highest in the testes, and was reduced in testes of males lacking a gremline (*Tudor*, light green). In the case of *CG31406* (A), expression was also reduced in flies carrying a meiotic arrest mutation (*Tombola*, red), suggesting it may be functioning in the post-meiotic germline. Finally, we found that male larvae express all four genes at a higher level than females (pink versus light blue).

expressed, we also determined if whether the expression pattern was conserved or had changed since the origin of the gene (Figure 4.3).

CG31909 This gene is found within an intron of *wnt4*, an essential gene in *D. melanogaster*. *CG31909* has one near-identical putative paralog in *D. melanogaster*, and has annotated orthologs in both *D. simulans* and *D. sechellia*. These orthologs are expressed in a similar pattern in these species (Figure 4.3), and are similar at the sequence level. However, the entire CDS is clearly absent in *D. yakuba* and *D. erecta* (Figure 4.3). Multiple primer pairs designed to nearby sequences in *D. yakuba* and *D. erecta* show the sequence is as annotated. There are no matches to raw sequencing reads from *D. yakuba* or *D. erecta,* implying misassembly is not the reason for absence. RT-PCR primers designed to flanking sequences in *D. yakuba* showed no expression of the gene regions (data not shown). The 5' UTR of *CG31909* includes a sequence that is found in the 5'UTR of several other genes (*CG33664, CG33665, CG33666, CG33667, CG33668, CG33669*). These genes all have a similar expression pattern to *CG31909* and its paralog (Daines et al. 2011), but show no protein-coding similarity to *CG31909*. We hypothesize that *CG31909* originated due to movement of this small element into the *wnt4* intron, and subsequent induction of transcription of the *CG31909* gene region.

CG33235 This gene is not colinear to any annotated genes. Unannotated, colinear and weakly similar sequences are present in *D. yakuba/D. erecta.* These regions are transcribed in *D. yakuba, D. erecta, D. simulans* and *D. sechellia* and lack in-frame stop



* D. erecta has a very short match to the 3' end, not expressed

Figure 4.2 - Stepwise gene model evolution of four *D. melanogaster* lineage specific genes
We used BLASTZ alignments as well as our own MAUVE alignments to infer the evolution of four *D. melanogaster de novo* genes. The current *D. melanogaster* gene model is shown on top, and blocks of sequence that are colinear and alignable to parts of the *D. melanogaster* gene (by BLASTZ) are shown below. Blue blocks represent potential protein coding sequence, grey blocks non-coding sequence. *D. simulans, D. yakuba,* and *D. ananassae* colinear blocks are shown as appropriate, with the size of the block indicating the relative size of each block. The proportion of *D. melanogaster* bases aligned and the sequence similarity of aligned bases are shown on each block (proportion/similarity). Large scale deletions are shown using vertical lines. Inferred status of the gene model at the nodes are also shown as faded blocks. Finally, expression was measured (using RT-PCR) in each species where colinear sequence could be found. Species where expression was detected are bolded on the phylogeny and the green dot indicates the inferred start of transcription. A red dot indicates cases where transcription - or the gene itself - was later lost.

codons (Figure: 4.3B, 4.4B), but also lack a start codon. *CG33235* has apparently expanded repeatedly within the *D. melanogaster* species subgroup as the *D. yakuba* colinear region covers only 20% of the total length of the *D. melanogaster* protein, and *D. simulans* covers only 60% (Figure 4.3B). We confirmed that most of the gene is truly absent (not simply misassembled) in *D. yakuba and D. erecta* by aligning *CG33235* to raw sequencing traces from these species (www.ncbi.nlm.nih.gov/Traces). *CG33235* appears to have arisen as a small, transcribed region in the ancestor of *D. yakuba* and *D. melanogaster*, and then underwent a rapid series of repeat expansions, ultimately evolving into a gene consisting of a single, long (4.7kb) open reading frame.

CG31406 Like *CG33235*, *CG31406* evolved through stepwise acquisition of increasing complexity along the lineage leading to *D. melanogaster*. However, it appears that this gene evolved through acquisition of novel exons not present in the original gene, not expansion of an existing repeat region as seen with *CG33235*. *CG31406* is structurally similar in *D. melanogaster, D. simulans,* and *D. sechellia,* with *D. sechellia* and *D. simulans* lacking one exon present in *D. melanogaster* (Figure 4.2C). There is some homology to a *D. yakuba* gene, but the structure of the *D. yakuba* gene (based on EST data) is substantially different from the *D. melanogaster* copy. The *D. yakuba* gene has only three exons whereas the *D. melanogaster* gene has five and the D. yakuba annotated protein is about half the total length (Figure 4.4C). No gene is annotated in *D. erecta,* and we confirmed that the small section of *D. erecta* colinear sequence was not expressed (Figure 4.2, Figure 4.3). Current annotations of the gene region showing that most of the gene (everything after the first exon) is not present in *D. erecta*, were

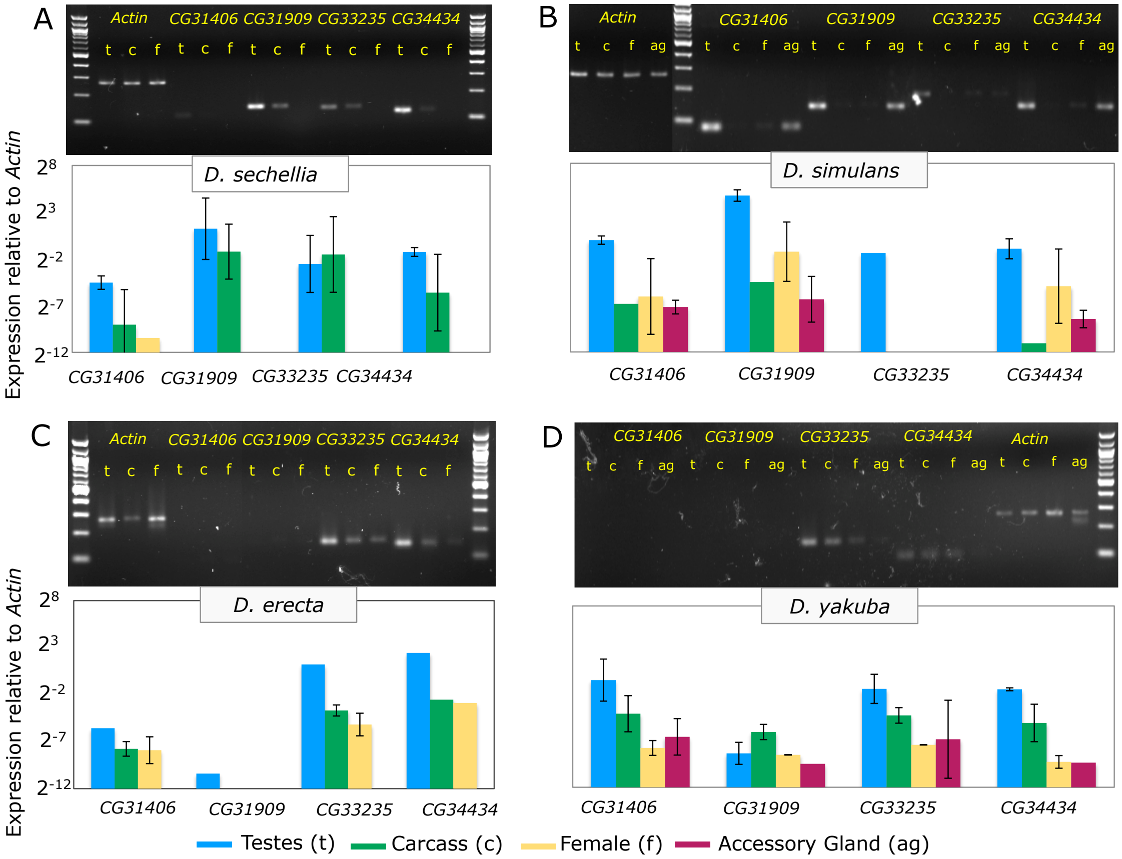


Figure 4.3 - Testes biased expression of *de novo* genes is conserved across species

We compared the expression of sequences or genes that were colinear to *D. melanogaster de novo* genes across a number of tissues in the five species of the *melanogaster* subgroup. In *D sechellia* and *D. erecta,* we dissected male reproductive tracts from flies, and compared expression across the reproductive tracts (testes, blue), the remainder of the male (carcass, green), and whole females (Females, yellow). In *D. yakuba* and *D. simulans,* we further dissected male reproductive tracts into testes and accessory glands (purple). When available, two biological replicates are shown. Expression is relative to *Actin* (based on deltaCt measurements).

confirmed with direct sequencing (data not shown). Thus it is likely that the gene was lost in *D. erecta* after first originating in the common ancestor of *D. yakuba and D. melanogaster*.

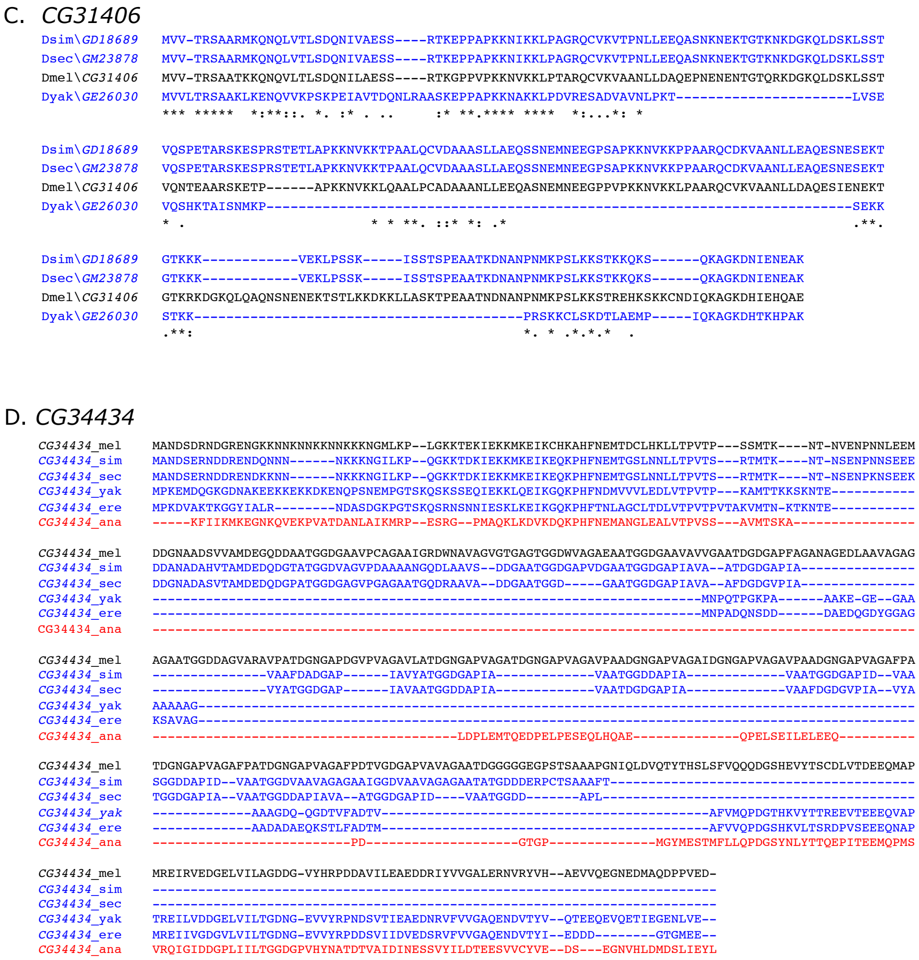
CG34434 The coding sequence of *CG34434* partly overlaps with genes in *D. simulans, D. sechellia, D. yakuba,* and *D. erecta*. Small parts of the first and second exons are identifiably similar to sequences annotated as non-coding in *D. annanassae*. Each of the annotated genes is substantially different in structure (Figure 4.2D, Figure 4.4D). Partially, this is due to a misannotation of the *D. sechellia, D. simulans*, and *D. yakuba* genes. The *D. sechellia* gene (*GM12640*) was misannotated with a mutation that caused the second exon to be frameshifted relative to *D. melanogaster*. We found using RT-PCR (and for simulans, testes RNA-seq data (Artieri et al. 2011) that the *D. yakuba* gene is similar in structure to the *D. erecta gene,* and the *D. simulans* gene (*GD16225*) is similar in structure to the *D. sechellia* gene. After this reannotation, it became clear that the first exon of *CG34434* is present in these 5 species, and is also expressed in *D. annanassae*. However, the *D. simulans/D. sechellia* and *D. yakuba/D. erecta* second exons differ substantially from one another. Interestingly, the *D. simulans/D. sechellia* exon 2 overlaps with a different part of *CG34434* exon 2 than exon 2 in *D. yakuba* and *D. erecta* (Figure 4.3D, 4.4D). This suggests that the full length *CG34434* exon 2 is ancestral to all five species, but was only retained in full in *D. melanogaster*. The *D. annanassae* region aligns poorly to *D. melanogaster* exon 2, and no gene is annotated. However, we were able to detect expression of exon 2 in *D. annanassae*, and the alignable regions in *D. annanassae* contained no stop codons (Figure 4.4D). Therefore,

A. *CG31909*

```
Dmel\CG31909   MEPSFSNTNKSELQNQLRVERALQDKYLKEFSSMADELNHLIDGTTPSHIETYEYLDTVDTESGRSLEEEKKLTNFFYKLKSDLQKLYREIVENPKYLNE
Dsec\GM16355   MEPSLSNTNKSELLDELRAERALQDKYLKDFCSMADKLRHLVDGTVPSKIETYEYLESSDGDSNRSLEEEKKLTIFFYKMRCDLEKNLREVQENLKYLDE
Dsim\GD23456   MEPSLSNTNKSELLDELRAERALQDKYLKDFCSMTDKLRHLVYGTVPTKIETYEYLESSDGDSNRSLEEEKKLTIFFYKMRCDLEKNLRQIQENPKYLDE
               ****:******** ::**.**********:*.**:*:*.**: **.*::*******:: * :*.********** ****::.**:*  *:: ** ***:*


Dmel\CG31909   IRDSTIIRDNIDLLDEVDAFEGSMNML
Dsec\GM16355   VRELGVIWDNIQYIW------------
Dsim\GD23456   VRQFGVIWDNID---------------
               :*:  :* ***:
```

