THE UNIVERSITY OF CHICAGO

THE EVOLUTION AND FITNESS EFFECTS OF SPECIES-SPECIFIC GENES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

COMMITTEE ON GENETICS

BY

NICHOLAS WILLIAM VANKUREN

CHICAGO, ILLINOIS

DECEMBER 2016

To Sydney

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am extremely grateful for all of the people who provided scientific and personal support over the last 6 years. Science is collaboration, and no project can be completed in isolation.

I am deeply indebted to Maria D. Vibranovski, who gave me opportunities to quickly become involved in her Long lab projects. She is a patient and encouraging mentor who not only helped me get on my feet in the lab, but has continued working with me throughout my entire graduate study. Her mentorship and friendship has been invaluable, and I look forward to continuing to work with her over the coming years. I thank Manyuan for his guidance and unquestioning support and encouragement throughout the last 6 years. I am especially grateful for his emphasis on telling *complete* stories that contribute to "big pictures" in science, and advance a general concept rather than a specific idea. This is a truly important mindset for doing meaningful and impactful scientific research, and I thank him for helping me change the way I think about science.

I thank Marty Kreitman, Marcus Kronforst, Chip Ferguson, and Dick Hudson for their guidance throughout and for being some of the nicest people I have gotten to work and talk with. I wish I had taken a bit more advantage of having such fantastic scientists mentoring me. I particularly thank Dick for his help and guidance with some of the population genetic analyses in Chapter 4, and Chip for being a fantastic role model for me and many other young scientists in the Molecular Biosciences cluster. I also thank Robert Ho and Joy Bergelson for allowing me to join their "Professional Grants and Pubs" course that taught me to really critically think, write, and evaluate writing. Their mentorship and the other members of that course strongly shaped how I approached my graduate work.

Past and current members of the Long Lab have been fantastic lab mates. I am grateful for discussions with Ben Krinsky, Patrick Landback, Iuri Mateuzzo-Ventura, Alex Advani, UnJin Lee, Qian Yang, Sidi Chen, and Haiwang Yang. I particularly thank Grace Lee for her guidance on population genomics, Andrea Gschwend for keeping the lab social, and Claus Kemkemer and Li Zhang for advice and discussion. I am grateful to have had a large group of friends in the Molecular Biosciences and Darwinian clusters that are as excited, frustrated, and in love with biology as me. I thank them all for endless inane, but extremely stimulating discussions.

I have had such an excellent group of administrators in both the Molecular Biosciences and Ecology and Evolution programs. They make everything run smoothly and I feel lucky to have never had to worry about where my next stipend check is coming from or how I will pay for that lab supply. Thanks especially to Jeff Wisniewski and Bonnie Brown in the E&E front office, Ian Miller and Mike Guerra, and particularly Sue Levison. She does many things for her MolBio students, from organizing social events to making sure we are actually going to leave GGSB at some point, to being there for a chat anytime you need. In the same vein, I thank Lucia Rothman-Denes for her administration of the NIH Genetics and Regulation Training grant. I also thank Barry Aprison for including me in so many outreach events, and for helping me get a NSF GRF.

I acknowledge my funding sources, including an NIH Genetics and Regulation Training Grant T32GM007197, an NSF Graduate Research Fellowship, and NIH Grant R01GM116113.

Obviously, I could not have even made it to graduate school without the support from my mom, dad, and sister. I appreciate all of their help and encouragement throughout my last decade in school. I thank them and my wife's parents for bearing with us being hundreds or thousands of miles away from any of them, and appreciate all their continuing support.

Above all, I am immensely grateful for my wife, Sydney, and her unwavering support over the last ten years. She never stopped supporting or feeding me, even through some late nights and a few angry scowls, and I look forward to all of our future adventures together.

ABSTRACT

Biologists have strived to understand the origins of morphological and behavioral differences between individuals since long before Darwin's 1859 "Abstract" to his Big Book. While Darwin's work made it abundantly clear that huge amounts of variation exist within and between species, the mechanism by which this variation was produced and passed on from generation to generation was not clear. We now know that DNA is the hereditary material. Furthermore, the complete DNA sequences from an immense range of organisms has revealed an enormous amount of DNA variation that could cause the morphological and behavioral differences Darwin highlighted in his books. Which of this genetic variation causes phenotypic variation? And what roles do natural or sexual selection play in its evolution? This thesis explores the role that evolutionarily new genes, the functional units of DNA, play in shaping the evolution of fitness and the genome through an investigation of genes found in a single species of fruit fly. I use precise genetic manipulation to show that at least 27% of *Drosophila melanogaster*-specific genes, while young, have essential roles in fly development and reproduction. Furthermore, I use empirical population genetics analyses to show that species-specific genes are frequently strongly selected, and, combined with my functional data, suggest that new genes are likely primarily selected for their male-beneficial functions. Altogether, this work shows that the genes underlying important processes such as development and reproduction can rapidly change and that this process is strongly influenced by selection.

CHAPTER 1

INTRODUCTION

1.0.1 Natural Selection Requires A Mechanism of Inheritance

Darwin's theory of evolution by natural selection, developed over nearly 30 years and finally published in 1859, is based on three simple facts: 1) there is no such thing as an unlimited resource; 2) individuals differ from one another; and 3) individuals inherit traits from their parents [1, 2]. Individuals that are better able to collect and utilize their limited resources will tend to produce more offspring and pass on the traits that make them so successful, resulting in morphological, behavioral, and other phenotypic changes in a lineage over many generations. Natural selection is thus "daily and hourly scrutinizing, throughout the world, every variation, even the slightest; rejecting that which is bad, preserving and adding up all that is good; silently and insensibly working, whenever and wherever opportunity offers, at the improvement of each organic being in relation to its organic and inorganic conditions of life" [1]. While he provided ample evidence that there is indeed competition between individuals for limited resources and that there are vast amounts of variation between individuals and species, the mechanism by which offspring inherited traits from their parents was not known to Darwin and his Victorian contemporaries [1, 3, 4].

1.0.2 From Naturalists to Geneticists

The mechanism by which traits are inherited was, ironically, described by Gregor Mendel in 1866, six years before Darwin's final edition *Origin*. But Mendel's laws did not become widely known for another 34 years, long after the deaths of Mendel and Darwin. However, the application of Mendel's laws of segregation and independent assortment to a wide variety of problems in heredity spurred rapid experimental and theoretical advances in understanding the physical causes of inheritance and the effects of natural selection on trait changes over time [5].

T. H. Morgan's 1910 discovery of a single male fruit fly with white eyes instead of the normal red eyes not only convince him that Mendelism and Darwinism were consistent (which he and many others had denied up until then), it prompted a series of experiments that precisely dissected the phenotypic effects of pieces of chromosomes [6, 7]. This impressively careful and patient work throughout the early 20^{th} century definitively showed that chromosomes, and particularly specific pieces of chromosomes, carried the information needed to produce fully functioning organisms [6, 8–11]. With knowledge of the precise mechanism of inheritance, precise mathematical models were developed to definitively show that natural selection could work by altering the frequencies of mutant chromosomes in populations over time [12–14]. Thus, by the end of the 1940s this "Modern Synthesis" had definitively linked Darwin's natural selection and Mendel's laws, experimentally and theoretically establishing Darwin's third fact [5].

1.0.3 Chromosomes to Genes to Genomes

Work throughout the mid-1900s showed clearly that chromosomes are comprised of DNA, and that DNA is a sequence of hundreds to billions of nucleotides. The pieces of DNA (loci) that control traits are generally defined as genes, and genes are considered to be the functional units of an organism's genome (the organism's full complement of DNA and the genes it contains). Genes are transcribed into RNAs that are translated into the functional units of the cell, proteins (i.e. genes are expressed). Genes are expressed, and therefore function, at different times during an organism's development, in different cell types, and in different environmental conditions. The times and places in which a gene is expressed and functions are regulated by other proteins and RNAs in the cell. Work over just the last 20 years has clearly shown that no gene functions in isolation, and that it is the set of genes active in a particular time or place, and the interactions between them and DNA that control cellular processes [15]. Amazingly, this fact was apparent to Wright (1930), who specifically said that "... it may safely be assumed that there are always important epistatic effects [between genes]. Genes favorable in one combination, are, for example, extremely likely to be unfavorable in another." [13].

The sequencing of entire genomes from a huge variety of organisms has revealed an astounding array of genetic diversity both within and between species. At this time, the central repository for genome sequence data contains 3,707 complete eukaryotic genomes, ranging in size from 660,000 nucleotides in the clam *Corbicula fluminea* to 27.6 billion nucleotides in the tree *Pinus lambertiana* (NCBI Genome, accessed 5 October, 2016). Beyond differences in size, comparisons between genomes from closely related organisms revealed that a massive number of single nucleotide differences, small DNA insertions and deletions, fusions and splits of chromosomes, and duplications and deletions of large regions of the genome are common between species and even between individuals of the same species. Darwin's second fact has now clearly been demonstrated to be true at the genomic level, too.