B. *CG33235*

```
dmel_CG33235   MNFSKKLLGDKNLVRLDSDATRTFCPQVSEGRLTLENVIFDRGGSVTTKKPRKRPTPKPVIRNFNNLSVPEEVDVDNLQGQRKGVAVNSNNDEMMISSSD
dsim_CG33235   TNFSMNVLGDKN------------------------RVISDRGGSVTTKKHGKRPAPYPMSRNSNNWGVLEEVDVENLLG--------------------
dsec_CG33235   TNFSMKVLGDKN------------------------RVISDRGGSVTTQKQGKRPAPYPMSRNSNNWGELEEVDVENLLG--------------------
dyak_CG33235   ----------------------------------------------------------------------------------------------------
dere_CG33235   ----------------------------------------------------------------------------------------------------

dmel_CG33235   SSDSSDDYSSFGDDIFTPGPETSDTSDGDSSCEDELKIPDFKSSATSKDEKLIPSSKWNFTLTKDIIPPGEGKKSHIGASLPEPVNRNFNNKSVLDLQGQ
dsim_CG33235   ---------------------------------------------------------------------------------------------------Q
dsec_CG33235   ---------------------------------------------------------------------------------------------------Q
dyak_CG33235   ----------------------------------------------------------------------------------------------------
dere_CG33235   ----------------------------------------------------------------------------------------------------

dmel_CG33235   RNQGSRFGAQGVAVNSNKDEMMISSSDSSDSSDDYSSFGDDIFTPGPETSDTSDGDSSCEDELKIPDFKSSATSKDKKMIPSSKRNFTLTKDIIPPGEGK
dsim_CG33235   RHQGSTFGAQDFAVNSNEDGKMIPGKDISDT---------------------------------------------------------------------
dsec_CG33235   RYQGSTFGAQDFAVNSKEDGKMIPGSDSSDTKNDDS----------PSIPSTLEN--------VISDRGGGVTTKT-------------------PGE--
dyak_CG33235   ----------------------------------------------------------------------------------------------------
dere_CG33235   ----------------------------------------------------------------------------------------------------

dmel_CG33235   KSHIGASLPEPVNRNSNNKSVLDLLGQRNQGVNSNKDELTILSSDTSDTSKEEKMIPSSKRNFFLTKDIIPPGEGKKSHIGASLPEPVNRNSNSKSVLDL
dsim_CG33235   ----------------------------------------------------------------------------------------------------
dsec_CG33235   -----RPTPDPMSRNSNNSSVLEEVDVDN----------------------------------------------------------------------L
dyak_CG33235   ----------------------------------------------------------------------------------------------------
dere_CG33235   ----------------------------------------------------------------------------------------------------

dmel_CG33235   LGQRNQGVNSNKDELTILSSDTSDTSKEEKMIPSSKRNFFLTKDNIPPGEGKKSHIGASLPEPVDRTSNNKSVLDLQGQRNQGSKDEKMIPSSKRNFTLT
dsim_CG33235   ----------------------------------------------------------------------------------------------------
dsec_CG33235   LGQRHQGSTFGAQDFAVNSKE------DGKMIPGSDSSDSMDDDSLS----IISILEVVISDREDDVTTKK-----------------------------
dyak_CG33235   ----------------------------------------------------------------------------------------------------
dere_CG33235   ----------------------------------------------------------------------------------------------------

dmel_CG33235   KDIIPPGEGKKSHIGASLPEPLNRNSNNKSVLDLQGQRDQGSKDEKMIPSSKRNFTLTKDIIPPGEGKKSHIGASLPEPVNRNSNNKSVLDLQGQRDQGS
dsim_CG33235   ----------------------------------------------------------------------------------------------------
dsec_CG33235   -----PEE-------RPTPDPMSRNSNNSSVLEEVDVDN----------------------------------------------------LLGQRHQG-
dyak_CG33235   ----------------------------------------------------------------------------------------------------
dere_CG33235   ----------------------------------------------------------------------------------------------------

dmel_CG33235   KDEKMIPSSKRNFTLTKDIVPPGEGKKSHIGASLPEPVNRNSNNKSVLDLQGQRNQGSKDEKMIPSSKRNFTLTKDIIPPGEGKKSHIGASLPEPVNRNS
dsim_CG33235   ----------------------------------------------------------------------------------------------------
dsec_CG33235   ---------------------------------------------STFGAQDFAVNSKEDGKMIPGSDSSDSMDDDSLS---------------------
dyak_CG33235   ----------------------------------------------------------------------------------------------------
dere_CG33235   ----------------------------------------------------------------------------------------------------

dmel_CG33235   NNKSVLDLQGQRNQGSKDEKMIPSSKRNFTLTKDIVPPVEGKKSHIGASLPEPVNRNSNSKSVLDLLGQRNQGVNSNKDELTILSSDTSDTSKEEKMIPS
dsim_CG33235   ----------------------------------------------------------------------------------------------------
dsec_CG33235   -IISILEV------------VISDREDDVTTKKPEERPT-----------PDPMSRNSNNSSVLEEVDVDN-----------------------------
dyak_CG33235   ----------------------------------------------------------------------------------------------------
dere_CG33235   ----------------------------------------------------------------------------------------------------

dmel_CG33235   SKRNFFLTKDNIPPGEGKKSHIGASLPEPVDRTSNNKSVLDLQGQRNQGSKDEKMIPSSKRNFTLTKDIIPPGEGKKSHIGASLPEPVNRNSNSKSVLDL
dsim_CG33235   ----------------------------------------------------------------------------------------------------
dsec_CG33235   -----------------------------------------LLGQRHQGST-------------------------------------------------
dyak_CG33235   ----------------------------------------------------------------------------------------------------
dere_CG33235   ----------------------------------------------------------------------------------------------------

dmel_CG33235   LGQRNQGVNSNKDELTILSSDTSDTSKEEKMIPSSKRNFFLTKDNIPPGEGKKSHIGASLPEPVNRNSNNKSVLDLQGQRNQGSKDEKMIPSSKRNFTLT
dsim_CG33235   ----------------------------------------------------------------------------------------------------
dsec_CG33235   FGAQDFAVNSKEDGKMIPGSDSSDSMDDD----------------------------------------SLSIISIL---------EVVISDREDDVTTK
dyak_CG33235   ----------------------------------------------------------------------------------------------------
dere_CG33235   ----------------------------------------------------------------------------------------------------

dmel_CG33235   KDIIPPGEGKKSPFGATLPEPVGRNSNNKSVLDLLGQRNQGVNSNKDELTILSSDTSDTSKEEKMIPSSKRNFFLTKDNIPPGEGKKSHIGASLPEPVNR
dsim_CG33235   ----------------------------------------------------------------------------------------------------
dsec_CG33235   K----PGE-------RPTPDPMSRNSNN-SVLEEVDVDN-------------------------------------------------------------
dyak_CG33235   ---------------------------MKFKI-LIDGKN---------IVRVECPGYATQRFCVKVHAGR----ITLENVFPNRGGFTRGGAYITASKPK
dere_CG33235   ---------------------------MKFKIHLINGKN---------IIRVECPGYETQRFAVQVRGGK----ITLLNVLPNQGEGSSGGSSATT----

dmel_CG33235   NSNNKSVLDLQGQRNQGSKDEKMIPSSKRNFTLTKDIIPPGEGKKSPFGATLPEPVGRNSNNKSVLDLLGQRNQGVNSNKDELTILSSDTSDTSKEEKMI
dsim_CG33235   ----------------------------------------------------------------------------------------------------
dsec_CG33235   ---------LLGQRHQGST-------------------------------------------------FGAQDFAVNSNEDGEMIPGSDSSDSMDD----
dyak_CG33235   TVASKKSVKRPATEPVRRRPIQFSLQDGVDVWMDTEMRQPSEGQR-------------------------FGSQGLGGYGDRLFKL--------------
dere_CG33235   AVSSAKGLSRVRIKMIFGKPSVP--------------TAPSSSRQ-------------------------AGEAGLS-----------------------

dmel_CG33235   PSSKRNFFLTKDIIPPGEGKKSHIGASLPEPVNRNSNNKSVLDLQGQRNQGSKDEKMIPSSKRNFTLTKDIIPPGEEKKSHIGASLPEPVNRNSNNKSVL
dsim_CG33235   ----------------------------------------------------------------------------------------------------
dsec_CG33235   ---------------------------------------------GSLSIISILEVVISDREDDVTTKK----PGER-------PTPDPMSRNSNNSSVL
dyak_CG33235   -----------------------------------DVWQPIPASREVAKNAASRRKTLSRKKENFPLRATITKKKQFDESSMPTATSTPHQKRRKTGLLS
dere_CG33235   -----------------------------------SEWGDIPAMGDLGQQQEEEEPPLPGGSE---LGAIIRAHQMSSATLMNELLETQNKLEIVEKKFQ

dmel_CG33235   EEVDVDNLLSQRNQGNRFGAQRLVVNSNKDELIILGSDSSDTSGDANLCKDEIKIPSSKRKLTLMNDMVSPSKRAHTEAVGPTTPKPWAPVLADKKDNDY
dsim_CG33235   ----------------------------------------------------------------------------------------------------
dsec_CG33235   EEVDVDNLLGQRHQGSTFGAQDFTVNSNEDGKMIPGKDSSDTSEDYSSFEDEIMSPGSES-----SDTSEDYSSFEDEIMSPGS---------DSSDTSE
dyak_CG33235   ETTERKPTQKVYSPSLRVDQTIILGKRN----------------ADGLSAKAFSMMTIRPSLVKA----ETKQRSQDLMTSLNDLGKHQVTTSNGQRHSA
dere_CG33235   KKMEK-----------EMEKDEMEKKKM----------------EKQLAKKMLKHKVAKEKMAKK----MAKQK--------------------------

dmel_CG33235   PYGGKEWARKFLEKKKTKAEVANPTLAEVSPQEDESKVVKELLETQEQLDMVEEMERKEMEKEKIDEWRKTSKQVLEKNKKEEKLNLQKPKMAKGKVQKK
dsim_CG33235   ----------------------------------------------------------------------------------------------------
dsec_CG33235   DY------------RSFEDEIMSPGSESSDTSEDYSSFEDEIMSPGSDS----------------------------SDTSEDNRSFEDEIMSPGSESGD
dyak_CG33235   SLKSLEREQKRLTKKKIEKKMAKKKIEKKMA-----------KKIEEMKMANEKKEKREMLRQKMVKEEKERKMMLRQKMVKEKMAKIAKKIAKEKMAKI
dere_CG33235   -----------IAKEKMARKMAKEKMIR--------------LKMAKEMLENQKMKKEKMIKLRMIKLKIEKETIAKENMEKEKMIKMK--MAKEKME--

dmel_CG33235   KSKDRQCGKTLKIAKGKLEALKTAKDTMAAKVLEKQMIMKKKMAQKMSEIEALERHKKYVEIKEKMVMGNGKRNRSAPYRYLKK
dsim_CG33235   ------------------------------------------------------------------------------------
dsec_CG33235   TSEDYXS-----------------------------------------------------FEDEVMSPGSNSSDTSEDYRIFRR
dyak_CG33235   ----------------------AKKMAKEKMAKIAKKMAKEKMAKIAKKMAKE-------------------------------
dere_CG33235   ------------------------KEEMKKLKMAKEKMEKQKMTKQLRGIM---------------------------------
```



C. *CG31406*

```
Dsim\GD18689    MVV-TRSAARMKQNQLVTLSDQNIVAESS----RTKEPPAPKKNIKKLPAGRQCVKVTPNLLEEQASNKNEKTGTKNKDGKQLDSKLSST
Dsec\GM23878    MVV-TRSAARMKQNQLVTLSDQNIVAESS----RTKEPPAPKKNIKKLPAGRQCVKVTPNLLEEQASNKNEKTGTKNKDGKQLDSKLSST
Dmel\CG31406    MVV-TRSAATKKQNQVLTLSDQNILAESS----RTKGPPVPKKNVKKLPTARQCVKVAANLLDAQEPNENENTGTQRKDGKQLDSKLSST
Dyak\GE26030    MVVLTRSAAKLKENQVVKPSKPEIAVTDQNLRAASKEPPAPKKNAKKLPDVRESADVAVNLPKT----------------------LVSE
                *** *****  *:**::. *. :* . ..     :* **.**** ****  *:...*: *

Dsim\GD18689    VQSPETARSKESPRSTETLAPKKNVKKTPAALQCVDAAASLLAEQSSNEMNEEGPSAPKKNVKKPPAARQCDKVAANLLEAQESNESEKT
Dsec\GM23878    VQSPETARSKESPRSTETLAPKKNVKKTPAALQCVDAAASLLAEQSSNEMNEEGPSAPKKNVKKPPAARQCDKVAANLLEAQESNESEKT
Dmel\CG31406    VQNTEAARSKETP------APKKNVKKLQAALPCADAAANLLEEQASNEMNEEGPPVPKKNVKKLPAARQCVKVAANLLDAQESIENEKT
Dyak\GE26030    VQSHKTAISNMKP-------------------------------------------------------------------------SEKK
                * .                       * * **. ::* *: .*                                           .**.

Dsim\GD18689    GTKKK------------VEKLPSSK-----ISSTSPEAATKDNANPNMKPSLKKSTKKQKS------QKAGKDNIENEAK
Dsec\GM23878    GTKKK------------VEKLPSSK-----ISSTSPEAATKDNANPNMKPSLKKSTKKQKS------QKAGKDNIENEAK
Dmel\CG31406    GTKRKDGKQLQAQNSNENEKTSTLKKDKKLLASKTPEAATNDNANPNMKPSLKKSTREHKSKKCNDIQKAGKDHIEHQAE
Dyak\GE26030    STKK-----------------------------------------PRSKKCLSKDTLAEMP-----IQKAGKDHTKHPAK
                .**:                                         *. * .*.*.*  .
```

D. *CG34434*

```
CG34434_mel     MANDSDRNDGRENGKKNNKNNKKNNKKKNGMLKP--LGKKTEKIEKKMKEIKCHKAHFNEMTDCLHKLLTPVTP---SSMTK----NT-NVENPNNLEEM
CG34434_sim     MANDSERNDDRENDQNNN------NKKKNGILKP--QGKKTDKIEKKMKEIKEQKPHFNEMTGSLNNLLTPVTS---RTMTK----NT-NSENPNNSEEE
CG34434_sec     MANDSERNDDRENDKKNN------NKKKNGILKP--QGKKTDKIEKKMKEIKEQKPHFNEMTGSLNNLLTPVTS---RTMTK----NT-NSENPKNSEEK
CG34434_yak     MPKEMDQGKGDNAKEEKKEKKDKENQPSNEMPGTSKQSKSSEQIEKKLQEIKGQKPHFNDMVVVLEDLVTPVTP---KAMTTKKSKNTE-----------
CG34434_ere     MPKDVAKTKGGYIALR--------NDASDGKPGTSKQSRNSNNIESKLKEIKGQKPHFTNLAGCLTDLVTPVTPVTAKVMTN-KTKNTE-----------
CG34434_ana     -----KFIIKMKEGNKQVEKPVATDANLAIKMRP--ESRG--PMAQKLKDVKDQKPHFNEMANGLEALVTPVSS---AVMTSKA----------------

CG34434_mel     DDGNAADSVVAMDEGQDDAATGGDGAAVPCAGAAIGRDWNAVAGVGTGAGTGGDWVAGAEAATGGDGAAVAVVGAATDGDGAPFAGANAGEDLAAVAGAG
CG34434_sim     DDANADAHVTAMDEDQDGTATGGDVAGVPDAAAANGQDLAAVS--DDGAATGGDGAPVDGAATGGDGAPIAVA--ATDGDGAPIA---------------
CG34434_sec     DDGNADASVTAMDEDQDGPATGGDGAGVPGAGAATGQDRAAVA--DDGAATGGD-----GAATGGDGAPIAVA--AFDGDGVPIA---------------
CG34434_yak     ------------------------------------------------------------------------MNPQTPGKPA------AAKE-GE--GAA
CG34434_ere     ------------------------------------------------------------------------MNPADQNSDD------DAEDQGDYGGAG
CG34434_ana     ----------------------------------------------------------------------------------------------------

CG34434_mel     AGAATGGDDAGVARAVPATDGNGAPDGVPVAGAVLATDGNGAPVAGATDGNGAPVAGAVPAADGNGAPVAGAIDGNGAPVAGAVPAADGNGAPVAGAFPA
CG34434_sim     ---------------VAAFDADGAP------IAVYATGGDGAPIA-------------VAATGGDDAPIA-------------VAATGGDGAPID--VAA
CG34434_sec     ---------------VYATGGDGAP------IAVAATGGDDAPIA-------------VAATDGDGAPIA-------------VAAFDGDGVPIA--VYA
CG34434_yak     AAAAAG----------------------------------------------------------------------------------------------
CG34434_ere     KSAVAG----------------------------------------------------------------------------------------------
CG34434_ana     ----------------------------------------LDPLEMTQEDPELPESEQLHQAE--------------QPELSEILELEEQ----------

CG34434_mel     TDGNGAPVAGAFPATDGNGAPVAGAFPDTVGDGAPVAVAGAATDGGGGGEGPSTSAAAPGNIQLDVQTYTHSLSFVQQQDGSHEVYTSCDLVTDEEQMAP
CG34434_sim     SGGDDAPID--VAATGGDVAAVAGAGAAIGGDVAAVAGAGAATATGDDDERPCTSAAAFT----------------------------------------
CG34434_sec     TGGDGAPIA--VAATGGDDAPIAVA--ATGGDGAPID----VAATGGDD-------APL-----------------------------------------
CG34434_yak     -------------AAAGDQ-QGDTVFADTV-------------------------------------------AFVMQPDGTHKVYTTREEVTEEEQVAP
CG34434_ere     -------------AADADAEQKSTLFADTM-------------------------------------------AFVVQPDGSHKVLTSRDPVSEEEQNAP
CG34434_ana     --------------------------PD--------------------GTGP--------------MGYMESTMFLLQPDGSYNLYTTQEPITEEMQPMS

CG34434_mel     MREIRVEDGELVILAGDDG-VYHRPDDAVILEAEDDRIYVVGALERNVRYVH--AEVVQEGNEDMAQDPPVED-
CG34434_sim     --------------------------------------------------------------------------
CG34434_sec     --------------------------------------------------------------------------
CG34434_yak     TREILVDDGELVILTGDNG-EVVYRPNDSVTIEAEDNRVFVVGAQENDVTYV--QTEEQEVQETIEGENLVE--
CG34434_ere     MREIIVGDGVLVILTGDNG-EVVYRPDDSVIIDVEDSRVFVVGAQENDVTYI--EDDD--------GTGMEE--
CG34434_ana     VRQIGIDDGPLIILTGGDGPVHYNATDTVAIDINESSVYILDTEESVVCYVE--DS---EGNVHLDMDSLIEYL
```

Figure 4.4 - Protein alignments illustrate rapid structural and sequence evolution

We aligned each *de novo* protein (A. *CG31909*; B. *CG33235*; C. *CG31406*; D. *CG34434*) to colinear annotated genes when available using ClustalW2 (blue text). When no gene was annotated in a given species, but colinear sequence was present, we extracted the sequences from the BLASTZ alignments and translated them before aligning to the gene of interest (red text).

transcription of the gene - and possibly also translation - most likely arose in the common ancestor of *D. annanassae* and *D. melanogaster*.

RNAi of *D. melanogaster* lineage specific genes affects viability and male fertility

Testes-biased expression led us to hypothesize that these genes may be involved in male fertility. However, RNAi lines from the VDRC (Dietzl et al. 2007) crossed with a ubiquitous GAL4 driver (Actin5c GAL4 stock#4414), produced no offspring for the three genes assayed (*CG31406*, *CG34434*, *CG33235*, all using the KK lines), though the control gene *Gr22c* caused no lethality. This suggests that these genes are important for viability. Using a driver line that included a GFP marker (*Actin*GAL4,UAS:GFP/CyO, donated by S. Chen), we found that lethality occurred in all three cases at the late pupal "pharate" stage (Figure 4.5). We measured the extent of RNAi knockdown in two biological replicates of control and RNAi larvae, and found that RNAi samples had lower expression than control samples, though knockdown of *CG31406* trending but not significant (Figure 4.5C, p = 0.078 for *CG31406* KK, p<0.05 for *CG34434* KK and *CG33235* KK). Our observation of pharate-stage lethality is consistent with previous work showing RNAi of *CG31406* leads to pharate-stage death (Chen et al. 2010) along with 30% of other new genes using the same driver. We crossed the RNAi lines to an additional ubiquitious driver (GAL4 *Tubulin*, Bloomington#5138) as well as a driver that targeted testes and various larval tissues (Larval fat body, gut, leg discs, and salivary glands, Bloomington #6982) with the same result of complete lethality. We obtained GD RNAi lines, which are generally thought to produce weaker knockdown, for all four genes. Using the same design as above, all GD lines produced viable progeny. We were able to confirm partial knockdown (Figure 4.5C) of the target genes in adults from three of the crosses ($p < 0.05$), but *CG31909* did not show knockdown ($p = 0.42$). *CG34434* GD-RNAi showed robust (~40-fold, Figure 4.5C) knockdown and a semi-lethal phenotype in adults, with males more affected than females while *CG31406, CG31909* and *CG33235* GD RNAi had no affect on viability (Figure 4.6A). In addition, *CG34434* GD-RNAi males had a dramatically reduced lifespan (Figure 4.6B).

Using males collected from all three lines that survived RNA knockdown (*CG31406* GD, *CG34434* GD, and *CG33235* GD), we proceeded to compare control and RNAi fertility using a sperm exhaustion assay (Sun et al. 2004) to measure potential effects on male fertility and on sperm production. We found that total fertility was reduced with RNAi of *CG34434* (Figure 4.7, $P < 0.05$). In particular, the number of males that failed to produce offspring was higher than expected. *CG34434* males appeared weaker overall as indicated by their shortened lifespan (Figure 4.6) and hence were unable to mate successfully. *CG31406* and *CG33235* RNAi males performed similarly to control flies (Figure 4.5B, C).

Because we were unable to knock down expression of *CG31909* using RNAi, we produced Tilling lines for *CG31909* (Cooper et al. 2008), obtaining an allele with a premature termination codon (PTC, predicted to truncate 40% of the protein) as well as a number of nonsynonymous and regulatory mutations. The PTC allele did not alter expression, which was not unexpected as nonsense mediated decay (Nagy and Maquat 1998) rarely effects PTCs that occur within ~400bp of the polyA signal (Gatfield et al. 2003). None of the alleles appeared to affect viability. We used sperm exhaustion to determine whether the PTC or a regulatory mutation reduced fertility and saw no effect of

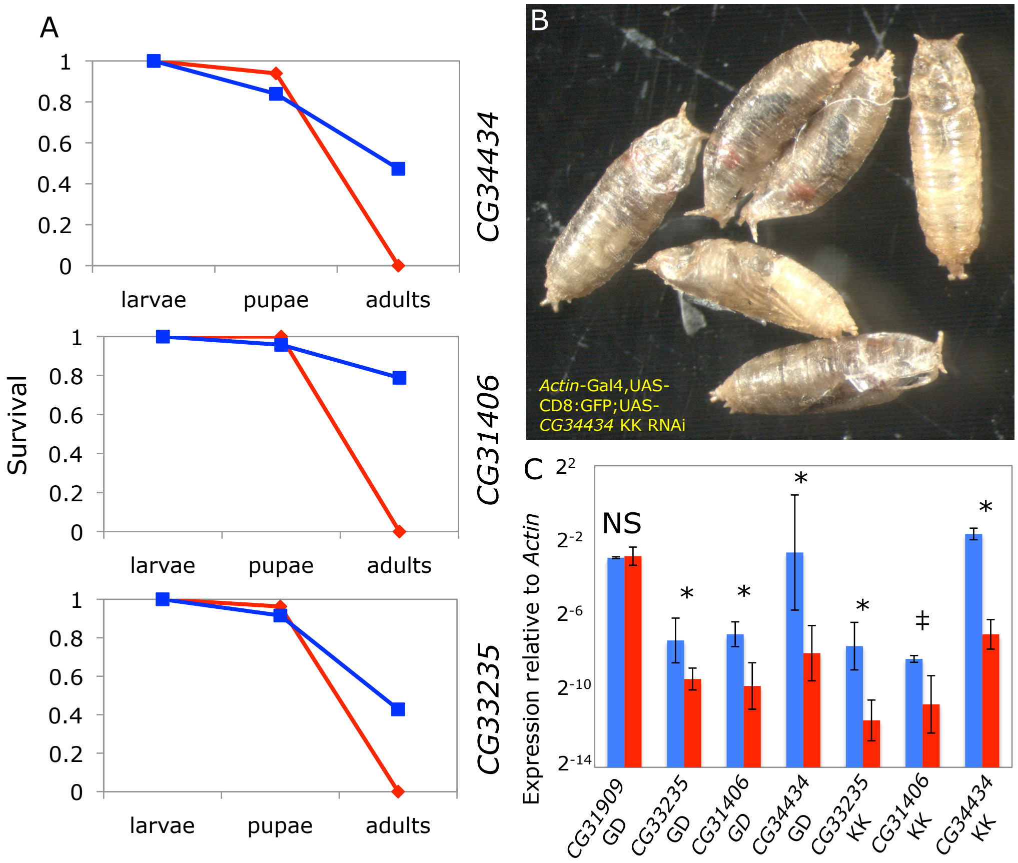


Figure 4.5 - RNAi of three *D. melanogaster de novo* genes causes arrest at the pharate stage

We knocked down expression of three *de novo* genes using KK RNAi lines and found that adult flies of the RNAi genotype did not eclose. (A) Using a GFP marked line, we found that RNAi (red, diamond) and control (blue, square) flies had similar death rates before the adult stage. (B) Flies that died appeared fully developed, but failed to eclose. (C) Extent of RNAi knockdown is compared between RNAi (red) and control (blue) flies for all lines used. Knockdown was measured in adults for the GD lines and in larvae for the KK lines - significance is indicated above each control/RNAi pair (*P<0.05, ‡P<0.10, NS P>0.10).

A

	RNAi males	Control males	RNAi females	Control females	RNAi/control	Fisher's Test P-value
CG31406-GD	77	58	87	77	1.215	0.5595
CG31909-GD	42	35	64	37	1.472	0.28
CG33235-GD	80	82	112	78	1.2	0.0868
CG34434-GD	22	81	46	95	0.386	0.065
CG34434-GD set 2	556	599	193	355	0.474	0.0001

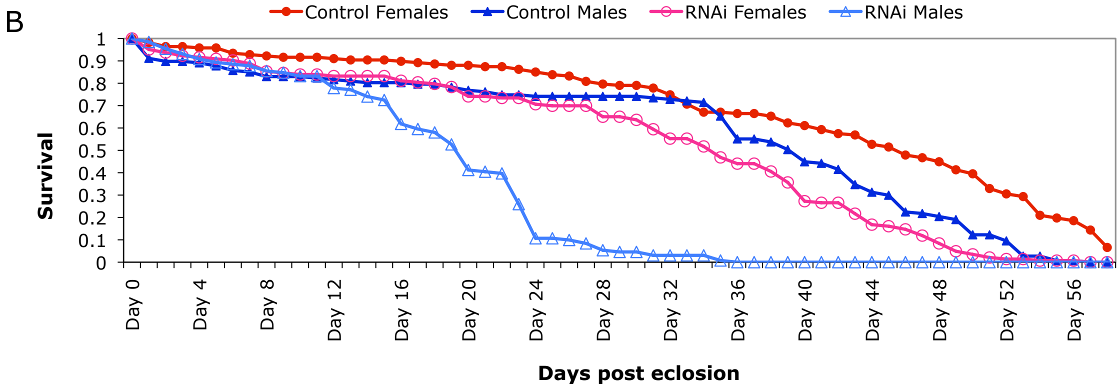


Figure 4.6 - RNAi of *CG34434* using the GD stock is semi-lethal and affects males disproportionately

(A) The GD RNAi lines for four *de novo* genes were crossed to *Actin*-GAL4 (Bloomington #4414) and progeny was counted. *CG34434* RNAi flies emerged at less than half the rate of control flies. A larger experiment (set 2) shows that males are disproportionately affected by RNAi. (B) Flies emerging from the *CG34434*-RNAi cross were sorted and kept in small vial populations (5-10 flies) as they emerged. RNAi males (light blue) died much more quickly than their female RNAi siblings (pink) or either control group (red, blue).

the flyTILL lines on performance compared to controls (a D->N mutation at position 118 And w^{1118}, Figure 4.7C). This could be for a number of reasons. First, *CG31909* has a near duplicate that is also testes-expressed according to MODencode and EST data (BT023668), though it is not annotated as a gene. This duplicate allele's function may be redundant with *CG31909* and sufficient to complement our tilling mutant. Secondly, *CG31909* may not function in fertility as a protein-coding gene. It has been suggested that some small ORFs may actually function as long-non-coding RNAs (Ponting 2008), though there is competing evidence that previously unrecognized small ORFs may be functional (Frith et al. 2006). Furthermore, *CG31909* may not affect fertility in a measurable way in this assay. Given that loss of testes-specific genes sometimes affects performance in a sperm competition assay (Yeh et al. 2012), this would be a reasonable next step.

***De novo* genes are evolving rapidly compared to nearby regions**

Each of these *de novo* genes has undergone large-scale changes both in gene model and sequence evolution across the five species of the *D. melanogaster* subgroup (Figure 4.3, Figure 4.4). We tested for a role of positive selection in the recent evolutionary history of these genes. We aligned the *D. simulans* and *D. melanogaster* extended gene region and compared with polymorphism data from *D. melanogaster* (www.dpgp.org, lines collected from Raleigh, "NA" and Malawi, "AF") using Variscan (Hutter et al. 2006). Divergence divergence (Figure 4.8, *K*, black bars) was always highest over the part of the region including the gene, whereas polymorphism was usually lower or similar to background

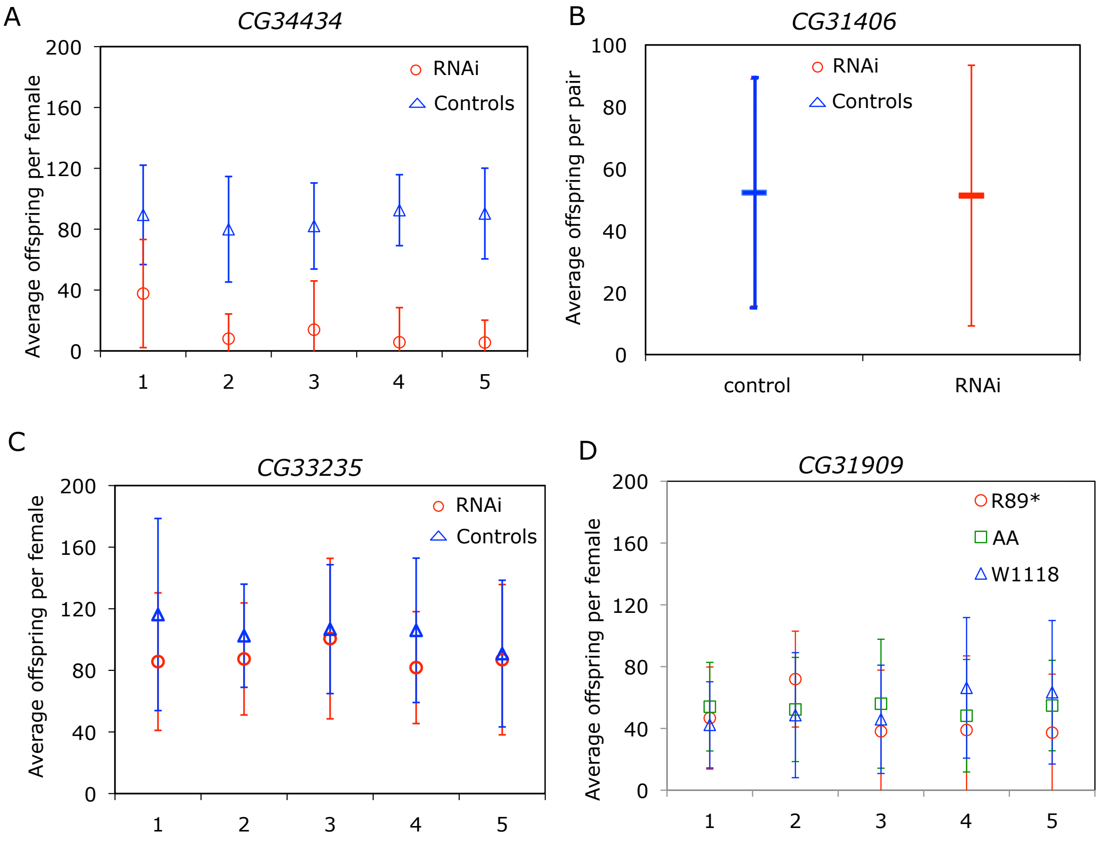


Figure 4.7 - RNAi of *CG34434* causes reduced performance of males in a sperm exhaustion fertility assay
We used the sperm exhaustion assay to measure male reproductive fitness of flies from GD RNAi crosses compared to controls (A,B). Using FlyTILL, we developed a mutant line for *CG31909* because we found the only RNAi line for that gene did not reduce expression (Figure 4.5C). This line contains a premature stop codon that truncates the putative protein by 39 amino acids, and is compared in the sperm exhastion assay (D) to an amino acid mutation (AA, position 118 D->N) and w^{1118}. For *CG31406*-RNAi we mated single RNAi and control males to wild type females and compared the total number of offspring produced. For *CG34434-RNAi*, *CG33235*-RNAi and *CG31909* we used a sperm exhaustion assay, and compared fertility on each day across the groups. Only *CG34434* (A) had a significant effect on male fertility.

levels (Figure 4.8, π, dotted lines). Compared to flanking regions, an estimate of divergence controlled for mutation rate (*K*/ π) was higher in the gene regions than in most other parts of the extended region, except for *CG31406*, which showed slightly higher *K*/ π in the intron of the nearby gene *jumu* (Table 4.1). This pattern is generally consistent with positive selection acting on a gene. However, polymorphism-based metrics (Tajima's *D* and Fu and Li's *D* and *F*, Tajima 1989; Fu and Li 1993) did not show significant deviation from neutrality (Table 4.1) for blocks containing the *de novo* genes. The only significant deviation from neutrality was a block containing the gene *CG1751*, which flanks *CG33235*.