1.0.4 Each Genome Contains a Unique Set of Genes

Analysis of whole genome sequences has clearly shown that the number of genes each genome contains is often very different [16]. For example, while the soybean (*Glycine max*) genome contains more than 115,000 protein-coding genes, the genome of the bacterium *Candidatus Tremblya princeps* contains only 155 genes (NCBI Genome, accessed 5 October, 2016). Furthermore, even genomes with similar numbers of genes can have very different sets of genes [17]. These observations show that there must exist a widespread process by which genes are gained and lost from genomes over evolutionary time. These observations also raise important questions about gene gain and loss and its contribution to evolution. In particular, what molecular mechanisms and evolutionary forces govern gene gain and loss? And what is the benefit of adding new genes to genomes?

1.0.5 The Process of New Gene Origination

New genes are those that were formed recently in evolutionary time and are therefore present in one group of closely-related organisms but absent in all others [16]. The process by which new genes originate consists of several overlapping phases. First, a mutation occurs in a single germ cell's DNA to form a new locus. The new locus then spreads through the population under the influence of evolutionary forces such as (natural or sexual) selection and genetic drift until it is fixed. Concurrent with and following its formation and fixation, the new locus can accumulate mutations that alter its sequence, structure, and/or expression patterns that cause it to contribute positively to the organism's fitness and to be maintained in the genome by selection [16].

There are thus two aspects to new gene origination: evolution of the *locus* and evolution of the *function*. The evolution of the new locus can be split into two phases that are delineated by fixation. Before fixation, the new locus is a polymorphism that is subject to the effects of evolutionary forces such as selection, drift, recombination, and mutation, and its journey to fixation can simplistically be described using classic population genetic models. Overlaid on this population genetic process is the evolution of the gene's function. A gene's function(s) include its expression pattern, biochemical function, and its interactions with other genes (which in turn dictates the functions of those genes). Mutations that cause new genes to become favored may accumulate at any point in the new locus' evolution, from mutation to fixation and beyond.

1.0.6 Models of New Gene Origination

Perhaps unsurprisingly, the structures, functions, and evolution of new genes have attracted the interests of pioneers in genetics and evolution since the early 20^{th} century. Sturtevant (1925) was one of the first to identify a gene formed by duplication of a piece of chromosome, the *Bar* duplication in the fruit fly*Drosophila melanogaster*, from which Muller developed the first prevalent model of new gene evolution [8, 11]. Muller predicted that a new duplicate gene could mutate, acquire a novel beneficial function, and be preserved in the genome. He further claimed that "there remains no reason to doubt the application of the dictum 'all life from pre-existing life' and 'every cell from a pre-existing cell' to the gene: 'every gene from a pre-existing gene"' [11]. Ohno further developed and argued for Muller's model, and emphasized the fact that the vast majority of duplicate genes will be inactivated by mutations and lost from the genome by additional mutations [18, 19]. In Muller's and Ohno's formulations, duplication produces two functionally redundant gene copies. As long as one copy retains the ancestral, important function, the other copy is free to accumulate mutations that confer on it a novel, beneficial function (neofunctionalization) [18].

Our increased understanding of duplication mechanisms and, especially, the complexity of gene expression regulation suggests that gene duplicates are probably rarely functionally redundant. To explain the high rates of duplicate gene retention after duplication (particularly whole genome duplication), models were developed that posit that selection acts on new genes at all stages of their origination [20, 21]. The adaptive radiation (AR) model predicts that duplication of a gene is beneficial because it increases gene expression levels (dosage). Repeated duplication results in many duplicate copies that are then free to accumulate mutations and neofunctionalize [20]. Piatigorsky and Wistow (1991) and Hughes (1994) stimulated development of a series of models assuming that new genes originate from existing genes with multiple functions [22, 23]. The Innovation-Amplification-Divergence (IAD) model predicts that a multifunctional gene's low-level secondary function may become beneficial with environmental change [21]. Repeated amplification increases dosage of the beneficial function, and different duplicates can then optimize the ancestral or novel functions. Similarly, escape from adaptive conflict (EAC) posits that duplication may be favored for a gene with two functions that cannot simultaneously be optimized by selection [24–26]. Thus, duplication allows partitioning of functions between the two copies and independent optimization. In each of these cases, natural or sexual selection is assumed to be operating on each stage of new gene origination.

Many clear examples of new genes that have evolved according to each of these neofunctionalization and subfunctionalization models have been published and contribute to new gene origination (see refs. [16, 27] for reviews).

1.0.7 New Gene Formation Mechanisms and Patterns

Classic studies of new genes focused on the contribution of DNA-based (copy-and-paste) duplication of genes, but it is now clear that a wide variety of mechanisms frequently generate new loci, including retrotransposition of mRNAs [28, 29], fusion or splitting of existing genes [30, 31], and even *de novo* gene origination from non-coding DNA [32] (see references [16, 33, 34] for reviews). The predominant mechanism of new locus formation varies between organisms. For example, while tandem duplication is most prominent in *Drosophila* [35], dispersed DNA-based duplication is most common in primates [36]. However, each mechanism has been experimentally demonstrated to contribute to genome evolution across the tree of life, and comparative genomic studies have estimated that new DNA- and RNA-based duplicates arise at high rates [17, 37–41]. For example at least 15 genes per million years are gained by Drosophila genomes [35, 42, 43], while ∼30 genes per million years are gained by mammalian genomes [44]. Thus, new genes of many different types originate at appreciable rates in all taxa studied to date.

Two general characteristics of new genes have emerged from such large-scale studies. First, new genes are frequently expressed more highly in males than in females (i.e. they have male-biased expression) [43, 45]. Second, new genes are non-randomly distributed in the genomes of flies, silkworms, mice, and humans [43, 46–49]. New genes tend to be located on autosomes, but to have been formed via duplication of sex chromosome-linked genes (X or Z) [28, 46–48]. That is, new genes tend to move X→A. While the exact cause(s) of X→A movement are still debated, the consequence of this gene traffic is a dearth of male-biased genes on the X chromosome in worms, flies, and mice [50–53]. Thus, the process of new gene origination has also shaped the overall distribution of genes among chromosomes.

1.0.8 New Gene Phenotypes

What is the advantage to adding a new gene to a genome? As genes age, they accumulate mutations that obscure the structural or evolutionary signals from their early history [33, 54]. Thus, the advantage to studying genes at the early stages of their evolution is that the sequence and expression differences that initially cause new genes to become important components of the genome are not likely to have been obscured by continued mutation accumulation [54]. This idea and the ability to assign ages to genes through comparative genomics allowed rigorous experimental analyses of new gene functions to begin in the 1990s.

The last two decades have produced an enormous number of studies of the phenotypic effects of new genes (for reviews, see refs [27, 55–57]). These studies have shown that new genes can quickly become involved in both conserved and novel processes and that new gene origination is frequently driven by selection. Arctic fish, for example, have evolved novel antifreeze proteins through duplication and neofunctionalization [58–60]. *Sphinx* is a novel chimeric gene found specifically in *D. melanogaster* that prevents male-male courtship [61]. Distinct regulatory regions were partitioned between *Gal3* and *Gal1*, yeast duplicate genes, through an escape from adaptive conflict to produce fine-scale control over induction of the galactose utilization pathway [24]. Ding and her colleagues used detailed experiments to show that a ∼10 million year old gene in *D. melanogaster* and its closest relatives rapidly evolved a unique and essential role in spermatogenesis through novel expression and subcellular localization [62].

Altogether, just these few studies highlight the fact that new genes, even those formed by complete gene duplications, can quickly produce novel phenotypes and participate in pathways and networks controlling critical cellular processes. Furthermore, selection played a large role in each of the above cases. Finally, these studies highlight the observation in metazoans that new genes disproportionately affect males.

1.0.9 New, Yet Essential Genes

Recently, Chen and his colleagues showed that constitutive knockdown of the expression of new genes in Drosophila, less than ∼35 million years old, frequently caused flies to fail to complete development [63]. That is, ∼30% of the new genes they studied were essential for fly development. Furthermore, they found that the essentiality of a new gene appears to be independent of its parent [63].

These results raised two important questions. First, how, molecularly, did new genes become essential for fly development? Developmental processes are assumed to be ancient and highly conserved, so it was unclear how new genes could be so quickly and tightly integrated into the gene networks that control these processes [15, 63]. Second, what are the evolutionary forces that drove this process, and what model(s) of new gene evolution can explain it? Like many studies of young genes, Chen et al. (2010) found that young essential genes exhibited elevated rates of amino acid substitutions [63]. But because the genes they studied were relatively old (i.e. > 3 million years), it is not clear if these protein sequence changes or other layers of divergence (e.g. expression or structural) may have initially caused new genes to become essential.

This thesis explores these questions through direct and indirect tests of the fitness effects of genes at the extreme earliest stages of their evolution: species-specific genes. I sought to answer two main questions using genes that were formed specifically in *Drosophila melanogaster*:

1. What are the phenotypic effects of *D. melanogaster*-specific genes?

2. What are the evolutionary forces that governed and continue to govern *D. melanogaster*-specific gene evolution?

I will use the term 'essential' frequently in this thesis. To be very clear, throughout this work I consider a gene to be essential if the loss of that gene's function causes a statistically significant reduction in fly survival and/or fertility. Arguments for and against this definition can certainly be made. For example, should a gene whose knockdown causes a 20%, but significant, reduction in fly survival be defined as essential gene? I believe the answer is yes because a variant with a selection coefficient of -0.2 would quickly be purged from a population and that, in evolutionary terms, is essential for fitness.

The remainder of this work describes my experimental and computational investigations of the answers to these questions. Chapter 2 describes a broad-scale screen of *D. melanogaster*-specific gene fitness effects and evolution. Chapter 3 discusses a more detailed investigation of one *D. melanogaster*-specific gene pair. Chapter 4 switches to an investigation of new genes segregating within *D. melanogaster* populations. Finally, Chapter 5 briefly summarizes my results, and some potential future directions.

CHAPTER 2

DROSOPHILA MELANOGASTER-SPECIFIC GENES CAN RAPIDLY BECOME ESSENTIAL FOR FLY DEVELOPMENT

2.0.1 Abstract

Essential genes are those whose functions are required for an organism to properly develop and successfully reproduce. Such genes are thought to be the keystones in the biological networks that produce fit organisms, and the prevailing view is that only genes found in many species, which have been conserved over long periods of evolutionary time by purifying selection, can be essential. Yet recent studies clearly show that new genes can become essential within just several millions of years. Here we investigate the essentiality of species-specific duplicate genes, the youngest class of new genes, to understand how new genes quickly become essential. We find 27% of *Drosophila melanogaster*-specific gene duplicates we tested are essential for fly development, while their parent copies are not, and that divergence between duplicates' expression patterns across development likely caused new duplicates to become essential. Furthermore, selection had a large direct or indirect role in this process, as 62% of *D. melanogaster*-specific duplications are found in recent, strong selective sweeps. Our results show that species-specific genes can quickly become essential, highlighting the fact that a gene's age or protein sequence conservation is not necessarily indicative of its importance in conserved processes such as development and reproduction.

2.0.2 Introduction

New genes are genes found in one group of closely-related organisms but absent in all others, and are common features of all species' genomes [16]. Recently, detailed molecular and phenotypic studies of new genes have shown that they can be critical components of the pathways and gene-gene interaction networks controlling novel phenotypes, sometimes within just several million years [27, 62–68]. *Eud-1*, for example, is a gene found specifically in the nematode *Pristionchus pacificus* and its sister species that promotes development of a novel mouth form (and therefore novel feeding strategy) in *P. pacificus* [69]. *Umbrea*, in contrast, was formed ∼15 million years ago in Drosophila, but has since evolved a novel, yet essential function promoting proper chromosome segregation during cell division through binding to centromeres in a species-specific fashion [63, 70].

The existence of young essential genes like *Umbrea* poses a practical problem because detailed studies of gene functions tend to focus on genes that are highly conserved among a broad range of taxa. A gene's function experimentally determined in a model system is then extrapolated to be that gene's function in a non-model system, like humans. This practice has produced the prevailing view that only ancient genes, shared by many taxa and maintained by selection over long periods of time, have essential functions, and has also led to a dearth of studies on and resources for studying new genes [45]. Certainly, influential authors and classroom textbooks, such as Jacob [71] and Lewin's *Genes*, state that "highly conserved genes tend to have more basic functions" [15] and that "functionally essential genes are not organism specific, nor are their functions protected by gene duplication" [72]. However, new essential genes do exist and it is therefore crucial to understand how they became essential and whether their essential functions are novel like *Umbrea*'s.

Most new genes are formed by duplication and are usually assumed to be functionally redundant in the early stages of their evolution [18, 73]. In contrast to this classic assumption, recent models make it clear that there are a broad range of conditions under which new duplicates may immediately or quickly become beneficial to their hosts via neofunctionalization [74, 75] or selection-driven partitioning of ancestral functions between duplicate copies [23, 24, 55]. Thus, a new gene may become essential either by gaining a novel essential function or retaining an ancestral essential function after duplication [21, 23, 75, 76]. New genes frequently exhibit signs of rapid amino acid sequence and expression pattern evolution, indicative of the role of selection in the early stages of new gene evolution and potentially rapid acquisition of beneficial, possibly essential, functions [35, 63, 77].

In addition to examples like *Umbrea* and *eud-1*, Chen et al. (2010) showed that new genes in Drosophila could become essential for fly development in ∼35 - 6 million years [63]. However, it remained unknown if and how often the youngest new genes, those found in a single species, are essential. This knowledge is critical, though, because the mutations and signatures of the evolutionary forces that initially caused these new genes to become essential are unlikely to have been obscured by additional mutation accumulation [54].

Here we investigate the essentiality and evolution of species-specific duplicate genes in *D. melanogaster*, which are therefore less than ∼2 million years old, to understand how new genes quickly become essential. We find that at least 27% of *D. melanogaster*-specific genes are essential for fly development using available mutant lines and RNA interference (RNAi). Essential *D. melanogaster*-specific gene expression patterns quickly diverged from their parent copies' and drove divergence of their gene-gene interaction network positions. Furthermore, 62% of *D. melanogaster*-specific duplications reside in recent and strong selective sweeps, suggesting that selection plays a large role in these earliest stages of new gene evolution. Our results clearly show that species-specific duplicate genes can rapidly become essential components of the networks controlling development, and they have significant implications for understanding the evolution of developmental processes and the genes and gene networks controlling them. In addition our findings highlight the fact that, while conservation certainly implies functional importance, functional importance does not necessarily require conservation.

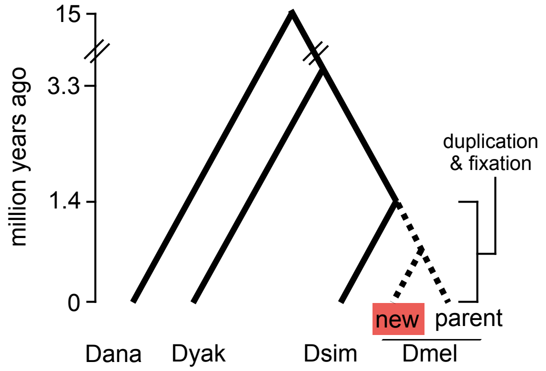


Figure 2.1: *Drosophila melanogaster*-specific genes are less than ∼2 million years old. A hypothetical gene genealogy depicting the relationship between a *D. melanogaster*-specific (new) gene and its homologs. The new copy is defined as the copy with the greatest amino acid divergence from the orthologous protein in *D. simulans* (Dsim) and *D. yakuba* (Dyak). Dashed branches show uncertainty in precisely determining when duplication and fixation occurred. Divergence time estimates are taken from [78]. Dana: *D. ananassae*; Dmel: *D. melanogaster*.

2.0.3 Results

D. melanogaster-Specific Duplicate Genes Can Be Essential for Fly Development

We compiled a high-confidence set of 17 *D. melanogaster*-specific protein-coding genes and their parents by filtering candidate genes from previous studies [35, 43, 65]. These genes were formed by 13 duplications since *D. melanogaster* and *D. simulans* diverged (Figure 2.1; Table 2.1; Table A.1) [78]. Most (11/13) duplications are tandem, and one of the two adjacent copies cannot be defined as ‘new’ in terms of age. However, consistent with

Table 2.1: *D. melanogaster*-specific genes and their parents.

New gene	K^{e}_{new}	**Parent gene(s)**	K^{e}_{parent}	**Type**	**Chrom.**[f]	**Freq.**[g]
CG31958[a]	0.0652	*CG31960*	0.0351	tandem	2L	0.89
Ada1-1	0.0167	*Ada1-2*	0.0126	tandem	2L	1
CG18789	0.0417	*CG18787*	0.0405	tandem	2L	1
Qtzl[b,c]	-	*CG12264/escl*	-/-	tandem	2L	1
ProtB[a]	0.0351	*ProtA*	0.0357	tandem	2L	1
CG31683	0.0156	*CG18858*	0.0093	tandem	2L	0.94
CG31687[b]	-	*Cdc23/CG31688*	-/-	tandem	2L	0.94
RpS15Ab	0.0254	*RpS15Aa*	0.0000	RNA	2R←X	1
CG33470	0.0226	*IMPPP*	0.0226	tandem	2R	0.89
CG30059	0.0154	*CG18278*	0.0154	tandem	2R	0.89
CG32165[a]	0.0397	*CG32164*	0.0302	tandem	3L	1
CG12592[b]	-	*CG18545/sle*	-/-	tandem	3R	1
CG11659	0.0172	*CG6300*	0.0173	tandem	3R	1
tHMG1[a]	0.2462	*tHMG2*	0.1176	tandem	3R	1
CG11700[a,d]	0.1037	*Ubi-p5E*	0.0000	tandem	X	1
CG32588[a]	0.3617	*CG33252*	0.2245	dispersed DNA	X←X	1
CG9123[a]	0.0906	*CG12608*	0.0440	tandem	X	1

a: Greater gene structure divergence than parent relative to outgroups - see Figure A.1 for detail.

b: Chimeric gene, formed by partial duplication and fusion of two nearby genes.

c: Studied in detail by [81]

d: Studied in detail by [82]

e: Polarized amino acid divergence from *D. yakuba* and *D. simulans* orthologs.

f: New gene chromosome ← parent gene chromosome.

g: Frequency in 17 re-sequenced fly genomes from the Drosophila Population Genomics Project Phase 2 [83].