We next tested whether the protein-coding regions of these genes showed evidence of adaptive evolution. Each gene had high levels of both synonymous and nonsynonymous divergence when compared to *D. simulans* or (Table 4.2). Results of a Macdonald-Kreitman test and a calculation of the Neutrality Index and Direction of Selection (DoS) are shown in Table 4.2. None of the genes tested show significant evidence indicating that they are evolving under strong positive or purifying selection, so we cannot reject neutral evolution as an explanation for their rapid rate of evolutionary change. The DoS estimates indicate that *CG31909* is the most likely of the four to be evolving under positive selection, though the Macdonald-Kreitman test was not significant. On the other hand *CG33235* and *CG34434* show evidence of purifying selection (DoS is negative). This makes sense, as these two genes are essential for viability *in D. melanogaster.* Similarly, there are no variants segregating in any of these four genes that lead to disruption of the open reading frame, which further indicates they are likely not simply pseudogenes. Likewise, a broad (~100 allele) PCR-based survey of

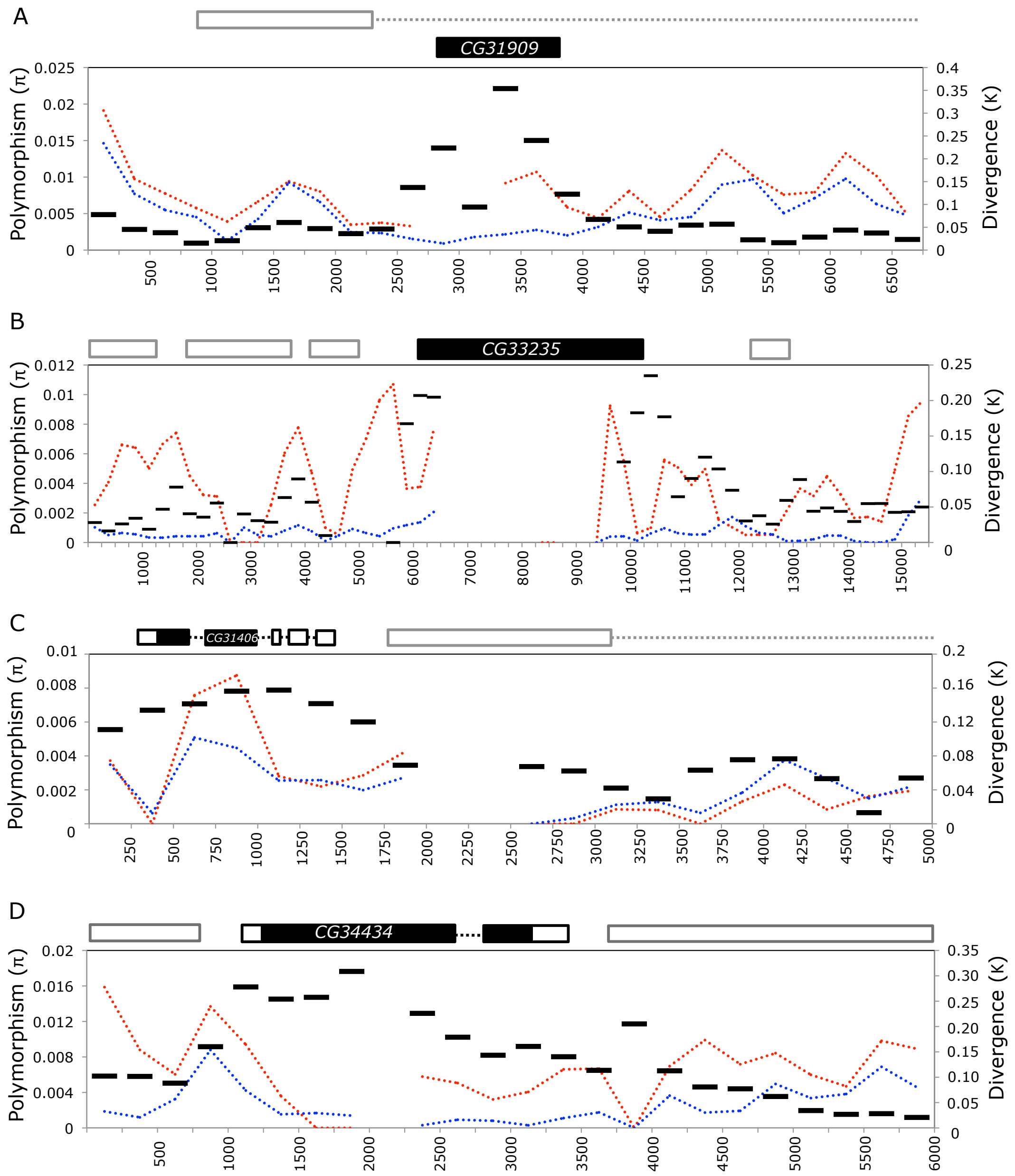


Figure 4.8 - *D. melanogaster de novo* genes are highly diverged relative to neighboring sequences but carry little standing variation
The expanded gene region (5-15kb) surrounding each *de novo* gene was aligned to the colinear sequence from *D. simulans* (using MAUVE) and to *D. melanogaster* genomes from the Drosophila Population Genomics Project (www.dpgp.org). We used Variscan (Hutter, 2006) to calculate divergence to simulans (*K*, black bars) as well as polymorphism (π) from both the North American (blue) and African (red) populations. The large black block shows the position of the focal *de novo* gene, and surrounding outlined boxes are other genes in the region. Over all, the *de novo* genes show elevated divergence (but not polymorphism) relative to surrounding sequences, indicating they may be under the influence of positive selection.

a natural population of *D. melanogaster* for deletions of *CG31909* found that in all cases, the gene was intact (Begun and Saeleo, *personal communication*).

CONCLUSIONS

Recent studies of *de novo* genes have differed in both their methods of identifying candidates, and their definition of what it takes to be considered a truly "new" gene. Work in Drosophila has been predicated on the idea that a *de novo* gene should be a gene that truly has no relatives - hence genes qualify only if they lack sequence similarity to any but the most closely related genomes. Subsequent papers applied this definition to other species groups (Cai et al. 2008; Toll-Riera et al. 2009; Li et al. 2010a; Xiao et al. 2009; Yang and Huang 2011). On the other hand, some studies in primates (Knowles and McLysaght 2009; Wu et al. 2011) defined *de novo* genes by carefully identifying the exact point mutations needed to extend a previously intact (but shorter) open reading frame. Finally, *de novo* genes in mice were identified by looking for lineage-specific transcription in mice (Heinen et al. 2009). None of these latter genes would be considered *de novo* by the first definition.

Ultimately, these definitions may not be in conflict, but may simply be identifying different stages in the evolution of lineage-specific genes across species. A new protein must be both transcribed and translated in order to function, so we may expect that transcription could in some cases occur first, and a protein only evolve subsequently. Alternatively, transcription and translation may originate nearly simultaneously. Logically, a novel protein is more likely to arise form a region that contains a nascent open reading frame - no matter how ill-adapted - than one without.

Table 4.1 – Metrics of neutrality for genes and surrounding regions

			Region				**Tajima's**		**Fu and**	**Fu and**	
		Population	**size**	***S***	***π***	***ϑ***	***D***	***K***	**Li's *D***	**Li's *F***	**K/π**
CG31909	entire region	NA	6414	66	0.006	0.006	-0.113	0.063	0.187	0.123	11.5
		AF	6067	82	0.007	0.007	-0.109	0.058	0.135	0.141	7.7
	intergenic	NA	806	18	0.012	0.012	0.189	0.061	0.946	0.945	4.9
		AF	803	25	0.017	0.017	0.143	0.069	0.506	0.537	3.9
	wnt4 exon	NA	1801	13	0.004	0.004	0.000	0.034	-0.350	-0.481	8.8
		AF	1801	14	0.004	0.004	-0.117	0.033	0.203	0.236	8.0
	CG31909	NA	559	2	0.002	0.002	-0.710	0.182	-1.507	-1.597	102.2
		AF	266	4	0.007	0.008	-0.754	0.199	-0.368	-0.489	26.5
	wnt4 intron	NA	2997	32	0.006	0.006	-0.297	0.049	0.289	0.240	8.7
		AF	2961	39	0.007	0.007	-0.176	0.046	-0.089	-0.097	6.2
CG33235	Entire region	NA	8660	79	0.001	0.002	-2.378	0.061	-3.421	-3.617	85.7
		AF	8706	101	0.004	0.005	-0.190	0.061	-0.540	-0.609	14.8
	Cyp4g15	NA	1331	10	0.001	0.002	-1.810	0.026	-1.054	-1.335	39.9
		AF	1335	19	0.004	0.006	-0.817	0.026	-0.803	-0.784	5.8
	intergenic1	NA	396	3	0.000	0.002	-1.507	0.061	-2.368	-2.476	148.7
		AF	399	8	0.008	0.008	0.039	0.062	-1.190	-1.309	7.4
	CG1749	NA	1553	13	0.000	0.002	-2.296*	0.037	-3.477*	-3.582*	79.7
		AF	1558	15	0.003	0.004	0.687	0.037	-0.633	-0.784	10.8
	CG1751	NA	1018	3	0.000	0.001	-1.196	0.030	NA	NA	89.2
		AF	1018	4	0.001	0.002	-0.612	0.032	1.095	1.301	25.7
	intergenic3	NA	1173	17	0.001	0.003	-2.258	0.167	NA	NA	168.0
		AF	1173	24	0.008	0.008	0.395	0.167	NA	NA	21.0
	CG33235	NA	332	6	0.001	0.004	-1.732	0.200	-1.566	-1.889	213.2
		AF	366	8	0.007	0.009	-0.817	0.205	1.095	0.976	29.9
	intergenic4	NA	1838	18	0.001	0.002	-2.021	0.121	-1.739	-2.041	142.9
		AF	1838	17	0.003	0.004	-0.834	0.122	-0.378	-0.422	40.1
	Drak2	NA	830	7	0.001	0.002	-1.967	0.025	-3.185	-3.371	44.7
		AF	830	2	0.001	0.001	-0.710	0.025	NA	NA	41.7
CG31406	3' intergenic	NA	323	3	0.006	0.005	-0.710	0.110	-0.368	-0.326	19.4
		AF	318	3	0.005	0.005	0.168	0.111	1.441	1.492	21.2
	Entire region	NA	4011	15	0.002	0.002	-0.265	0.085	-0.624	-0.655	41.2
		AF	3946	17	0.002	0.002	0.346	0.085	-0.112	-0.098	35.8
	CG31406	NA	990	6	0.004	0.003	-0.069	0.144	0.151	0.133	40.8
		AF	943	7	0.004	0.004	0.956	0.146	0.168	0.197	34.5
	jumu exon	NA	403	0	0.000	0.000	NA	0.059	-0.913	-0.976	NA
		AF	395	2	0.003	0.003	-0.710	0.056	-1.201	-1.279	22.0
	jumu intron	NA	2019	5	0.001	0.001	-0.314	0.055	-0.803	-0.863	44.6
		AF	2014	4	0.001	0.001	0.168	0.054	-0.678	-0.665	50.5
CG34434	*CG34435*	NA	890	13	0.002	0.004	-1.533	0.094	-0.282	-0.667	47.4
		AF	890	21	0.009	0.009	-0.300	0.094	0.192	0.112	10.2
	Entire Region	NA	4506	63	0.003	0.004	-0.615	0.101	-0.919	-0.894	34.3
		AF	4568	109	0.009	0.009	-0.366	0.106	-0.706	-0.824	12.1
	CG34434	NA	1495	14	0.001	0.002	-1.336	0.202	-2.433	-2.556	142.7
		AF	1495	30	0.008	0.008	-0.071	0.202	-0.447	-0.590	26.3
	RhoGap	NA	1852	32	0.005	0.004	0.216	0.047	-0.292	-0.092	10.1
		AF	1853	49	0.009	0.010	-0.566	0.048	-1.256	-1.351	5.2

Table 4.2 – Neutrality Index, Direction of selection estimates for four *de novo* genes

							DoS		
					NI		Dn/(Dn+Ds)		
	Dn	Pn	Ds	Ps	(Pn/Ps)/(Dn/Ds)	α	- Pn/(Pn+Ps)	MK test (G)	P-value
CG33235	375	13	411	8	1.781	-0.781	-0.142	1.660	0.198
CG31406	52	8	35	7	0.769	0.231	0.064	0.217	0.641
CG31909	37	3	28	6	0.378	0.622	0.236	1.783	0.182
CG34434	103	16	68	7	1.509	-0.509	-0.093	0.765	0.382

Of the four genes considered in this study, it appears that the open reading frame may already have been present, as there exist *potential* open reading frames in every species where transcription occurs. These ORFs are not always annotated as genes in the other species (*CG33235* in all species, and *CG34434* in *D. annanassae*) and are highly diverged in sequence and structure. However, previous studies on *de novo* genes have raised the question of whether these genes are likely to produce proteins. Heinen and colleagues (2009) annotated their newly evolved transcripts as non-coding RNAs despite the presence of open reading frames in these genes. They argued that it is unlikely that a protein from a novel RNA would be functional, and failed to stain the putative protein with a custom peptide antibody. On the other hand, some human *de novo* proteins have evidence of translation from peptide databases, suggesting that these genes are indeed translated (Wu et al. 2011).

All genes considered in the present study are predicted to produce proteins, but currently there is no biological data in support of this annotation. We designed peptide antibodies to *CG34434* and *CG31909* but were unable to stain a target of the appropriate size (data not shown). Given the failure rate of peptide antibodies, we cannot reject the possibility that a protein is produced, but it is possible that as proposed by Heinen and colleagues, RNA is the final product of these genes.

We show that in Drosophila*, de novo* genes are important to the function of the organism, contributing to the most basic functions of life - survival and reproduction. This is consistent with a result in yeast (Cai et al. 2008), which found that loss of a *de novo* gene in a synthetic lethal screen was lethal. A genetic mutation in the youngest *de novo* gene we analyzed (*CG31909*) did not show any obvious viability or fertility effects, leading to the intriguing hypothesis that newer *de novo* genes might be less likely to be essential. Instead, as *de novo* genes age, they may begin to evolve new functions. All four of these genes exhibit testes-biased expression and expression is greatly reduced in males lacking a germline. This further supports the hypothesis that the testes may be a fertile ground for the origin of novel genes. Our results indicate that *CG34434* males have reduced fertility, and *CG31406* showed reduced expression in a meiotic arrest mutant, implying it may be functioning in the sperm or sperm precursor cells. Both genes have become essential in *D. melanogaster*. Perhaps these genes were first expressed in the male germline at a high level, and only later evolved expression and function in other tissues. In this scenario, their current expression pattern is either an evolutionary remnant of their origin, or these genes are maintaining multiple functions, one of which is essential to survival. Evolutionary theory predicts that genes with multiple functions are often under genomic conflict, because each function would presumably push the gene to evolve in a different way. One proposed way out of this conflict is for a gene to duplicate (Hittinger and Carroll 2007) - each copy could then evolve to maximize function. Interestingly, we find that one of the genes (*CG31909*) has a near duplicate copy nearby.

Finally, our study revealed differences in potential origination mechanisms for Drosophila *de novo* genes. Early evolution of a *de novo* gene may be typified by *CG31909*. This gene may have originated due to movement and duplication of a small 5'UTR element that is associated with the expression of several male-biased genes and may have led to the origin of its transcription. In contrast to *CG31909's* sudden appearance, the other genes have increased in complexity gradually over time. This pattern has been predicted previously as a potential evolutionary trajectory for *de novo* genes - a chance event leading to initial transcription, followed by the acquisition of new coding domains and functions. Future work will show whether this is a general phenomenon, or simply specific to these genes.

Acknowledgements

We thank S. Chen for the ActinGAL4:UASGFP stock and H. White-Cooper for the *tombola* mutant flies, as well as personal communications regarding FlyTED. We thank D. Begun for comments on the manuscript. We thank Teni Coker, Betty Wanjiru, Perot Saelao, and Hung-Jui Shih for technical assistance.

CHAPTER FIVE: TWO *DE NOVO* OPEN READING FRAMES EVOLVED FROM PREVIOUSLY TRANSCRIBED SEQUENCES

Authors: Josephine A Reinhardt and Corbin D Jones

ABSTRACT

De novo genes have now been found in a number of species, but the origins and functions of these genes remain obscure. For example, it is not known whether most *de novo* genes encode proteins as soon as they arise, or exist as non-coding RNAs and only later gain protein-coding function. Here we analyzed the evolutionary history and function of two *D. melanogaster de novo* (*CG32690* - now known as *CR32690* - and *CG32582*). Both genes are expressed at a higher level in males than females, and RNAi of *CG32582* leads to lethality during metamorphosis. We find that the gene regions colinear to the *D. melanogaster* gene are transcribed in other closely related species even though they are highly diverged and no open reading frame exists, suggesting that transcription evolved prior to the evolution of the open reading frame.

INTRODUCTION

De novo genes - genes that evolved from previously non-coding sequences - have shifted from being thought of as rare evolutionary novelties to being viewed as an important source of variation. Prior to their initial discovery in *D. melanogaster*, most thought it unlikely that *de novo* genes comprised an appreciable proportion of functional genes. Today, few deny their importance, with some (Zhou et al. 2008; Toll-Riera et al. 2009) claiming *de novo* evolved genes make up as much as 11% of all new genes (e.g. new duplicates, chimeras, and retroposed genes). Yet there is substantial debate about what constitutes a *de novo* gene, how *de novo* genes originate, and how they subsequently evolve. For example, some have defined *de novo* genes as proteins that evolved from conserved sequences through the loss of a stop codon (Knowles and McLysaght 2009). Others have used novel transcription as the key factor (Heinen et al. 2009). Finally, some have suggested that *de novo* genes should not be similar to *any* sequence in other more distantly related species in order to rule out the possibility that unannotated genes might be present in other species (Levine et al. 2006; Begun et al. 2007b; Cai et al. 2008).