Note: *Ada1-2*, *CG18789*, and *Qtzl* were formed by one tandem duplication; *CG33470* and *CG30059* were formed by one tandem duplication; *CG31687* and *CG31683*were formed by one tandem duplication.

any definition of homology among genes found in more than one species, we define a new tandem duplicate gene as the gene copy with greater amino acid sequence divergence from the single-copy orthologs in *D. simulans* and *D. yakuba* [35, 43, 79, 80]. In addition, three genes were formed by gene fusion (i.e. they are chimeric), and half of the remaining genes have unique insertions or deletions relative to their parents and orthologs, unambiguously defining them as the 'new' gene copies (Table 2.1, Figure A.1). Three duplications segregate in *D. melanogaster* at frequencies > 0.89 and are thus copy number variants (CNVs; Tables 2.1 and A.1).

We tested the essentiality of *D. melanogaster*-specific duplicates using constitutive RNAi. We constructed 16 new RNAi lines predicted to specifically target new or parent copies, then used these and other available RNAi lines to constitutively knock down target gene expression and monitored egg-to-adult survival, external adult morphology, and ability to produce progeny (Figure 2.2; Tables A.2, A.3, and A.4) [63, 84, 85]. We also collected and confirmed the phenotypes of available stocks carrying transposable element (TE) insertions or deficiencies predicted to disrupt the functions of *D. melanogaster*-specific genes or their parents (Table A.5). Altogether we were able to test the essentiality of 14 new and 15 parent genes using at least one type of evidence.

We observed no external morphological defects or significant sterility in any of the RNAi lines we screened. However, we found that 0/3 polymorphic and 3/11 (27%) fixed *D. melanogaster*-specific genes we tested are essential for proper fly development (Figure 2.2a; Tables A.3 and A.4). Knockdowns of *CG9123* and *tHMG1* with multiple lines and at least one of the two drivers caused complete lethality which was manifested in the pupa (*CG9123*) or prior to hatching (*tHMG1*; Figure 2.2b). Knockdown of *CG32165*, which is clearly incomplete (Figure 2.2c), caused a 20% reduction in fly survival to adulthood (Welch's t-test, Benjamini-Hochberg corrected $p < 0.05$ for both drivers; Table A.4). In addition, 3/15 (20%) parent genes we tested are essential by these criteria (*CG12264*,*Cdc23*, and *RpS15Aa*). The proportions of essential new and parent genes are not different (FET $p = 1$) and there are no duplicates which are both essential, though our numbers are small and the power to detect differences is low.

It is likely that either *CG18789* or *CG18787* and *Qtzl* are essential as well, but we could not confirm this. Constitutive, but non-specific knockdown of *CG18789* and its parent *CG18787* with TRiP line 55663 was completely lethal, but knockdown of *CG18787* 55375 was not lethal (Figure 2.2a). These results suggest that either *CG18789* is essential or the pair of genes together is essential. Furthermore, TE insertion in the 5' UTR of *Qtzl* at least

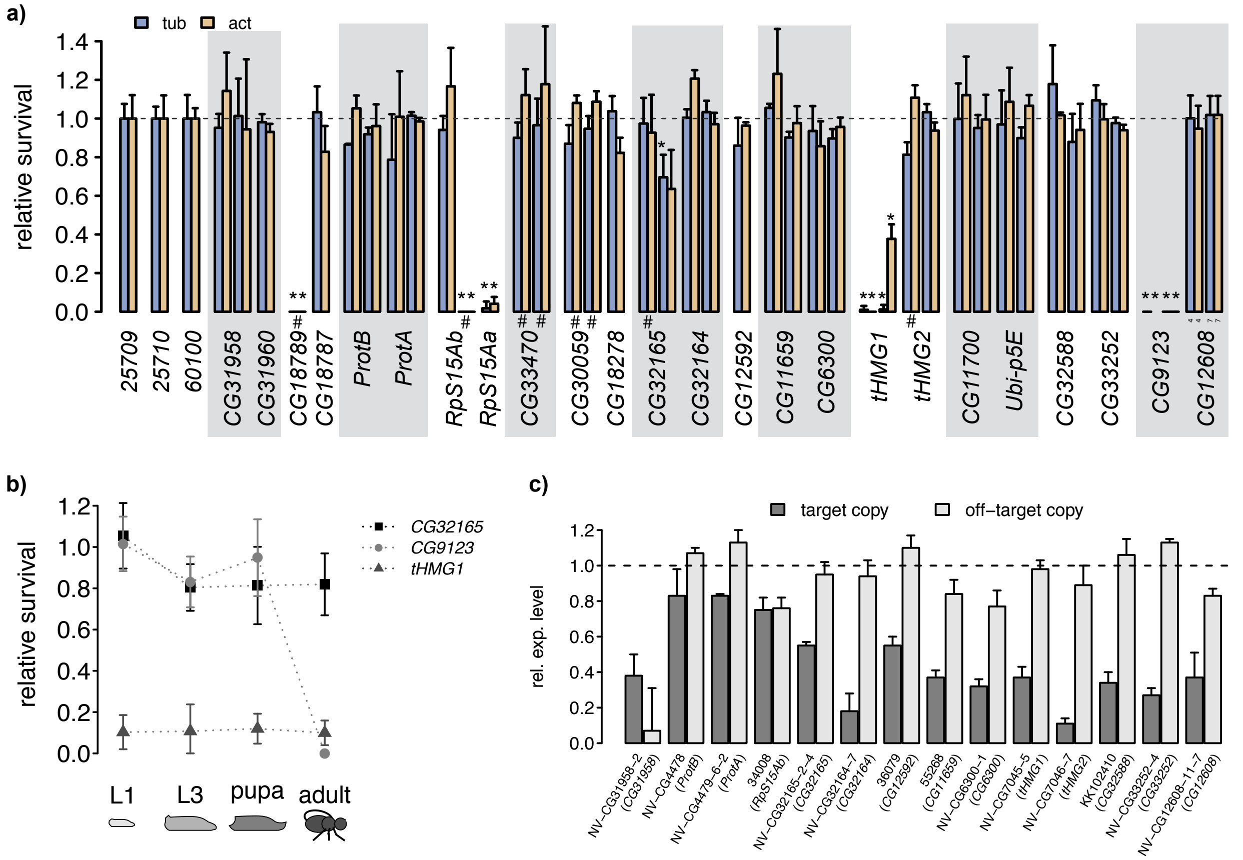


Figure 2.2: Constitutive gene expression knockdowns suggest that some *D. melanogaster*-specific genes are essential for proper fly development. **a)** The mean and 2 SEM from up to two RNAi lines are shown for each target gene for each of two constitutive drivers, *Act5C::GAL4* (act) and *αTub84B::GAL4* (tub). Each bar is comprised of 2 - 8 replicate crosses. # Indicates that the RNAi line targets both gene copies and is therefore not specific. Line names and full results for these and additional lines are shown in Tables A.3 and A.4. *Welch's two-sample t-test, Benjamin-Hochberg-corrected p < 0.05. **b)** The timing of lethality with essential gene RNAi. We used a GAL4 driver line carrying a GFP-marked balancer to track RNAi vs. non-RNAi F1 individuals. Each point summarizes the mean and 2 SEM of three replicate crosses of RNAi lines to *αTub84B::GAL4* drivers. The following lines were used: 60100-ϕC31 (control), NV-CG32165-2-4, NV-CG9123-8, NV-CG7045-5 (*tHMG1*). **c)** Relative expression levels of target and non-target (i.e. the other duplicate copy) genes in RNAi knockdowns versus controls. Quantitative PCR (qPCR) results were normalized using the $\Delta\Delta C_T$ method to *RpL32*. Bars show means and 2 SEM for three replicates. All results are from *αTub84B::GAL4* crosses except NV-CG7045-5, which yielded complete lethality with that driver but partial lethality with *Act5C::GAL4*. Labels are RNAi line names and target genes in parentheses.

causes sterility because it is and has always been maintained in a balanced stock (see ref [81] for discussion).

Altogether, these experiments show that species-specific genes, less than ∼2 million years old, can be essential components of *D. melanogaster* developmental programs. Their essentiality is independent of their parents' and the proportion of essential new genes is similar to the estimated proportion of essential old genes (25%-35%) [86, 87].

D. melanogaster-Specific Gene Pairs Rapidly Diverged in Expression Pattern

New duplicate gene pairs can rapidly diverge in their expression patterns, sequences, and structures [35, 77, 88–93]. At least 40% (6/14) of *D. melanogaster*-specific duplicate pairs we tested are not functionally redundant because specific knockdown or disruption of one copy causes lethality. To understand how gene pairs quickly became non-redundant, we next investigated divergence between *D. melanogaster*-specific duplicate genes in their expression patterns, gene structures, and gene sequences.

We first summarized duplicate expression patterns using whole transcriptome sequencing data from the modENCODE project available in FlyBase (Figure 2.3) [94–96]. Qualitatively, most gene pairs have different expression patterns among tissues and developmental stages (Figure 2.3a). Two thirds of *D. melanogaster*-specific new genes are most highly expressed in the testis and/or L3 imaginal disc, as are 47% of parent genes. Consistent with its RNAi phenotype, *CG9123* is most highly expressed in L3 larval imaginal discs, groups of cells that rapidly divide and differentiate to form adult fly tissues during metamorphosis, and during late larval to late pupal stages (Figure 2.3a). While *CG32165* and *tHMG1* are also highly expressed in imaginal discs and testis, these genes appear to exert their effects earlier in development (Figure 2.2a).