What is clear is that for a gene to evolve entirely “from scratch”, it must evolve both protein-coding potential and transcriptional potential. In principle, these events could occur in either order (Figure 5.1). The recent discoveries of many long non-coding RNAs across species have led some to suggest that transcription is likely to occur first (Levine et al. 2006; Tautz and Domazet-Loso 2011). Logically, this makes sense. If a new open reading frame evolves within a transcribed region (such as a lncRNA), it is more likely to ultimately be translated than an ORF that evolves in a region of untranscribed DNA (“transcription first” model, Figure 5.1 left). This idea was put forward with the discovery of the first *de novo* genes in Drosophila (Levine et al. 2006), yet no studies to date have only speculated on possible mechanisms of origin. Here we present a detailed analysis of the evolutionary history and function of two *D. melanogaster de novo* genes previously reported in the literature (*CG32690* - now known

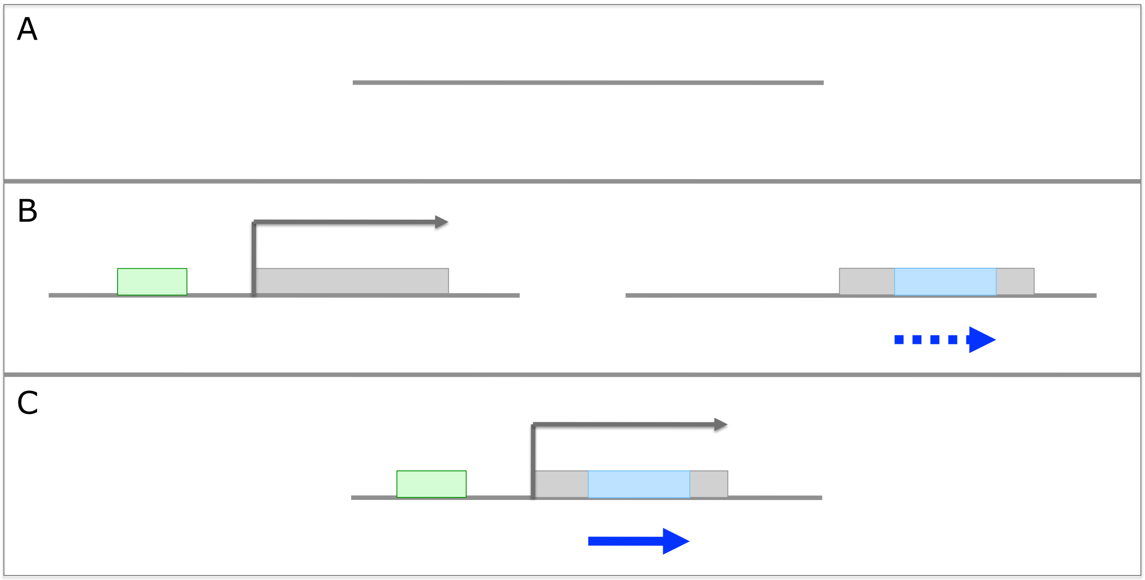


Figure 5.1 - Two models for the origin of *de novo* genes
De novo genes may emerge and evolve into protein coding genes (C) from non-coding sequences (A) through one of several intermediate steps (B). Left - a novel non-coding RNA becomes transcribed after a new promoter (green) is recruited. Right - a "cryptic" ORF (blue) is present prior to the origin of transcription.

as *CR32690* - and *CG32582*). We find that although the open reading frames of these genes are unique to *D. melanogaster* these gene regions are transcribed in other closely related species. Both genes arose recently - within the *D. melanogaster* species subgroup - from a region ancestrally lacking protein-coding DNA, likely via a non-coding RNA intermediate. Subsequently, an open reading frame evolved through numerous steps in the lineage leading to *D. melanogaster*. Both genes exhibit a male-biased expression pattern that is conserved across species, although expression can be detected in both males and females. Finally, as was previously found for a large number of other new *D. melanogaster* genes, RNAi of *CG32582* causes lethality during late pupal development, implying this new gene it has recently become essential.

MATERIAL AND METHODS

Annotation, molecular evolutionary and population genetic analyses

We downloaded BLASTZ (Chiaromonte et al. 2002) alignments of the extended gene regions surrounding *CG32582* and *CR32690* from the UCSC genome database (Fujita et al. 2011). We used these alignments to determine which parts of the *D. melanogaster* putative lineage-specific genes - and their flanking sequences were colinear to sequences in each of the other species. We extracted any portion of the alignment overlapping transcripts of *CG32582* and *CR32690* and realigned pairs of sequences (*D. melanogaster* against each other species) using the water pairwise alignment program (Rice et al. 2000). We calculated the total sequence similarity and the proportion of alignable bases between sections of each gene (e.g. CDS, UTRs, etc) from these pairwise alignments.

We also performed a global pairwise alignment of the *D. melanogaster* and *D. simulans* extended gene regions (extracted from FlyBase genbank files) using progressiveMAUVE (Darling et al. 2004; Darling et al. 2010). We counted the number of fixed differences between *D. melanogaster* and *D. simulans* in 500 bp windows along the alignment, then aligned 39 *Drosophila melanogaster* Raleigh genomes and 6-9 African genomes (www.dpgp.org) to these regions and calculated polymorphism (π) and divergence (K) in each window. We calculated Tajima's D (Tajima 1989) and Fu and Li's D and F (Fu and Li 1993) for 500 base pair windows across the region using Variscan (Hutter et al. 2006).

Tissue collection and dissection

Male reproductive tracts were dissected on ice from whole flies (*D. yakuba, D. simulans, and D. melanogaster*) in PBS. Male reproductive tracts and carcasses were each pooled and then flash frozen in liquid nitrogen. Whole females and males of each species were also collected and flash-frozen. *D. melanogaster* and *D. yakuba* male reproductive tracts were further dissected into accessory glands and testes in PBS and flash frozen. *D. melanogaster* third instar larvae were sexed by identification of genital discs following *Drosophila protocols* (Blair 2000), then flash-frozen. Testes were also dissected from males carrying a null mutation at the gene *tombola* (*tomb*^{GS12862}, stock generously supplied by Dr. Helen White-Cooper), and sons of females mutant for the *tudor* gene (Bloomington stock #1786 - these flies lack a male germline).

Gene expression analyses

We extracted RNA from two or more biological replicates of each dissected tissue using TRIZOL reagent (Invitrogen, Grand Island, NY #15596-026), and synthesized cDNA using M-MLV reverse transcriptase (Invitrogen, Grand Island, NY #28025013). We performed relative qRT-PCR quantification using gene-specific primers and a control primer that worked across all species (*Actin5c*). All qRT-PCR was averaged across two technical replicates.

In addition to our own data, we mined expression information from online databases - FlyAtlas (Chintapalli et al. 2007), modENCODE RNAseq data (Graveley et al. 2011), Baylor RNAseq data (Daines et al. 2011), and FlyTED: Testes expression database (Zhao et al. 2010).

RNAi knockdown

Virgin females from *Actin*-GAL4 (P[Act5C-GAL4]25FO1, Bloomington #4414) were collected and crossed to a line carrying UAS-RNAi constructs for *CG32582* (VDRC #105051). *CyO* (control) and straight winged (RNAi) progeny of both sexes were counted and collected. We collected larvae in the wandering stage and compared expression of the target gene using RT-PCR.

Viability assays

To estimate effects on adult viability, we counted the number of control (*CyO*) and RNAi (straight-winged) progeny eclosing from the RNAi cross (described above). To determine the stage at which lethality was occurring, we crossed the same RNAi line to a stock with the same *Actin*-GAL4 and *CD8*::UAS-GFP on the same chromosome (kindly donated by S. Chen). RNAi or control status can be ascertained at any stage (RNAi larvae/pupae/adults will express GFP). We collected larvae from the cross during the late third instar ("wandering")/prepupal stage, and sorted by GFP expression and sex (Blair 2000). We then allowed each type to continue development and counted the number that survived, or that died prior to pupation or prior to eclosion.

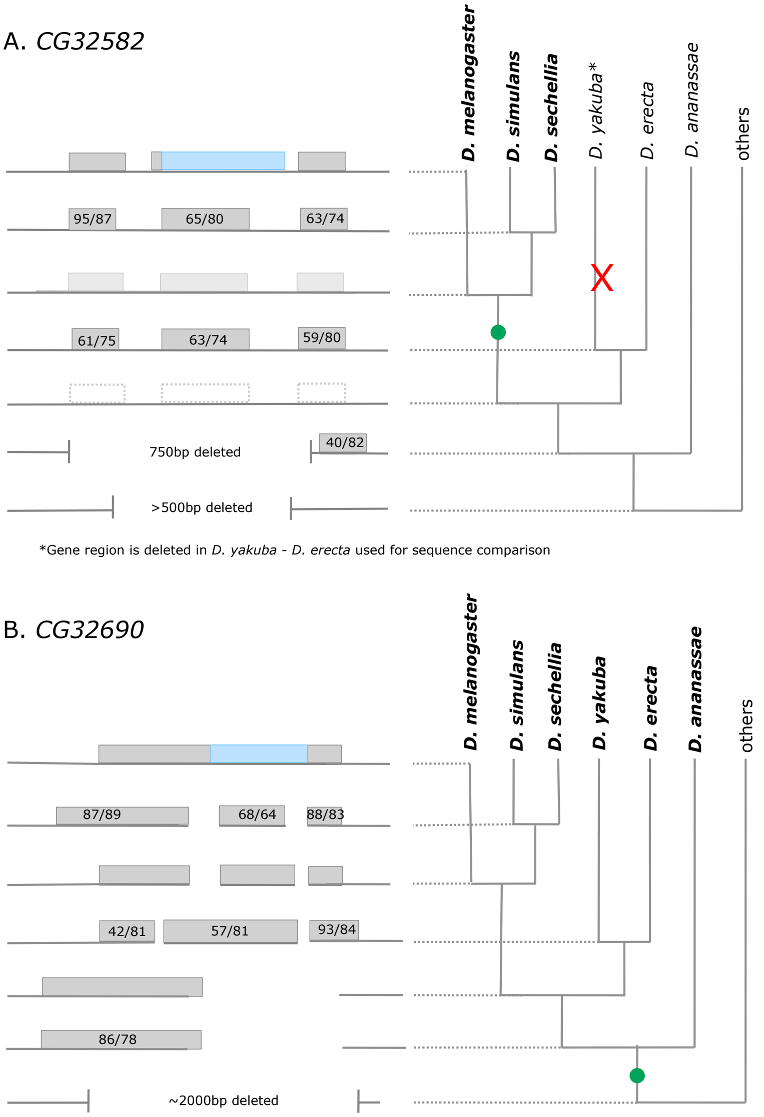
RESULTS

CG32582 **and** ***CR32690*** **are recently evolved transcripts that contain** ***D. melanogaster*** **specific open reading frames**

CG32582 is a multiexonic, putatively protein-coding gene in *D. melanogaster*. It was identified in two previous studies as being a *D. melanogaster* specific *de novo* gene (Chiaromonte et al. 2002; Levine et al. 2006; Zhou et al. 2008). No orthologs for this gene are annotated, and sequences homologous to the CDS are not observed in species more distantly related than *D. yakuba* and *D. erecta* (UCSC BLASTZ chained alignment, Chiaromonte et al. 2002). The 5' UTR of *CG32582* is much better conserved across species than either the CDS or the 3' UTR (Figure 5.2A, 5.3A), but the region is almost entirely missing in *D. yakuba*, *D. annanassae* and all other species. Sequences colinear to the *CG32582* CDS in *D. simulans, D sechellia,* and *D. erecta* carry multiple disabling mutations relative to *D. melanogaster* (Figure 5.3A). Analysis of the four species alignment using a parsimony-based approach demonstrates that at least three mutations were required to create the CDS as it exists today in *D. melanogaster* from non-coding sequences in *D. simulans/D. sechellia:* 1) a large deletion near the start of the gene leading to a frame shift, 2) a two-base pair insertion at the start of the second exon, 3) a

C->T mutation leading to the emergence of a stop codon (though this last only effects one isoform of the gene).

CR32690 was previously (prior to Flybase 5.28, FlyBaseGenomeAnnotators 2010) annotated as a short, single exon protein-coding gene including only the CDS (*CG32690*). Its annotation was revised to a CR (coding RNA) in the Flybase 5.28 update, when evidence was found that more sequence surrounding the CDS was transcribed, making the CDS only 29% of the total (still single exon) transcribed region. The transcribed region of the revised gene model was confirmed with RNAseq data (Daines et al. 2011). Prior to its new CR annotation, *CG32690* was identified in two studies as being a *D. melanogaster*-specific *de novo* protein-coding gene (Levine et al. 2006; Zhou et al. 2008). No genes are annotated as orthologous to *CR32690.* UCSCs BLASTZ (Chiaromonte et al. 2002) alignment (Figure 5.2B) shows sequences exist that are colinear to portions of the *CR32690* transcript in *D. simulans, D. sechellia, D. yakuba,* and *D. erecta,* and *D. annanassae* but not other species. The putative CDS is by far the least similar part of the transcript (Figure 5.2B), and the CDS is deleted in its entirety in D. *erecta* and *D. annanassae*. Sequences colinear to the putative CDS in *D. yakuba, D. simulans,* and *D. sechellia* carry multiple disabling mutations relative to *D. melanogaster* (Figure 5.3B). Comparison to *D. simulans* and *D. sechellia* indicates that that at least three changes in *D. melanogaster* were required to create the 80 AA open reading frame in *D. melanogaster:* 1) insertion of 11 base pairs at the start of the gene including the start codon, 2) deletion of 51 base pair region carrying several stop codons


A. *CG32582*

*Gene region is deleted in *D. yakuba* - *D. erecta* used for sequence comparison

B. *CG32690*

Figure 5.2 - Gene model evolution of *CG32582* and *CG32690*
We used BLASTZ alignments and our own MAUVE alignments to examine the recent evolution of *CG32582* and *CG32690*. The current *D. melanogaster* gene model is shown on top, and blocks of sequence that are colinear to parts of the *D. melanogaster* gene (as determined by BLASTZ global alignment) are shown below. Blue blocks represent protein coding sequence, grey blocks non-coding sequence. *D. simulans, D. yakuba,* and *D. ananassae* colinear blocks are shown at the appropriate nodes. The proportion of *D. melanogaster* bases aligned and the sequence similarity of aligned bases are shown on each block (proportion/similarity). The relative size of blocks indicates the length of the aligned sequence in each species. The inferred gene models of the internal nodes are shown as faded blocks. Finally, expression was measured (using qRT-PCR) in each species where colinear sequence could be found. Species where expression was detected are bolded on the phylogeny and the green dot indicates the inferred start of transcription.

and 3) insertion of a 63 basepair segment encoding 21 amino acids (Figure 5.3B). The proposed RNA harbors potential polyadenylation sites and a polyadenylated mRNA is readily detected (see below). Whether this sequence is likely to be translated is unknown. It is worth noting that even with the new annotation, the FlyBase record indicates that the gene may produce a small protein. The transcribed region contains nine initiation codons upstream of the previously predicted CDS, all of which encode shorter ORFs than the previously annotated *CG32690* protein.

CG32582 **and** ***CR32690*** **are expressed as non-coding RNA outside of** ***D. melanogaster***

No open reading that could produce proteins of similar length to CG32582 or *CR32690* exist in species outside of *D. melanogaster*. However, the novel ORFs may have evolved from sequences that were previously transcribed. We used qRT-PCR to measure expression of *CG32582* and *CR32690* in *D. simulans, D. sechellia, D. erecta, D. yakuba*. Despite the fact that only a tiny portion of the *D. annanassae* genome could be aligned to the gene regions of *CG32582* and *CR32690* (Figure 5.3), we also measured expression of this region. Parts of the sequences colinear to *CR32690* were expressed in all species tested, even the tiny portion of *D. annanassae*. We therefore infer that transcription of this region first occurred in the ancestor of *D. annanassae* and *D. melanogaster* (Figure5.2B, green dot). *CG32582* was expressed only in *D. simulans, D. sechellia, and D. melanogaster*. We used two sets of primers to ensure that expression was not occurring in *D. yakuba* or *D. erecta*, and also tested multiple tissues (see below). Transcription of *CG32582* at its current levels therefore evolved in the common ancestor of *D. melanogaster* and *D. simulans* (Figure 5.2A, green dot).