Most *D. melanogaster*-specific duplicates were formed by tandem duplication and likely share the same chromatin environment and regulatory regions [97]. We calculated correlation coefficients between *D. melanogaster*-specific duplicate pairs as well as duplicates that arose

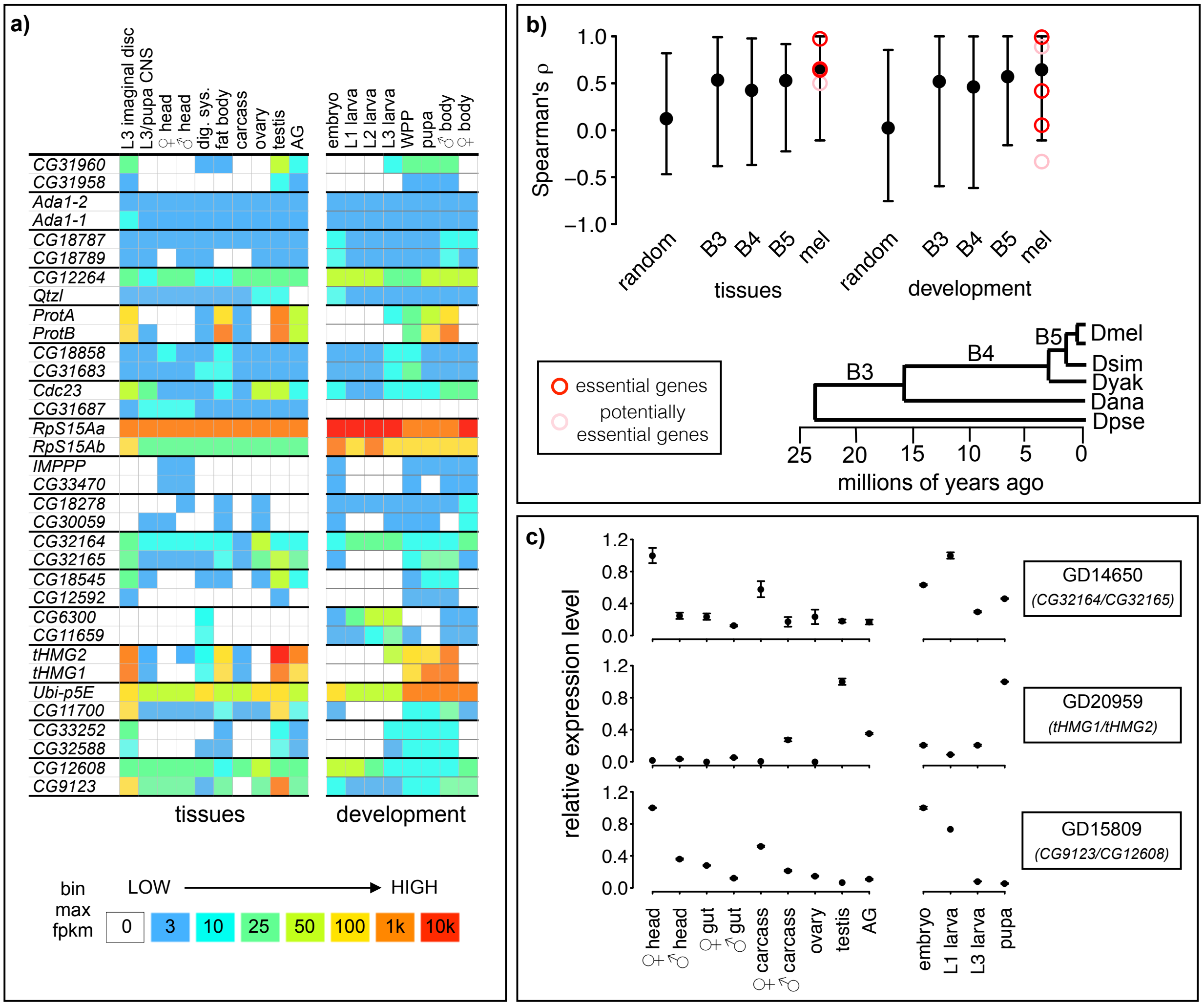


Figure 2.3: Expression patterns of *D. melanogaster*-specific genes and their parents. **a)** Heat map of new and parent gene expression levels in modENCODE tissue and development datasets [94, 95]. New gene/parent gene pairs are grouped, with parent gene above new gene (see also Table 2.1). Average FPKM values for redundant datasets are used (see Methods). WPP: white prepupa; CNS: central nervous system; dig. sys.: digestive system and larval salivary glands; AG: accessory gland. **b)** Pairwise correlations between duplicate gene pairs of different ages among the non-redundant modENCODE tissue and development RNAseq datasets. Means and 95% CIs are shown. Branch assignments are taken from ref. [43] and displayed in the tree. Correlations between 1,000 random protein-coding gene pairs are shown, too. Gene pair age is not a significant predictor of expression pattern correlation in either tissue ($F = 0.16$, $p = 0.69$) or development ($F = 1.59$, $p = 0.21$) datasets. **c)** Expression patterns of the *D. simulans* orthologs of *CG32165/CG32164*, *tHMG1/tHMG2*, and *CG9123/CG12608* in *D. simulans* strain *w501* tissues and developmental stages, relative to *RpL32*. Mean and 2 SEM are shown for 3 biological replicates. Relative expression levels are calculated as $2^{-(C_{T,G}-C_{T,R})}$, where $C_{T,G}$ and $C_{T,R}$ are the threshold cycles of the target gene and *RpL32*, respectively. Values are scaled to the maximum expression level within the tissue or development samples.

recently in the *D. melanogaster* lineage. We find no association between duplicate pair age and correlation coefficients among tissues (mean Spearman's Rank Correlation, $\bar{\rho} = 0.64$) or developmental stages ($\bar{\rho} = 0.73$), suggesting that expression divergence between duplicates occurs very soon after the duplication event (Figure 2.3b). Overall, the magnitude of expression divergence between *D. melanogaster*-specific gene pairs in their tissue or developmental expression patterns is not different (Wilcoxon Rank Sum $p = 0.13$). However, *CG32165* and *CG9123* have diverged especially rapidly from their parents across development, consistent with their phenotype ($\rho = 0.42$ and $\rho = 0.05$, respectively; Figure 2.3b). These particular low correlations are caused by asymmetric divergence in the new copies' expression patterns because their parents (*CG32164* and *CG12608*) have a qualitatively similar pattern to their *D. simulans* orthologs (Figure 2.3c).

Each gene performs its function in the context of other genes. To better understand the position of *D. melanogaster*-specific duplicates in gene-gene interaction networks, we constructed a gene co-expression network using modENCODE RNAseq data (Figure 2.4). We used centrality measures to assess the positions of *D. melanogaster*-specific duplicates in the network, and used their close interaction partners to assess what biological processes they are likely involved in (Figure 2.4; Tables 2.2). *D. melanogaster*-specific genes have similar numbers of direct interaction partners (degree) as all genes, suggesting that they have similar opportunities to participate in gene-gene interactions as their parents and all genes (Figure 2.4b). Essential genes have especially high betweenness values, suggesting they are essential because they have more pleiotropic effects (Figure 2.4c). Finally, no essential *D. melanogaster*-specific / parent gene pairs share any first degree interaction partners in this co-expression network, highlighting the large potential for rapid gene-gene interaction network divergence between gene pairs driven by expression divergence.

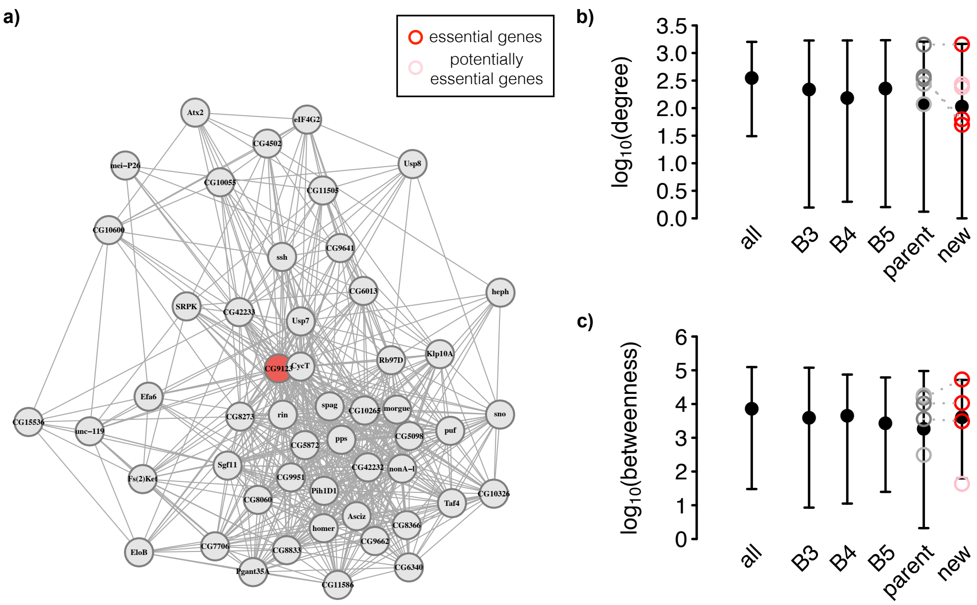


Figure 2.4: Expression pattern divergence and co-expression network characteristics of *D. melanogaster*-specific duplicates and other young genes in Drosophila. **a)** An example of the edges between one *D. melanogaster*-specific essential gene (*CG9123*, red) and its direct (first degree) interaction partners. Edges connect pairs of highly co-expressed genes (Spearman's $\rho > 0.78$ among all non-redundant modENCODE datasets, determined using permutations; Methods). Edge length is meaningless, but more tightly-connected genes are closer together. **b-c)** Co-expression network characteristics of all genes or genes of different ages. Gene ages (B3-B5) follow the tree in Figure 2.3b. New essential gene parents are shown in gray circles, and connected to their essential partners by dotted lines. Means and 95% CI are shown. *Qtzl* and *CG18789* are potentially essential genes (Figure 2.2; Tables A.3 and A.4). **b)** The distribution of degree, the number of first-degree interaction partners, for all protein-coding genes or genes of different ages. **c)** The distribution of betweenness, a measure of a gene's importance in connecting the network, for all protein-coding genes or genes of different ages.

Highly co-expressed genes are likely involved in similar biological processes. We used a Gene Ontology (GO) analysis of each *D. melanogaster*-specific duplicate's direct interactors to determine the biological process(es) in which that duplicate may be involved [98, 99] and combined these results with GO annotations based on sequence similarity and experimental evidence compiled by FlyBase (Tables 2.2 and 2.3). All duplicate gene pairs have the same FlyBase GO annotations [96], but 60% of pairs with GO annotations based on co-expression data for both copies have different annotations, including *CG32165/CG32164* and *CG9123/CG12608* (Table 2.2), further highlighting a large potential for functional di-

Table 2.2: GO analysis of co-expression network data. Only pairs where at least one duplicate has an annotation are shown.

Gene	**Biological Process GO**	**GO**	p^a	**enrichment**^{b}
CG31958	-	-	-	-
CG31960	TCA cycle	GO:0006099	0.02	6
Ada1-1	mitotic spindle elongation	GO:0000022	5.5E-27	32.7
Ada1-2	-	-	-	-
CG18789	mRNA splicing, via spliceosome	GO:0000398	0.89	1.5
CG18787	mRNA splicing via spliceosome	GO:0000398	5.5E-5	5.1
Qtzl	mRNA splicing, via spliceosome	GO:0000398	1.7E-9	9.8
CG12264/escl	mitochondrial electron transport/-	GO:0006120	6.70E-24	21.5
ProtB	microtubule-based movement	GO:0007018	4.3E-7	8.4
ProtA	microtubule-based movement	GO:0007018	1.4E-6	7.6
CG31687	-	-		
Cdc23/CG31688	anaphase-promoting complex/-	GO:0090302	1.5E-4	5.6
RpS15Ab	mitotic spindle elongation	GO:0000022	2.30E-29	42.4
RpS15Aa	translation	GO:0006412	1.00E-54	68
CG32165	protein dephosphorylation	GO:0006470	0.79	1
CG32164	MF:Ran GTPase binding	GO:0008536	0.01	3
CG12592	protein dephosphorylation	GO:0006470	2.80E-04	3.4
CG18545/sle	protein Ser-Thr phosphatase/	GO:0004722/	1.8E-5/	3.5/ 4.6
	centrosome organization	GO:005129	4.7E-5	
CG11659	metabolic process	GO:0008152	1.60E-03	2.8
CG6300	transmembrane transport	GO:0055085	3.30E-03	
tHMG1	microtubule-based movement	GO:0007018	1.90E-06	7.5
tHMG2	microtubule-based movement	GO:0007018	7.80E-07	7.7
CG11700	carbohydrate phosphorylation	GO:0046836	3.10E-03	3
Ubi-p5E	-	-	-	-
CG32588	TCA cycle	GO:0006099	0.02	6.5
CG33252	TCA cycle	GO:0006099	3.60E-03	7.7
CG9123	mRNA splicing, via spliceosome	GO:0000398	0.51	2.3
CG12608	pseudouridine synthesis	GO:0001522	1.90E-03	2.6

a: Benjamini-Hochberg corrected p-value

b: Enrichment score from DAVID 6.8 beta

vergence between pairs due to expression pattern differences.

We note that these analyses are assessing the genes that duplicates interact with on a broad scale. It is possible, or even likely that their (essential) functions are in a particular tissue or developmental stage with a particular gene or set of genes. Thus, here we are only describing the broad-scale potential for functional divergence, not inferring the exact causes

Table 2.3: GO terms associated with *D. melanogaster*-specific gene pairs based on sequence similarity and experimental evidence from FlyBase 2015_05. Pairs always have the same annotation.

New Gene	Parent Gene	GO Term	GO
CG31958	*CG31960*	-	-
Ada1-1	*Ada1-2*	histone H3 acetylation	GO:0043966
CG18789	*CG18787*	-	-
Qtzl	*CG12264/escl*	Fe-S clustering/H3K27 methylation	GO:0016226/GO:0070734
ProtB	*ProtA*	spermatogenesis	GO:0035093
CG31683	*CG18858*	lipid metabolic process	GO:0006629
CG31687	*Cdc23/CG31688*	regulation of mitosis/-	GO:0030071
RpS15Ab	*RpS15Aa*	cytoplasmic translation	GO:0002181
CG33470	*IMPPP*	antibacterial humoral response	GO:0019731
CG30059	*CG18278*	glycosaminoglycan metabolic process	GO:0030203
CG32165	*CG32164*	intracellular protein transport	GO:0006886
CG12592	*CG18545/sle*	-/nucleolus organization	GO:0007000
CG11659	*CG6300*	metabolic process	GO:0008152
tHMG1	*tHMG2*	-	-
CG11700	*Ubi-p5E*	protein ubiquitination	GO:0016567
CG32588	*CG33252*	biological_process	GO:0008150
CG9123	*CG12608*	cell proliferation	GO:0008283

of the non-redundant fitness effects.

Most *D. melanogaster*-Specific Duplications Are In Recent, Strong Selective Sweeps

Gene duplicates may diverge due to neutral or selection-driven processes. We next tested whether selection has driven *D. melanogaster*-specific duplicate divergence by analyzing rates of sequence evolution and signatures of selection in regions containing species-specific duplications.

One signature of neofunctionalization is an accelerated amino acid substitution [18, 54]. However, no *D. melanogaster*-specific duplicates exhibit accelerated evolution according to maximum likelihood analyses of the nonsynonymous to synonymous site substitution rates (d_N and d_S; Table A.8) [100]. Furthermore, only 3/37 new or parent genes contain a significant excess of nonsynonymous substitutions relative to polymorphisms according to McDonald-Kreitman tests (Table A.7) [101]. There is thus little evidence of rapid amino acid sequence evolution in *D. melanogaster*-specific genes or their parents. Additionally, all ancestral pre-duplication genes, except the *CG32588/CG33252* ancestor, exhibit $d_S > 0.03$ and $d_N/d_S << 1$ on the branch between *D. simulans* and the duplication event, suggesting that they were under strong purifying selection before the duplication occurred and that selection-driven subfunctionalization models do not describe *D. melanogaster*-specific gene pair evolution, at least at the gene sequence level (Table A.8).

Selection can cause new beneficial alleles to rapidly fix in a population. Alleles at tightly linked sites hitchhike with the selected allele and are also rapidly fixed, leading to a temporary reduction in nucleotide variation in genome regions near the selected site just after fixation (i.e. a selective sweep) [102, 103]. We expected that any sweep signatures would still be present in the regions surrounding *D. melanogaster*-specific duplications because if a sweep did occur, it must have occurred within the last 2 million years. Genome regions that recently underwent a selective sweep are expected to contain 1) an excess of intermediate and high-frequency derived alleles when the sweep completes, 2) an excess of low-frequency alleles shortly after sweep completion, and 3) a reduction in variation relative to divergence. These characteristics can be quantified using the population genetic statistics Fay and Wu's H (1), Tajima's D (2), and the Hudson-Kreitman-Aguadé test (3) [104–106]. To avoid the problem of low diversity in *D. melanogaster*-specific duplications themselves, we calculated these statistics in sliding windows across the *D. melanogaster* genome in 17 African fly genomes, used the window centered on the duplication to represent that duplication, and considered a duplication to be in a sweep if its window's statistic was in the lowest 5% of values for its chromosome arm (Figure 2.5; Table A.9) [83, 104–108].