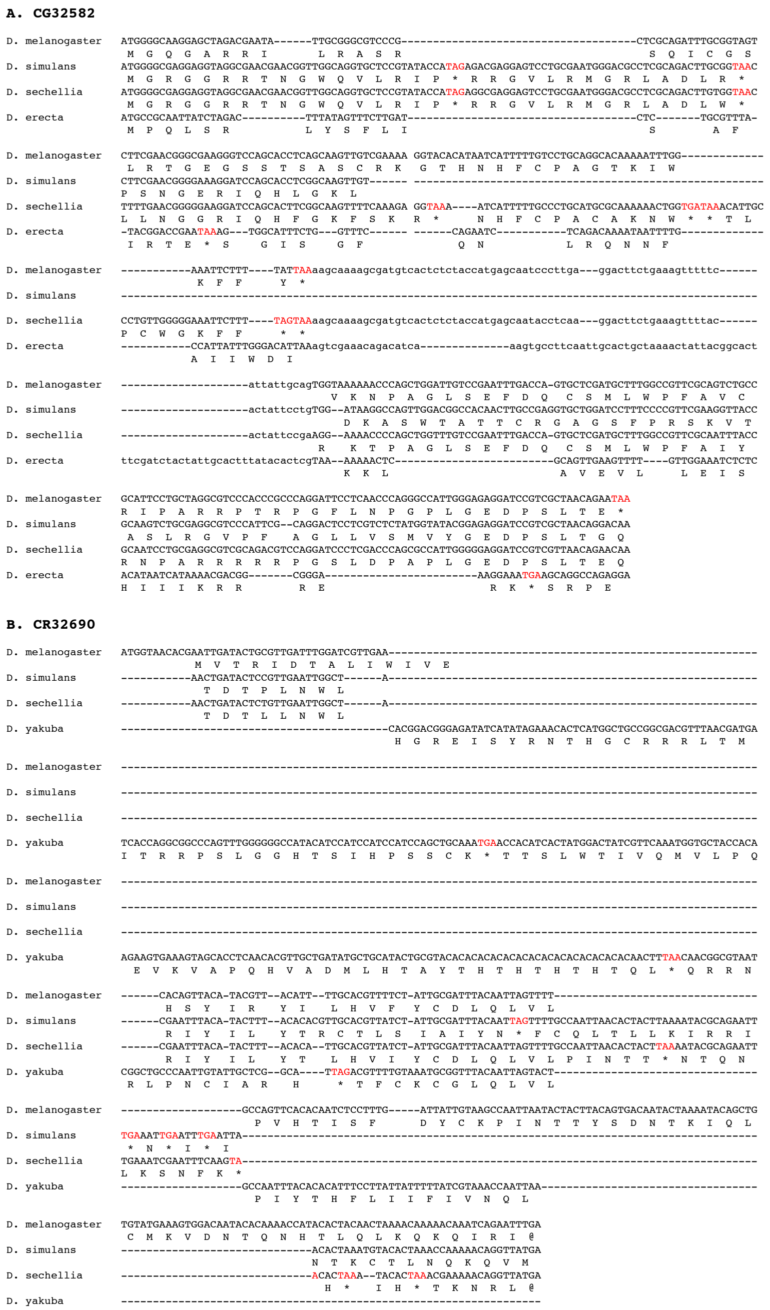


Figure 5.3 - Protein coding potential of *CG32582* and *CG32690* is *D. melanogaster* specific

We searched the UCSC BLASTZ alignments for regions of colinearity with (A) *CG32582* and (B) *CG32690*. In species where there was colinearity (*D. simulans, D. sechellia* and either *D. yakuba or D. erecta)* we translated the colinear sequences that there are multiple nonsense mutations in each species other than *D. melanogaster*. In-frame stop codons are shown in red. For *CG32582*, there are two potential open reading frames due to splicing variation. However, in-frame stop codons are present in each species prior to the *D. melanogaster* splice site (blue vertical line).

Testes biased expression is conserved across species

Prior work shows that *CR32690* and *CG32582* both exhibit male-biased expression, and are expressed at their highest levels in L3 larvae, pupae, and adult males (Graveley et al. 2011). Based on prior work (Levine et al. 2006) we expected expression of both genes to be restricted to the male reproductive system. We compared expression in *D. melanogaster* in adult testes, male accessory glands, the remainder of the male tissues, and adult females. In addition, we sexed L3 larvae (Blair 2000) and measured expression in male and female larvae. We found that expression of both genes was at its highest in the testes, and that male larvae expressed at a higher level than female larvae (Figure 5.4). We also found that lack of a male germline - measured in testes from male flies lacking a germline (Figure 5.4 *sons-of-tudor,* light green) - reduces expression of *CR32690* and *CG32582* in the testes (expression of CR32690 was undedectable in the tudor testes, while *CG32582* expression was reduced by approximately 500-fold). Expression of many testes-specific genes are under the control of meiotic arrest genes (e.g. *tombola*, Jiang 2007), but we found that expression of both genes was normal in a *tomb* background, indicating they function in parallel independently of the meiotic arrest pathways.

We then compared expression levels of colinear expressed sequences in tissues (testes, male carcass, and female) from *D. simulans, D. sechellia, D. yakuba* and *D. erecta* (Figure 5.5A-D). We measured *D. annanassae* expression only in cDNA from whole fly, and compared to amplification from genomic DNA (Figure 5.5E). *CR32690* expression was trending towards the same testes bias seen in *D. melanogaster* across five species. We found that expression of *CG32582* tended to be higher in the testes of *D.*

simulans and *D. sechellia* than in other tissues*,* but expression was either very low or not detectable in *D. yakuba, D. erecta, and D. ananassae*. This suggests that transcription of these genes has been testes-biased since the genes first originated.

RNAi of *CG32582* leads to arrest at the pharate pupal stage

We used RNAi to knock down expression of *CG32582* (VDRC#105051) by crossing the RNAi line to a ubiquitous driver (*Actin*-GAL4 BLM#4414). This driver has been used previously with the Vienna RNAi lines - including in a study where a number of new duplicates were found to be essential (Chen et al. 2010). This study found that roughly 30% of new duplicates and retroposed genes were essential, demonstrating that this driver is typically not lethal in combination with Vienna RNAi lines unless they have gene-specific effects. In our hands, constructs targeting non-essential genes (*Gr22c*, VDRC#104704) did not cause lethality when crossed to this driver.

No adult RNAi offspring resulted from our RNAi cross. To determine at what stage lethality was occurring, we used a line that carried the same Actin-GAL4 driver and a UAS-GFP marker (Kindly donated by S. Chen, Chen et al. 2010) and sorted larvae at the L3 stage. Surprisingly, we found no lethality had occurred before the late third instar, and the larvae pupated and appeared to develop as normal. However, no adults carrying the RNAi construct and driver eclosed from their pupae (Figure 5.6A). Observationally, these arrested flies were fully developed within their pupae with eyes, wings, and legs clearly visible (the "pharate" stage, Figure 5.6B).

This result suggests that *CG32582* is important for metamorphosis. However, *CG32582* may also be important at earlierstages, but the larvae may tolerate its absence better than the developing adult. For example, the gene product might be expressed at a weak level despite RNAi, and the amount in storage could be sufficient to sustain life through to pupation, but no further. According to expression databases, there are no detectable *CG32582* transcripts until at least the second larval instar, ruling out persistent maternal transcripts and implying that loss of this gene is most likely to be disrupting normal function only later in development, consistent with our results.

We collected RNA from RNAi and control L3 larvae, and measured expression of *CG32582* to confirm knockdown. Unfortunately, we were unable to confirm RNA knock down was occurring in the RNAi larvae even though no adult flies emerge from the RNAi cross. *CG32582* is expressed at a relatively low level, and expression is lower in male larvae than in adult males. Future work will test knockdown throughout late larval and pupal development. As lethality is occurring even when we cannot confirm knockdown, we need to test if lethality is a by-product of the lines we are using. Off targets of RNAi are one possibility - however, the *CG32582* open reading frame - to which the RNAi construct was designed - is unique, and there are no predicted off-targets for the construct used. We (and other groups, Chen et al. 2010) also confirmed that the Actin-GAL4 driver was not necessarily lethal when crossed to any VDRC RNAi lines. These results suggest that our failure to observe knockdown at L3 and our observation of pupal lethality simply reflects the need to measure knockdown later in development when *CG32582* is more highly expressed.

***CR32690* and *CG32582* are evolving rapidly**

CR32690 and *CG32582* were identified during a screen to find genes that are

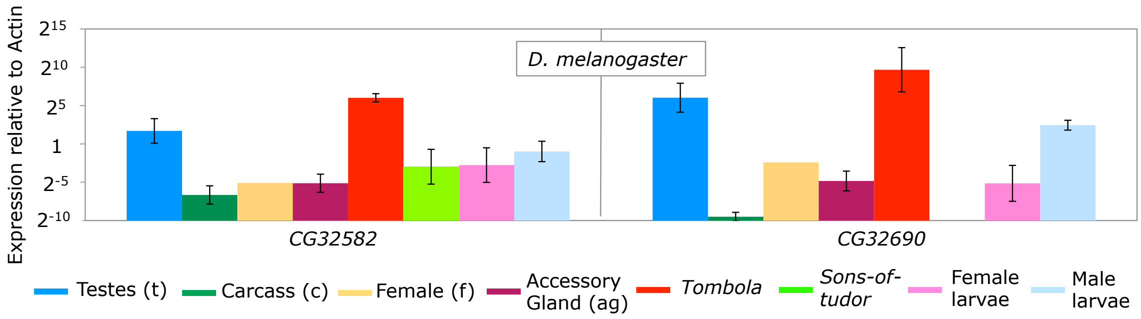


Figure 5.4 - *CG32582 and CG32690* genes exhibit male-biased and germline-dependent expression
We compared the expression of *CG32582* (A) and *CG32690* (B) in a variety of tissues dissected from *D. melanogaster* using qRT-PCR. Expression is relative to the reference gene *Actin*. Expression of both genes was highest in the testes, confirming results from online databases, and was reduced in testes of males lacking a gremline (*Tudor*, light green), but loss of the meiotic arrest gene *tombola* did not significantly affect testes expression. Finally, we found that male larvae express both genes at a higher level than females (pink versus light blue).

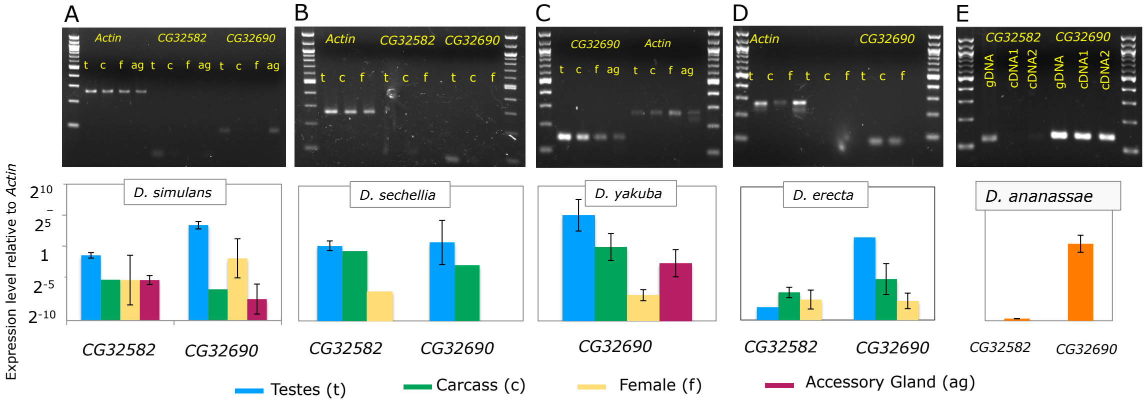


Figure 5.5 - *CG32582* and *CR32690* exhibit testes-biased expression in all species where they are transcribed
We compared the expression of sequences that were colinear to *CG32582* and *CR32690* across a number of tissues in the species of the *melanogaster* subgroup. In *D. sechellia* and *D. erecta,* we dissected male reproductive tracts from flies, and compared expression in the male reproductive tracts (Testes, blue), the remainder of the male (Carcass, green), and whole females (Females, yellow). In *D. yakuba* and *D. simulans,* we further dissected male reproductive tracts into testes and accessory glands (purple). When available, two biological replicates are shown.

lineage-specific based on a lack of sequence similarity to other species (Levine 2006). As such, it is not surprising that these genes are highly diverged at the sequence level even when compared to close relatives. We wanted to know whether natural selection or neutral processes (e.g. a mutational "hotspot") caused this rapid sequence evolution. As the open reading frames for these genes exist only in *D. melanogaster*, we are unable to determine the rate or type of amino acid divergence between species. We can however compare nucleotide polymorphism and divergence.

We compared polymorphism from an African and North American sample (www.dpgp.org) with nucleotide divergence relative to *D. simulans* across the gene regions containing *CR32690* and *CG32582* and also calculated a number of metrics (Table 5.1, Figure 5.7) using the population genetics package Variscan (Hutter et al. 2006). We found that for both genes, the CDS showed the highest ratio of divergence (*K*) to polymorphism (π), compared to other parts of the extended gene region (Table 5.1). Likewise, 500 base pair windows overlapping the gene regions also showed peak levels of divergence and low levels of polymorphism (Figure 5.7) compared to flanking sequences - whether or not they contained gene. These patterns are expected if a gene is undergoing repeated selective sweeps to fix new variants. However, polymorphism-based tests (Tajima's *D*, Fu and Li's *D* and *F*) did not show significant deviations from the null hypothesis of neutral evolution for the CDS of the gene (Table 5.1) or any other part of the gene region. The extremely low number of polymorphic sites and short gene regions analyzed - there were only seven and five polymorphisms in the region overlapping the *CG32582* and *CR32690*

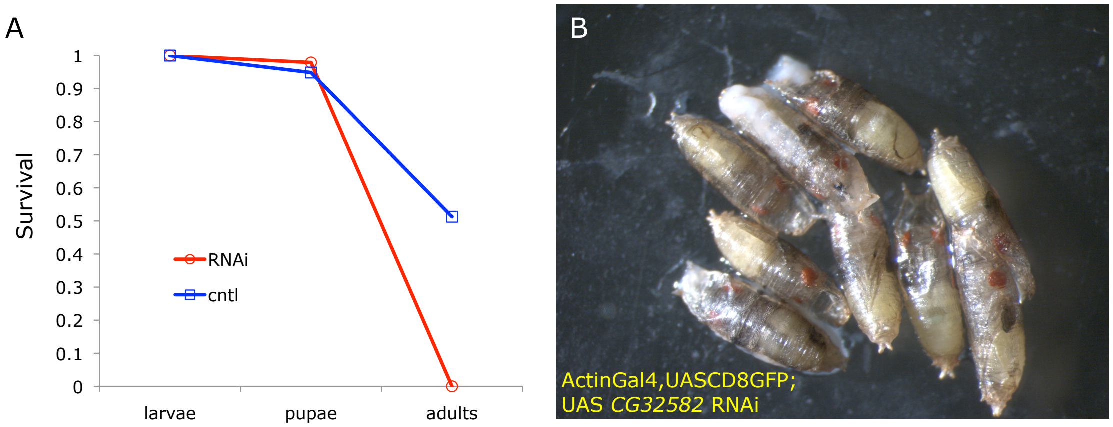


Figure 5.6 - *CG32582-RNAi* flies die during the pharate pupal stage
We crossed UAS *CG32582*-RNAi flies to a stock carrying an Actin-GAL4 driver and a GFP marker and tracked survival of GFP (RNAi, red line) and non-GFP (control, blue line) individuals at the wandering larval stage, the start of pupation, and eclosion (A). *CG32582*-RNAi pupae arrested just prior to eclosion (B) with a number of adult features visible (e.g., eyes, wings, legs). We attempted to detect knockdown of *CG32582* in the wandering larvae but were unable to detect a difference between control and RNAi samples, suggesting that the critical expression period may be after the third larval instar.

Table 5.1 – Metrics of neutrality for genes and surrounding regions

			Region				**Tajima's**		**Fu and**	**Fu and**	
		Population	**size**	***S***	***π***	**θ**	***D***	***K***	**Li's *D***	**Li's *F***	***K/π***
CG32582	Entire Region	NA	6101	39	0.0018	0.0017	0.142	0.050	-0.400	-0.128	27.31
		AF	6101	67	0.0039	0.0047	-0.261	0.050	-0.571	-0.634	12.85
	Chc-exons	NA	4243	28	0.0019	0.0016	0.726	0.025	-0.292	0.149	13.13
		AF	4243	44	0.0036	0.0041	-0.042	0.025	-0.323	-0.367	7.02
	intergenic	NA	900	6	0.0020	0.0022	-0.792	0.135	-0.686	-0.778	67.45
		AF	900	12	0.0057	0.0075	-0.389	0.130	-1.380	-1.466	22.74
	***CG32582* region**	NA	958	5	0.0011	0.0018	-1.272	0.151	0.010	-0.241	133.72
		AF	958	11	0.0044	0.0059	-0.824	0.150	-0.678	-0.799	34.28
	***CG32582* cds**	NA	518	3	0.0008	0.0019	-1.360	0.169	-0.784	-0.991	222.68
		AF	518	4	0.0028	0.0042	-0.780	0.171	-0.913	-0.976	61.54
CG32690	Entire Region	NA	10802	186	0.0030	0.0049	-1.387	0.071	-1.150	-1.513	23.99
		AF	10802	256	0.0089	0.0108	-0.298	0.072	-0.616	-0.658	8.07
	***CG32690* region**	NA	1470	13	0.0029	0.0039	-1.266	0.192	-0.464	-0.806	66.70
		AF	1470	14	0.0050	0.0066	-0.389	0.191	0.446	0.349	37.96
	***CG32690* cds**	NA	231	2	0.0012	0.0021	-1.256	0.251	-0.784	-0.991	206.51
		AF	231	3	0.0046	0.0051	0.592	0.257	1.095	0.976	55.24
	CG15309	NA	1719	15	0.0011	0.0021	-1.520	0.055	-1.469	-1.622	49.91
		AF	1719	25	0.0050	0.0057	-0.300	0.055	-0.350	-0.481	11.07

Table 5.2 – Polymorphisms in *CG32582* and *CG32690*

	Nucleotide position	Codon	Protein position	Frequency	Protein coding change
CG32582 PB	84	ACC -> ACA	28	1:47	Synonymous
	123	GGA -> GGT	41	1:47	Synonymous
	193	ACC -> GCC	65	1:35	T -> A
	202	GGA -> CGA	68	3:37	G->R
	228	GGG -> GGA	75	1:46	Synonymous
	253	TAT -> TTA	85	1:46	@->Y
CG32690	40	GAA -> AAA	14	1:46	E->K
	42	GAA -> GGG	14	1:46	Y->S
	47	TAC -> TCC	16	8:38	Synonymous
	169	CAG -> TAG	56	1:47	Q->@

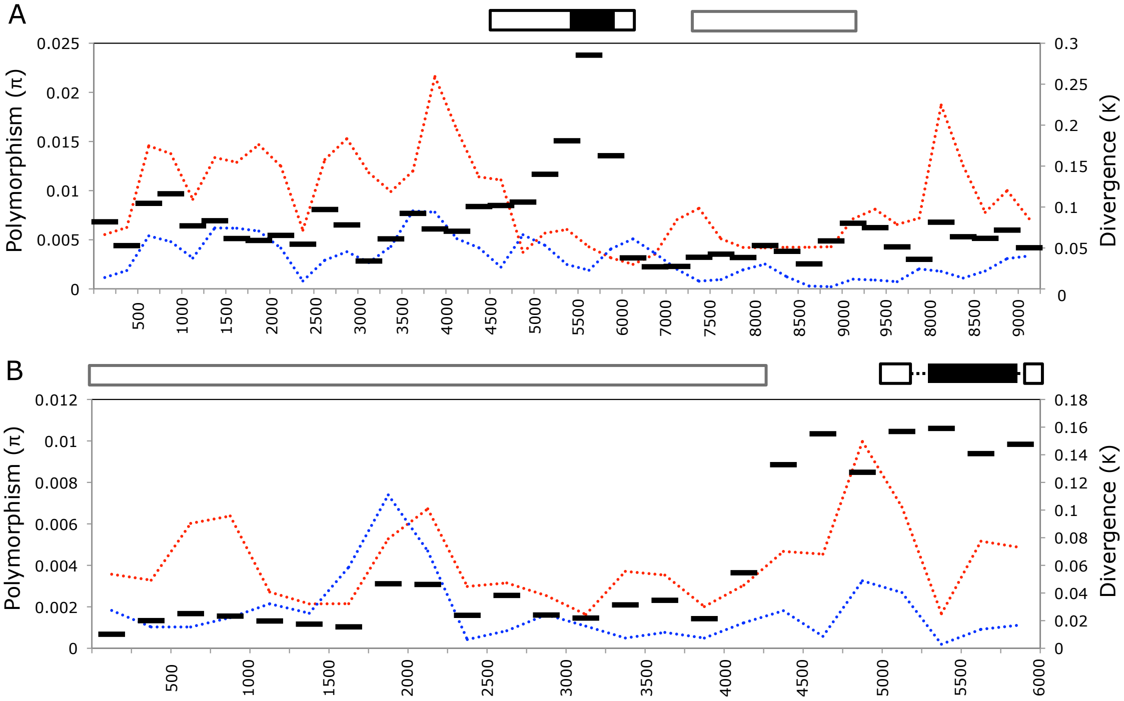


Figure 5.7 –*CG32582* and *CR32690* are diverging rapidly compared to the surrounding regions

The expanded gene region (5-15kb) surrounding each *de novo* gene was aligned to the colinear sequence from *D. simulans* (using MAUVE) and to *D. melanogaster* genomes from the Drosophila Population Genomics Project (DPGP.org). We used Variscan (Hutter 2006) to calculate divergence to *D. simulans* (*K*, black bars) as well as polymorphism (π) from both the North American (blue) and African (red) populations. The large black block shows the position of the focal gene, and surrounding outlined boxes are other genes in the region. Over all, the *de novo* genes show elevated divergence (but not polymorphism), indicating they may have evolved through repeated selective sweeps, or that they evolved rapidly, and are now under purifying selection.

CDS respectively - may have reduced our statistical power. Finally, we examined the coding properties of polymorphisms in the *CG32582* and *CR32690* CDS (Table 5.2). Two of six polymorphisms cause protein-coding changes in *CG32582*, and one line (Malawi-28) carried a mutation causing a small expansion of the *CG32582* open reading frame. Of the four polymorphisms *CR32690* harbors, two are nonsynonymous and one is a nonsense mutation leading to a 23 amino acid truncation. This may suggest that the *CR32690* ORF is under only weak purifying selection. In sum, while we cannot entirely reject neutral processes, our data at least suggests that selective forces are involved in determining the evolutionary trajectory of both genes since divergence from *D. simulans*.

DISCUSSION

Since they were first discovered, *de novo* genes have proved a puzzle to biologists. These are genes that are restricted to only one or a few species, and their origins and functions have remained elusive. Whether these genes produce proteins or instead function as non-coding RNAs has been particularly controversial. Some groups (Heinen et al. 2009) argue that the genes they discovered are almost certainly non-coding RNAs - others have found evidence that proteins are encoded by *de novo* genes (Wu et al. 2011). Even more surprisingly, a potential *de novo* gene in Drosophila was reported to be essential, along with a number of other new genes (Chen et al. 2010).

Here, we report a functional and molecular evolutionary analysis of two *de novo* genes in *D. melanogaster.* Both genes were first reported as protein-coding genes whose ORFs are disrupted in all other species, and cannot be found in any species further diverged than the *D. yakuba* and *D. erecta* species group. We find that despite the fact that the ORFs are disrupted, both gene regions are transcribed in *D. simulans* and *D. sechellia* and the *CR32690* region is transcribed in all 5 species where it can be found. Further, the expression pattern is similar across species. This indicates that regardless of their potential to function as protein-coding genes in *D. melanogaster*, this pair of *de novo* genes could have functioned as non-coding RNA first, supporting the "transcription-first" hypothesis for the origin of *de novo* genes (Levine et al. 2006; Tautz and Domazet-Loso 2011). If this is a general pattern, and new open reading frames arise commonly, we might expect to find more novel open reading frames evolving - and persisting - in genomic regions where transcription is more permissive, or indeed evolving from known non-coding RNAs. For example, some research suggests that portions of the X chromosome where the dosage compensation complex binds may be more transcriptionally complex than other portions of the X or autosomes (Marin 2000). We collected preliminary data that *CG32582* is essential - *CG32582* RNAi flies failed to eclose from their pupae. But how does loss of *CG32582* cause lethality? We know that RNA has been transcribed from this region prior to the emergence of the open reading frame, making the non-coding RNA the more likely candidate for essential function. Some long non-coding RNAs act by regulating other genes, often through base pair matching. We searched the *D. melanogaster* genome for potential targets and found small (~20-30bp) matches of the UTRs of *CG32582* to the 3' UTR of two genes, *CG8119/CR43299* and *CG8928*. If we can confirm the knockdown of *CG32582* in the future, these genes would be good candidates for further study.

Another question that remains open is whether these genes are typically translated into proteins. Some - but not all - human *de novo* genes show evidence of translation from peptide databases (Wu et al. 2011), and we were unable to find evidence of polypeptides from either *CG32582* or *CR32690* in existing databases (Takemori and Yamamoto 2009), though this represents a small sampling of the proteins in *D. melanogaster*. Heinen and colleagues (2009) found that a newly transcribed gene in mouse was important to fertility, and surmised because of the failure of an antibody to bind the protein that the gene was not translated into a protein as predicted. This could certainly be done in Drosophila as well, though Western Blots may fail to stain a target protein for a variety of reasons. We did, however, find a single premature stop codon in *CR32690* segregating in flies from a natural population*,* which suggests that if a protein is produced, it is most likely not as essential as *CG32582.* However, its low frequency in the population is comparable segregating stop codons in many other genes observed in a survey of stop codon polymorphisms (Chapter 2).

If they do encode proteins, these transcripts both existed as RNA first. The finding that *CG32582* RNAi causes lethality could in principle be explained by an essential function for the RNA or the novel protein. It seems more likely that the older gene product - the RNA - would perform the essential function, because it will have had more time to integrate into existing regulatory networks. We can imagine a scenario wherein the RNA continues to perform its essential function, while the protein-coding portion is free to evolve towards some other evolutionary end - provided this does not interfere with the function of the RNA. This mechanism mirrors the “RNA world” hypothesis put forward regarding the origin of the first protein-coding genes. RNA may remain a key intermediate in the origin of new proteins.

Acknowledgements

We thank S. Chen for the *Actin*GAL4:UASGFP stock and H. White-Cooper for the *tombola* mutant flies, as well as personal communications regarding FlyTED. We thank D. Begun for comments on the manuscript.

CHAPTER SIX: CONCLUSIONS

To understand how species evolve, we must understand the nature of the genetic variation that underlies differences between and within species. It is increasingly acknowledged that a model of gradual protein evolution may be insufficient to explain the entirety of adaptive change observed in nature. More extreme genetic changes occur regularly and some of these contribute to adaptation. For example, dramatic changes in gene regulation can lead to relatively radical phenotypic change (Wray 2007).

In my work, I have shown that a group of lineage-specific genes that exist in a few fruit fly species have become vital to the survival and function of these organisms. My dissertation shows that variants with striking changes in gene structure arise commonly within species and are tolerated (Chapter two), that very rapidly evolving genes that appear to be lineage-specific can nevertheless be essential (Chapter three), and that *de novo* genes - whether they arose first as protein-coding genes (Chapter four) or as non-coding RNA genes (Chapter five) - are often essential. The recent molecular evolution of these genes is also similar; all of the genes presented here show low levels of polymorphism but have recently diverged rapidly. We were surprised to find that - as a rule - lineage specific genes in *D. melanogaster* were transcribed in all species where orthologous sequence could be found, whether or not a gene was annotated, and that their expression pattern was invariably conserved across species. The one exception was *CG32582*, where we inferred that the evolution of transcription was followed later by acquisition of an open reading frame. However, it is still not clear whether a protein is produced by any of these genes, leaving open the question of whether they function as non-coding RNAs or proteins. Given the recent interest in long non-coding RNAs (Ponting et al. 2009), and the increasing availability of transcriptome data, comparison of long non-coding RNAs between species may uncover the evolution of more lineage-specific open reading frames that began as long non-coding RNAs in Drosophila and other species. Likewise, subsequent investigations of the molecular functions of the known Drosophila *de novo* genes will reveal how they have become essential in such a short time. Have these genes taken over a role once performed by another gene, or do they interact with existing genes and genetic networks in new ways?

In sum, we find that lineage-specific genes in Drosophila are diverse in their mechanism of origin but surprisingly similar in their putative functions and subsequent molecular evolution. While many mysteries remain to be solved regarding the exact molecular role lineage-specific genes play, there remains little doubt that they contribute novel molecular tools to the organism and that these tools are integrated into the fly's toolkit far more rapidly than was thought possible.

REFERENCES

Abdi H. 2007. Bonferroni and Sidak corrections for multiple comparisons. In: Salkind N, editor. Encyclopedia of Measurement and Statistics. Thousand Oaks, CA: Sage

Adams EM, Wolfner MF. 2007. Seminal proteins but not sperm induce morphological changes in the *Drosophila melanogaster* female reproductive tract during sperm storage. J Insect Physiol 53:319-331.

Aguadé M. 1998. Different forces drive the evolution of the *Acp26Aa* and *Acp26Ab* accessory gland genes in the *Drosophila melanogaster* species complex. Genetics 150:1079-1089.

Ahituv N, Zhu Y, Visel A, Holt A, Afzal V, Pennacchio LA, Rubin EM. 2007. Deletion of ultraconserved elements yields viable mice. PLoS Biol 5:e234.

Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics 136:927-935.

Artieri C, Mattiuzzo N, Malone J, Oliver B. 2011. Comparison of dissected reproductive tracts and remaining carcasses between *D. simulans* and *D. pseudoobscura*.

Avila FW, Wolfner MF. 2009. *Acp36DE* is required for uterine conformational changes in mated Drosophila females. Proc Natl Acad Sci USA 106:15796-15800.

Bates D, Maechler M. 2009. lme4: Linear mixed-effects models using S4 classes.

Begemann G, Paricio N, Artero R, Kiss I, Perez-Alonso M, Mlodzik M. 1997. muscleblind, a gene required for photoreceptor differentiation in Drosophila, encodes novel nuclear Cys3His-type zinc-finger-containing proteins. Development 124:4321-4331.

Begun DJ, Aquadro CF. 1993. African and North American populations of *Drosophila melanogaster* are very different at the DNA level. Nature 365:548-550.

Begun DJ, Holloway AK, Stevens K et al. 2007a. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biol 5:e310.

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

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

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

Behm-Ansmant I, Gatfield D, Rehwinkel J, Hilgers V, Izaurralde E. 2007. A conserved role for cytoplasmic poly(A)-binding protein 1 (PABPC1) in nonsense-mediated mRNA decay. Embo J 26:1591-1601.

Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, Haussler D. 2004. Ultraconserved elements in the human genome. Science 304:1321-1325.

Bjorklund AK, Ekman D, Light S, Frey-Skott J, Elofsson A. 2005. Domain rearrangements in protein evolution. J Mol Biol 353:911-923.

Blair SS. 2000. Imaginal Discs. In: William Sullivan MA, R. Scott Hawley, editor. Drosophila Protocols. Cold Spring Harbor Laboratory Press. p. 159-173

Boswell RE, Mahowald AP. 1985. tudor, a gene required for assembly of the germ plasm in *Drosophila melanogaster*. Cell 43:97-104.

Burkhart BD, Montgomery E, Langley CH, Voelker RA. 1984. Characterization of Allozyme Null and Low Activity Alleles from Two Natural Populations of *DROSOPHILA MELANOGASTER*. Genetics 107:295-306.

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

Cai JJ, Woo PC, Lau SK, Smith DK, Yuen KY. 2006. Accelerated evolutionary rate may be responsible for the emergence of lineage-specific genes in ascomycota. J Mol Evol 63:1-11.

Chang JC, Kan YW. 1979. beta 0 thalassemia, a nonsense mutation in man. Proc Natl Acad Sci U S A 76:2886-2889.

Chang YF, Imam JS, Wilkinson MF. 2007. The nonsense-mediated decay RNA surveillance pathway. Annu Rev Biochem 76:51-74.

Chapman T, Bangham J, Vinti G, Seifried B, Lung O, Wolfner MF, Smith HK, Partridge L. 2003. The sex peptide of *Drosophila melanogaster*: female post-mating responses analyzed by using RNA interference. Proc Natl Acad Sci USA 100:9923-9928.

Chapman T, Herndon LA, Heifetz Y, Partridge L, Wolfner MF. 2001. The Acp26Aa seminal fluid protein is a modulator of early egg hatchability in *Drosophila melanogaster*. Proc Biol Sci 268:1647-1654.

Charlesworth B. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat Rev Genet 10:195-205.

Chen S, Zhang YE, Long M. 2010. New genes in Drosophila quickly become essential. Science 330:1682-1685.

Chiaromonte F, Yap VB, Miller W. 2002. Scoring pairwise genomic sequence alignments. Pac Symp Biocomput:115-126.

Chintapalli VR, Wang J, Dow JA. 2007. Using FlyAtlas to identify better *Drosophila melanogaster models* of human disease. Nat Genet 39:715-720.

Clark AG, Begun DJ. 1998. Female genotypes affect sperm displacement in Drosophila. Genetics 149:1487-1493.

Clark AG, Eisen MB, Smith DR et al. 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature 450:203-218.

Cooper JL, Till BJ, Henikoff S. 2008. Fly-TILL: reverse genetics using a living point mutation resource. Fly (Austin) 2:300-302.

Daines B, Wang H, Wang L et al. 2011. The *Drosophila melanogaster* transcriptome by paired-end RNA sequencing. Genome Res 21:315-324.

Darling ACE, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14:1394-1403.

Darling AE, Mau B, Perna NT. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS ONE 5:e11147.

Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 4:P3.

Dietzl G, Chen D, Schnorrer F et al. 2007. A genome-wide transgenic RNAi library for conditional gene inactivation in Drosophila. Nature 448:151-156.