Eight out of 13 (62%) duplications reside in windows of extremely negative, H, D and/or $HKAl$ statistics, suggesting that they reside in recent selective sweeps (Figure 2.5). The

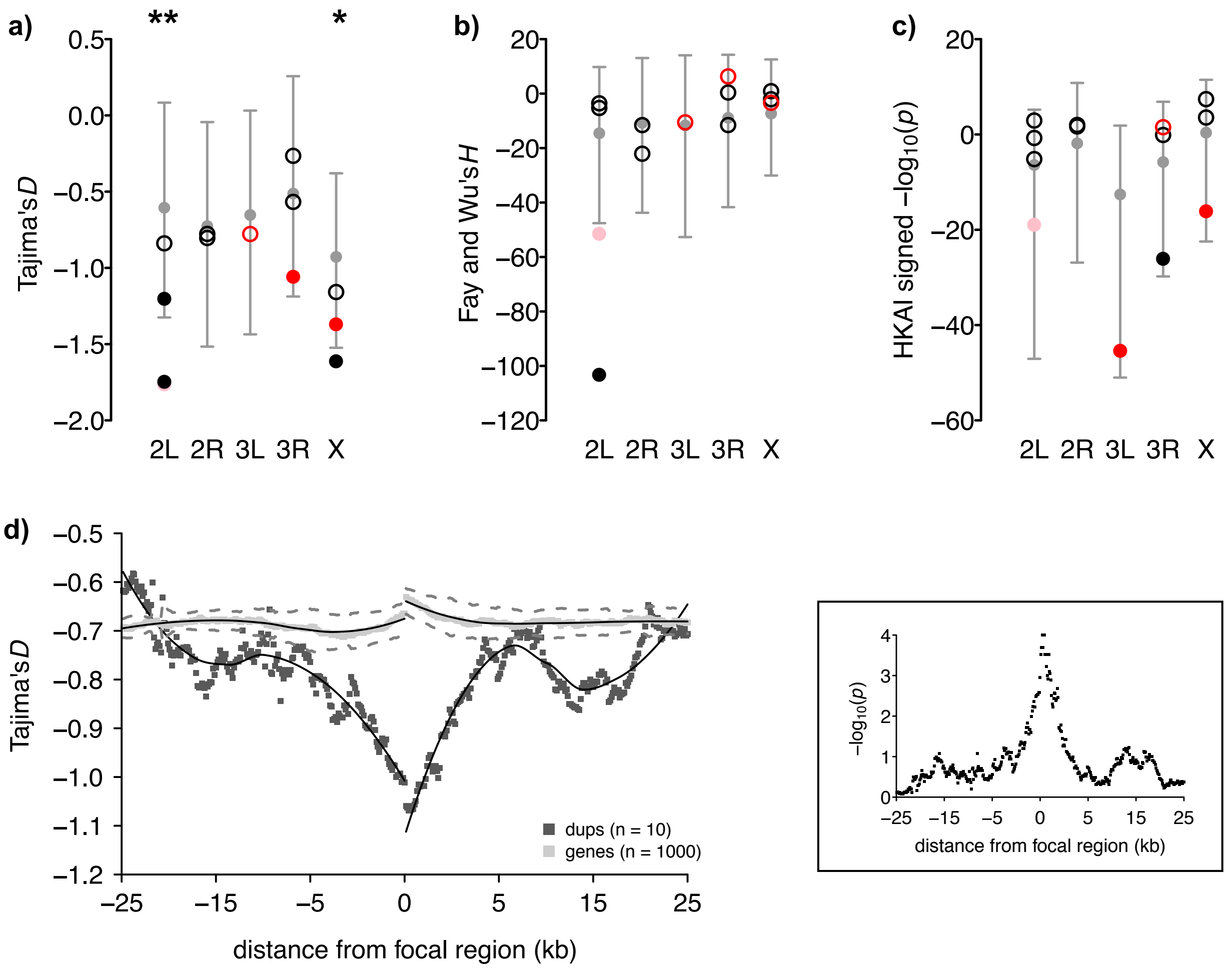


Figure 2.5: Population genetic summary statistics in regions surrounding *D. melanogaster*-specific duplications. **a-c)** Tajima's D, Fay and Wu's H, and the signed Hudson-Kreitman-Aguad-like test p-values in windows of 250 informative sites centered on *D. melanogaster*-specific duplications (points). The means and 95% CI of windows centered on annotated protein-coding genes are shown for reference in gray. Red points are values for duplications containing new essential genes. Filled circles indicate duplications in windows with statistics in the lowest 5% of statistics on their chromosome arm. Wilcoxon Rank Sum $p < 0.05$ (*) or $p < 0.01$ (**) are shown for significant comparisons. **d)** Mean Tajima's D values calculated in 5 kb sliding windows (100 bp step) in the regions flanking autosomal duplications (dups) or protein-coding genes with *D. melanogaster*-specific K_S between 0 and 0.01. Such genes have accumulated relatively neutral substitutions since *D. melanogaster* and *D. simulans* diverged and should roughly represent the expected level of D surrounding recently-fixed neutral variants. Lines are smoothed LOESS fits (solid) and 95% CI upper and lower bounds (dotted) for each window, calculated separately for left and right flanking regions. Probability values for each window (right) were determined using 10,000 permutations. We obtained similar results using regions flanking *D. melanogaster*-specific intergenic substitutions (Figure A.2). Duplicate or gene regions are not included in any of the calculations in **a-d**.

general patterns of these per-duplication results are clearly visible in plots of D in windows flanking duplications (Figures 2.5d and A.2). Low D and H values are likely due to recent strong positive selection because most windows also have significantly low ratios of polymorphism to divergence, indicated by negative *HKAl* statistics (Figure 2.5c). More duplications have extremely negative D values than H values, suggesting that most of these sweeps completed some time ago and are beginning to recover variation [105, 106].

Our window analysis prevents identification of the precise target of selection. However, six *D. melanogaster*-specific genes with demonstrated fitness effects (*CG11700*, *CG32165*, *tHMG1*, *Sdic*, *Qtzl*, and *CG9123*) all lie in sweeps. In fact, previous studies of the sweeps containing *CG32165*, *Qtzl*, and *CG9123* estimated they were caused by variants with selection coefficients of 0.007, 0.006, and 0.010 and to have completed roughly 50,000, 15,000, and 20,000 years ago, respectively [81, 91, 109]. Furthermore, 4/8 sweeps we identified here were also identified by ref. [83]. There is thus good evidence that 62% of *D. melanogaster*-specific duplications were recently involved in strong selective sweeps, and that *D. melanogaster*-specific duplications have large beneficial effects that arose concurrent with or shortly after the duplication event.

2.0.4 Discussion

Contrary to suggestions that "essential genes are not organism specific" [72], our work clearly shows that species-specific genes can rapidly become essential components of the genetic networks that control development. We used RNAi and available disruption lines to show that at least 27% of fixed new species-specific genes are essential for proper fly development and that essential gene knockdowns cause lethality between fertilization and eclosion. Importantly, while the common claim that protein sequence conservation indicates functional importance is certainly true, our results highlight the fact that the converse statement is not true: functional importance is not necessarily predicted by sequence conservation (and gene age) [45].

These results challenge the assumption that a gene's function in one organism can be inferred to be the function of its ortholog, and suggest instead that essential gene functions may evolve *de novo* or be partitioned into different gene copies in different lineages. How often are ortholog functions different in different taxa?

Note also that all copy-specific d_S values are well within the range of d values calculated for *D. melanogaster*-specific divergence at 4-fold degenerate sites or short introns calculated by Hu and her colleagues [110]. This suggests that these duplications are in fact *D. melanogaster*-specific. However, even if the duplication occurred before the *D. simulans/D. melanogaster* split and one duplicate was subsequently lost in *D. simulans*, these genes are still found only in *D. melanogaster* and these conclusions are unchanged.

Species-specific genes that are essential for development support the existence of species-specific network topologies. Duplicate genes, especially tandem duplicates like the ones we examined in this study, are likely expressed from the moment they are formed and therefore interact with co-expressed genes. New duplicates can be fixed as long as they provide a net benefit. However, the existing gene-gene interaction network will likely need to change to accommodate the addition of the new gene and any pleiotropic effects that addition may have, and rapidly drive the divergence of these networks between populations [65, 111, 112]. Our finding that essential *D. melanogaster*-specific / parent gene pairs quickly diverged in their developmental expression patterns is consistent with their loss-of-function phenotypes, and, combined with the fact that there is little amino acid sequence divergence between duplicates, suggests that expression divergence is likely the main driver of functional differentiation between the two copies. The fact that no essential new gene / parent gene pairs share any direct interaction partners in the co-expression network we constructed further highlights how pairs can diverge in interaction networks. Recent work has shown that young genes and even intraspecific genetic variants can significantly impact development and show that the genetic basis and control of this important process can change in short periods of time [63, 69, 113, 114].

Our observation that the pre-duplication genes were under strong purifying selection before the duplication event suggests that whatever benefit the duplication provides is large ($s \approx 0.01$) and arose concurrent with or shortly after duplication [81, 91, 109]. Neutral variants in *D. melanogaster* are expected to fix in, on average, ~400,000 years, so it is possible the new duplication drifted to fixation and subsequently swept specifically in *D. melanogaster*, but the population dynamics of variants with selection coefficients of magnitudes greater than $\sim 10^{-7}$ are dominated by selection in populations as large as *D. melanogaster*'s, and new duplicates probably rarely drift to fixation [115–117].