Ding Y, Zhao L, Yang S et al. 2010. A young Drosophila duplicate gene plays essential roles in spermatogenesis by regulating several Y-linked male fertility genes. PLoS Genet 6:e1001255.

Dubinin N, Romashov, DD, Heptner, MA, and Demidova, ZA. 1937. Aberrant polymorphism in *Drosophila fasciata* Meig. B. Zh., 6:311-354.

Durbin R. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061-1073.

Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA, Hurles ME, McVean GA. 2010. A map of human genome variation from population-scale sequencing. Nature 467:1061-1073.

Duret L, Mouchiroud D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc Natl Acad Sci U S A 96:4482-4487.

Dworkin I, Jones CD. 2009. Genetic changes accompanying the evolution of host specialization in *Drosophila sechellia*. Genetics 181:721-736.

Ellegren H. 2011. Sex-chromosome evolution: recent progress and the influence of male and female heterogamety. Nat Rev Genet 12:157-166.

Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M. 2008. Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*. Science 320:1629-1631.

Ewing B, Hillier L, Wendl MC, Green P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8:175-185.

FlyBaseGenomeAnnotators. 2010. Changes affecting gene model number or type in release 5.28 of the annotated *D.melanogaster* genome.

Friedman EJ, Temple BR, Hicks SN, Sondek J, Jones CD, Jones AM. 2009. Prediction of protein-protein interfaces on G-protein beta subunits reveals a novel phospholipase C beta2 binding domain. J Mol Biol 392:1044-1054.

Frith MC, Forrest AR, Nourbakhsh E, Pang KC, Kai C, Kawai J, Carninci P, Hayashizaki Y, Bailey TL, Grimmond SM. 2006. The abundance of short proteins in the mammalian proteome. PLoS Genet 2:e52.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. Genetics 133:693-709.

Fujita PA, Rhead B, Zweig AS et al. 2011. The UCSC Genome Browser database: update 2011. Nucleic Acids Res 39:D876-882.

Gatfield D, Unterholzner L, Ciccarelli FD, Bork P, Izaurralde E. 2003. Nonsense-mediated mRNA decay in Drosophila: at the intersection of the yeast and mammalian pathways. Embo J 22:3960-3970.

Graveley BR, Brooks AN, Carlson JW et al. 2011. The developmental transcriptome of *Drosophila melanogaster*. Nature 471:473-479.

Guo X, Su B, Zhou Z, Sha J. 2009. Rapid evolution of mammalian X-linked testis microRNAs. BMC Genomics 10:97.

Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P. 2005. Multilocus patterns of nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. Genome Res 15:790-799.

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

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

Hittinger CT, Carroll SB. 2007. Gene duplication and the adaptive evolution of a classic genetic switch. Nature 449:677-681.

Huang da W, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44-57.

Hutter S, Li H, Beisswanger S, De Lorenzo D, Stephan W. 2007. Distinctly different sex ratios in African and European populations of *Drosophila melanogaster* inferred from chromosomewide single nucleotide polymorphism data. Genetics 177:469-480.

Hutter S, Vilella AJ, Rozas J. 2006. Genome-wide DNA polymorphism analyses using VariScan. BMC Bioinformatics 7:409.

Ives PT. 1945. The Genetic Structure of American Populations of *Drosophila melanogaster*. Genetics 30:167-196.

Jiang J, Benson E, Bausek N, Doggett K, White-Cooper H. 2007. Tombola, a tesmin/TSO1-family protein, regulates transcriptional activation in the Drosophila male germline and physically interacts with always early. Development 134:1549-1559.

Jiggins FM, Kim KW. 2005. The evolution of antifungal peptides in Drosophila. Genetics 171:1847-1859.

Juni N, Yamamoto D. 2009. Genetic analysis of chaste, a new mutation of *Drosophila melanogaster* characterized by extremely low female sexual receptivity. J Neurogenet 23:329-340.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. Genome Res 20:1313-1326.

Katzman S, Kern AD, Bejerano G, Fewell G, Fulton L, Wilson RK, Salama SR, Haussler D. 2007. Human genome ultraconserved elements are ultraselected. Science 317:915.

Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, Blaxter ML. 2009. Analysis of the genome sequences of three *Drosophila melanogaster* spontaneous mutation accumulation lines. Genome Res 19:1195-1201.

Kelleher ES, Markow TA. 2009. Duplication, selection and gene conversion in a *Drosophila mojavensis* female reproductive protein family. Genetics 181:1451-1465.

Kleene KC. 2005. Sexual selection, genetic conflict, selfish genes, and the atypical patterns of gene expression in spermatogenic cells. Dev Biol 277:16-26.

Kleene KC. 2001. A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells. Mech Dev 106:3-23.

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

Kreitman M. 1983. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature 304:412-417.

Langley CH, Voelker RA, Brown AJ, Ohnishi S, Dickson B, Montgomery E. 1981. Null allele frequencies at allozyme loci in natural populations of *Drosophila melanogaster*. Genetics 99:151-156.

Larracuente AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, Sturgill D, Zhang Y, Oliver B, Clark AG. 2008. Evolution of protein-coding genes in Drosophila. Trends Genet 24:114-123.

Lazzaro BP. 2005. Elevated polymorphism and divergence in the class C scavenger receptors of Drosophila melanogaster and D. simulans. Genetics 169:2023-2034.

Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. 2005. Evolution of proteins and gene expression levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. Mol Biol Evol 22:1345-1354.

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

Li C-Y, Zhang Y, Wang Z et al. 2010a. A human-specific *de novo* protein-coding gene associated with human brain functions. PLoS Comput Biol 6:e1000734.

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

Li WH, Gojobori T, Nei M. 1981. Pseudogenes as a paradigm of neutral evolution. Nature 292:237-239.

Librado P, Rozas J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451-1452.

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

Madigan SJ, Edeen P, Esnayra J, McKeown M. 1996. att, a target for regulation by *tra2* in the testes of *Drosophila melanogaster*, encodes alternative RNAs and alternative proteins. Mol Cell Biol 16:4222-4230.

Marais G, Domazet-Loso T, Tautz D, Charlesworth B. 2004. Correlated evolution of synonymous and nonsynonymous sites in Drosophila. J Mol Evol 59:771-779.

Marin I, Siegal ML, Baker BS. 2000. The evolution of dosage-compensation mechanisms. Bioessays. 22(12):1106-14.

Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C. 2011. Protein 3D Structure Computed from Evolutionary Sequence Variation. PLoS One 6:e28766.

Matsuo T, Sugaya S, Yasukawa J, Aigaki T, Fuyama Y. 2007. Odorant-binding proteins OBP57d and OBP57e affect taste perception and host-plant preference in *Drosophila sechellia*. Plos Biology 5:985-996.

McBride CS. 2007. Rapid evolution of smell and taste receptor genes during host specialization in *Drosophila sechellia*. Proc Natl Acad Sci U S A 104:4996-5001.

McBride CS, Arguello JR, O'Meara BC. 2007. Five Drosophila genomes reveal nonneutral evolution and the signature of host specialization in the chemoreceptor superfamily. Genetics 177:1395-1416.

McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in Drosophila. Nature 351:652-654.

McGraw LA, Gibson G, Clark AG, Wolfner MF. 2004. Genes regulated by mating, sperm, or seminal proteins in mated female *Drosophila melanogaster*. Curr Biol 14:1509-1514.

Murali T, Pacifico S, Yu J, Guest S, Roberts GG, 3rd, Finley RL, Jr. 2011. DroID 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for Drosophila. Nucleic Acids Res 39:D736-743.

Nagy E, Maquat LE. 1998. A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. Trends Biochem Sci 23:198-199.

Obbard DJ, Welch JJ, Kim K-W, Jiggins FM. 2009. Quantifying adaptive evolution in the Drosophila immune system. PLoS Genet 5:e1000698.

Ohno S. 1970. Evolution by gene duplication. Berlin (Germany): Springer-Verlag

Ohno S, Wolf U, Atkin NB. 1968. Evolution from fish to mammals by gene duplication. Hereditas 59:169-187.

Parra G, Reymond A, Dabbouseh N, Dermitzakis ET, Castelo R, Thomson TM, Antonarakis SE, Guigó R. 2006. Tandem chimerism as a means to increase protein complexity in the human genome. Genome Res 16:37-44.

Peden J. 1999. Analysis of codon usage. PhD Thesis, University of Nottingham, UK.

Ponting CP. 2008. The functional repertoires of metazoan genomes. Nat Rev Genet 9:689-698.

Ponting CP, Oliver PL, Reik W. 2009. Evolution and functions of long non-coding RNAs. Cell 136:629-641.

Prud'homme B, Gompel N, Carroll SB. 2007. Emerging principles of regulatory evolution. Proc Natl Acad Sci U S A 104 Suppl 1:8605-8612.

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R. 2005. InterProScan: protein domains identifier. Nucleic Acids Res 33:W116-120.

Rand DM, Kann LM. 1996. Excess amino acid polymorphism in mitochondrial DNA: contrasts among genes from Drosophila, mice, and humans. Mol Biol Evol 13:735-748.

RDevelopmentCoreTeam. 2009. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

RDevelopmentCoreTeam. 2010. R: A Language and environment for statistical computing.

Retelska D, Iseli C, Bucher P, Jongeneel CV, Naef F. 2006. Similarities and differences of polyadenylation signals in human and fly. BMC Genomics 7:176.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276-277.

Richards S, Liu Y, Bettencourt BR et al. 2005. Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. Genome Res 15:1-18.

Rosenfeld PJ, Cowley GS, McGee TL, Sandberg MA, Berson EL, Dryja TP. 1992. A null mutation in the rhodopsin gene causes rod photoreceptor dysfunction and autosomal recessive retinitis pigmentosa. Nat Genet 1:209-213.

Rymarquis LA, Kastenmayer JP, Huttenhofer AG, Green PJ. 2008. Diamonds in the rough: mRNA-like non-coding RNAs. Trends Plant Sci 13:329-334.

Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG. 2007. Dynamic evolution of the innate immune system in Drosophila. Nat Genet 39:1461-1468.

Sandler L, Lindsley DL, Nicoletti B, Trippa G. 1968. Mutants affecting meiosis in natural populations of *Drosophila melanogaster*. Genetics 60:525-558.

Schmid KJ, Aquadro CF. 2001. The evolutionary analysis of "orphans" from the Drosophila genome identifies rapidly diverging and incorrectly annotated genes. Genetics 159:589-598.

Schully SD, Hellberg ME. 2006. Positive selection on nucleotide substitutions and indels in accessory gland proteins of the *Drosophila pseudoobscura* subgroup. J Mol Evol 62:793-802.

Sherry ST, Ward M, Sirotkin K. 1999. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res 9:677-679.

Spencer WP. 1947. Mutations in wild populations in Drosophila. Adv Genet 1:359-402.

Stephan W, Li H. 2007. The recent demographic and adaptive history of *Drosophila melanogaster*. Heredity (Edinb) 98:65-68.

Stoletzki N, Eyre-Walker A. 2011. Estimation of the neutrality index. Mol Biol Evol 28:63-70.

Subramanian S, Kumar S. 2004. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics 168:373-381.

Sun S, Ting CT, Wu CI. 2004. The normal function of a speciation gene, Odysseus, and its hybrid sterility effect. Science 305:81-83.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595.

Takemori N, Yamamoto MT. 2009. Proteome mapping of the *Drosophila melanogaster* male reproductive system. Proteomics 9:2484-2493.

Tamura K, Subramanian S, Kumar S. 2004. Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. Mol Biol Evol 21:36-44.

Tautz D, Domazet-Loso T. 2011. The evolutionary origin of orphan genes. Nat Rev Genet 12:692-702.

Temple BR, Jones CD, Jones AM. 2010. Evolution of a signaling nexus constrained by protein interfaces and conformational States. PLoS Comput Biol 6:e1000962.

Thompson JD, Gibson TJ, Higgins DG. 2002. Multiple sequence alignment using ClustalW and ClustalX. Curr Protoc Bioinformatics Chapter 2:Unit 2.3.

Timofeef-Ressovsky N. 1930. Das Genovariieren in verschiedenen Richtungen bei *Drosophila melanogaster* unter dem Einfluss der Rontgenbestrahlung. Naturwiss:434-437.

Timofeeff-Ressovsky H, Timofeeff-Ressovsky, NW. 1927. Genetische analyse einer freilebenden Drosophila melanogaster population. Arch. Entw. Mech. Org.:70-109.

Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, Albà MM. 2009. Origin of primate orphan genes: a comparative genomics approach. Mol Biol Evol 26:603-612.

Tsaur SC, Ting CT, Wu CI. 1998. Positive selection driving the evolution of a gene of male reproduction, *Acp26Aa*, of Drosophila: II. Divergence versus polymorphism. Mol Biol Evol 15:1040-1046.

Turner LM, Hoekstra HE. 2006. Adaptive evolution of fertilization proteins within a genus: variation in ZP2 and ZP3 in deer mice (Peromyscus). Mol Biol Evol 23:1656-1669.

Tweedie S, Ashburner M, Falls K et al. 2009. FlyBase: enhancing Drosophila Gene Ontology annotations. Nucleic Acids Res 37:D555-559.

Ulveling D, Francastel C, Hube F. 2011. When one is better than two: RNA with dual functions. Biochimie 93:633-644.

Van Valen L. 1973. A new evolutionary law. Evolutionary theory.

Vasudevan S, Peltz SW, Wilusz CJ. 2002. Non-stop decay--a new mRNA surveillance pathway. Bioessays 24:785-788.

Voelker RA, Schaffer HE, Mukai T. 1980. Spontaneous Allozyme Mutations in *DROSOPHILA MELANOGASTER*: Rate of Occurrence and Nature of the Mutants. Genetics 94:961-968.

Wagstaff BJ, Begun DJ. 2005a. Comparative genomics of accessory gland protein genes in *Drosophila melanogaster* and *D. pseudoobscura*. Mol Biol Evol 22:818-832.

Wagstaff BJ, Begun DJ. 2007. Adaptive evolution of recently duplicated accessory gland protein genes in desert Drosophila. Genetics 177:1023-1030.

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

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7:256-276.

Weiner J, 3rd, Beaussart F, Bornberg-Bauer E. 2006. Domain deletions and substitutions in the modular protein evolution. Febs J 273:2037-2047.

White-Cooper H, Bausek N. 2010. Evolution and spermatogenesis. Philos Trans R Soc Lond, B, Biol Sci 365:1465-1480.

Wilson GA, Bertrand N, Patel Y, Hughes JB, Feil EJ, Field D. 2005. Orphans as taxonomically restricted and ecologically important genes. Microbiology (Reading, Engl) 151:2499-2501.

Wong A, Albright SN, Wolfner MF. 2006. Evidence for structural constraint on *ovulin*, a rapidly evolving Drosophila melanogaster seminal protein. Proc Natl Acad Sci USA 103:18644-18649.

Wong A, Christopher AB, Buehner NA, Wolfner MF. 2010. Immortal coils: conserved dimerization motifs of the Drosophila ovulation prohormone *ovulin*. Insect Biochem Mol Biol 40:303-310.

Wray GA. 2007. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 8:206-216.

Wu DD, Irwin DM, Zhang YP. 2011. *De novo* origin of human protein-coding genes. PLoS Genet 7:e1002379.

Xiao W, Liu H, Li Y, Li X, Xu C, Long M, Wang S. 2009. A rice gene of *de novo* origin negatively regulates pathogen-induced defense response. PLoS ONE 4:e4603.

Yamaguchi-Kabata Y, Shimada MK, Hayakawa Y, Minoshima S, Chakraborty R, Gojobori T, Imanishi T. 2008. Distribution and effects of nonsense polymorphisms in human genes. PLoS One 3:e3393.

Yandell M, Mungall CJ, Smith C, Prochnik S, Kaminker J, Hartzell G, Lewis S, Rubin GM. 2006. Large-scale trends in the evolution of gene structures within 11 animal genomes. PLoS Comput Biol 2:e15.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555-556.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-1591.

Yang Z, Huang J. 2011. *De novo* origin of new genes with introns in Plasmodium *vivax. FEBS Lett 585:641-644.*

Yeh SD, Do T, Chan C et al. 2012. Functional evidence that a recently evolved Drosophila sperm-specific gene boosts sperm competition. Proc Natl Acad Sci U S A 109:2043-2048.

Yngvadottir B, Xue Y, Searle S, Hunt S, Delgado M, Morrison J, Whittaker P, Deloukas P, Tyler-Smith C. 2009. A genome-wide survey of the prevalence and evolutionary forces acting on human nonsense SNPs. Am J Hum Genet 84:224-234.

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

Zhao J, Klyne G, Benson E, Gudmannsdottir E, White-Cooper H, Shotton D. 2010. FlyTED: the Drosophila Testis Gene Expression Database. Nucleic Acids Res 38:D710-715.

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