An important question that we cannot answer is: What function was selected? We cannot definitively say whether the essential functions of *CG32165*, *tHMG1*, or *CG9123* are novel or beneficial and it is difficult to imagine how a new essential function can arise. A plausible scenario may be that a sex-specific benefit was selected and an ancestral essential function simply partitioned into the more divergent gene copy. Non-essential *D. melanogaster*-specific duplicates may have more subtle fitness effects than we measured, and there is strong evidence from the *Sdic* gene family [118, 119] and *CG11700* [82] that *D. melanogaster*-specific gene effects may often be slight and male-beneficial. For example, deletion of the *Sdic* gene family causes a $\sim 1.5\%$ reduction in sperm competitiveness [119]. These observations and the fact that many new genes in mammals and insects exhibit testis-biased expression patterns suggests that they are mainly beneficial to males [43, 56, 120]. Indeed, *ProtB* and *tHMG1* are important components of networks controlling spermatogenesis, and 60% of species-specific new genes we studied here have their highest expression in testis and/or imaginal disc [121, 122]. Experimental tests of the fitness effects of the pre-duplication, ancestral gene in closely-related outgroup species may help determine if these essential functions are novel or not.

Finally, while we only consider *CG32165*, *tHMG1*, and *CG9123* as essential species-specific genes, it is very likely that *Qtzl* and possibly *CG18789* are essential as well based on disruption lines and RNAi results. Furthermore, our qPCR experiments and others show that at least 40% of our experiments produce $< 20\%$ knockdown of target gene expression, and so we are probably not testing the effects of those genes [84]. Altogether, the fact that previous studies estimate the number of *D. melanogaster*-specific genes to be ∼60 [35, 43], the high false negative rate of RNAi, and our focus on gene essentiality suggest that there could be substantially more species-specific genes with strong fitness effects than we uncovered in this study.

2.0.5 Acknowledgments

We thank G.Y-C. Lee for advice on population genomics analyses, W. Zhang and U. Lee for help with constructing gene co-expression networks, and other members of the Long lab for valuable discussion. We thank D. Barbash and D. Begun for helpful comments on a previous version of the manuscript. NWV was partially supported by National Institutes of Health Genetics and Regulation Training Grant T32GM007197 and a NSF Graduate Research Fellowship. This work was also supported by NSF1026200, NIH R01GM100768-01A1 and R01GM116113.

2.0.6 Methods

D. melanogaster-Specific Gene Identification

Candidate Dmel-specific genes were collected from previous studies of new gene origination in Drosophila [35, 43, 66]. We removed from this initial list of 233 candidates 1) any genes whose Dmel release 6.05 (http://flybase.org) annotation status is 'withdrawn', 2) genes not located on the major chromosome arms 2L, 2R , 3L, 3R, or X, and 3) members of large tandem arrays, including the *Sperm dynein intermediate chain* [118, 119], *Stellate* (*Ste*), and X: 19,900,000-19,960,000 arrays that appear to be Dmel-specific but are impossible to specifically study. We checked syntenic whole-genome alignments of the remaining 84 genes manually using our multi-species alignments (see below) and the UCSC Genome Browser (http://genome.ucsc.edu/). We required that the Dsim, Dsec, Dyak, and Dere genome assemblies contained no assembly gaps, transposable elements, or repeats in the region containing the putative Dmel-specific gene. Dmel-specific gene formation mechanisms and parent genes were taken from the original studies and confirmed using BLAT and BLASTp. If a gene had multiple significant ($E < 10^{-10}$) full-length BLASTp hits, the hit that was most similar to the Dmel-specific gene was assumed to be the parent. Dmel-specific genes often have unique structures not present in outgroup orthologs (Table 2.1, Figure A.1). We also used available Dsim and Dyak next generation sequencing reads to test the presence of putative Dmel-specific tandem duplications in these two species (Figure A.3) [110, 123]. We found no uniquely-mapped read pairs supporting Dmel tandem duplications in any of 20 Dsim or 20 Dyak genomes, supporting the idea that these tandem duplications are specifically found in Dmel and are not simply missing from the Dyak and Dsim reference genome assemblies. We checked if any of the duplications in our final set are segregating within Dmel by analyzing whole genome re-sequencing data from the DPGP2 core Rwanda (RG) genomes [83]. We required tandem duplications to have at least one read uniquely mapped to each of the three unique breakpoints in order to be called as ‘present’ in a particular line. Ten of these genes are not found in any of 17 additional *D. melanogaster* genomes we analyzed, suggesting that they are singletons found specifically in the reference stock.

RNAi Strain Construction

We generated new, specific RNAi lines following ref [84]. We designed RNAi reagents using the E-RNAi server (http://www.dkfz.de/signaling/e-rnai3/) and kept constructs with 100% of the possible 19-mers uniquely matching the intended target gene and excluding designs with >1 CAN repeat [124]. Constructs were cloned into pKC26 following the Vienna Drosophila Resource Center’s (VDRC’s) KK library strategy (http://stockcenter.vdrc.at, last accessed 2 February 2016). We introgressed the X chromosome from Bloomington Drosophila Stock Center line 34772, which expresses ϕC31 integrase in ovary under control of the *nanos* promoter, into the VDRC 60100 strain. Strain 60100 carries attP sites at 2L:22,019,296 and 2L:9,437,482 [125]. We ensured that our RNAi constructs were inserted only at the 2L:9,437,482 site using PCR following [125]. RNAi constructs were injected into the 60100-ϕC31 at 250 ng/μL. Surviving adult flies were crossed to sna^{Sco}/CyO balancer flies (BDSC 9325) and individual insertion strains isolated by backcrossing.

RNAi Screen

We constitutively knocked down target gene expression using driver lines constitutively and ubiquitously expressing GAL4 under the control of either the *Actin5C* or *αTubulin84B* promoter. We replaced driver line balancer chromosomes with GFP balancer chromosomes to enable tracking of non-RNAi progeny. As controls in each of the following experiments, flies from the background strains 60100-ϕC31, 25709, or 25710 were crossed to driver strains and treated identically. Five males and five virgin driver females were used in each cross. Crosses were grown at 25°C, 40% - 60% humidity, and a 12h:12h light:dark cycle. F1 progeny were counted at day 19 after crossing, after all pupae had emerged. We screened F1 RNAi flies for visible morphological defects in the fly's general structure and coloring in addition to 1) wings: vein patterning and numbers, wing periphery; 2) notum: general bristle organization and number, structure and smoothness; 3) legs: number of segments. We monitored survival of RNAi F1s by counting GFP and non-GFP L1 and L3 larvae and pupae. We tested RNAi F1 sterility by crossing individual RNAi F1 flies to 60100-ϕC31 and monitoring vials for L1 production. Ten replicates for each sex for each line were performed. No significant differences were found between the fractions of fertile RNAi flies and control flies for any of the lines (no experiment or control produced more than 2/10 sterile vials).

RNAi Knockdown Specificity and Sensitivity

We sought to address two known problems of RNAi technology using RT-qPCR. First, since off-target effects are common in RNAi experiments [124, 126], we wanted to test that our lines are specifically knocking down target gene expression. Our constructs are computationally predicted to be specific. Second, since RNAi is often incomplete [84], we wanted to estimate how many genes we are actually able to test. We chose a sample of lines because either 1) they targeted an essential D. melanogaster-specific gene or its parent or 2) the target gene had no phenotype according to previous studies or ours. We collected qPCR primers from FlyPrimerBank [127]. For those genes not found in FlyPrimerBank we designed primers specifically targeting a 100 bp region of the target gene using Primer-BLAST (Table A.6). We confirmed primer specificity with PCR and Sanger sequencing. We extracted RNA from sets of 16 flies (8 females and 8 males) in triplicate from each RNAi cross using TRIzol (Thermo Fisher, USA), treated 2μg RNA with RNase-free DNase I (Invitrogen, USA), then used 2 μL treated RNA in cDNA synthesis with SuperScript III Reverse Transciptase (Invitrogen, USA) using oligo(dT)$_2$0 primers. cDNA was diluted 1:10 in water before using 2 μL as template in 10 μL qPCRs with iTaq^{TM} Universal SYBR Green Supermix (Bio-Rad, USA) and 400 nM each primer. Reactions were run on a Bio-Rad C1000 Touch thermal cycler with CFX96 detection system (BioRad, CA). Cycling conditions were 95°C for 30 sec, then 45 cycles of 95°C for 5 sec, 60°C for 30 sec, and 72°C for 15 sec. We normalized gene expression levels using the $\Delta\Delta C_t$ method and *RpL32* C_ts as reference gene expression levels. We tested the efficiency of qPCR primers using a 8-log$_2$ dilution series for each primer pair (Table A.6).

D. melanogaster SNPs

We called SNPs in the 17 primary core RG (Rwanda) samples from the *Drosophila* Population Genomics Project Phase 2 (DPGP2; www.dpgp.org) with less than 3% admixture [83]: RG2, RG3, RG4N, RG5, RG7, RG9, RG18N, RG19, RG22, RG24, RG25, RG28, RG32N, RG33, RG34, RG36, and RG38N. We downloaded raw sequencing reads from the NCBI Short Read Archive and mapped them to the release 6 genome assembly using bwa mem v0.7.12 with default parameters except the -M flag [128]. PCR duplicate reads were marked with Picard Tools v1.95 [129]. Alignments were processed following the Genome Analysis ToolKit's (GATK's) best practices workflow [130]. We used GATK v3.4-0 with default parameters to realign reads around putative indels and to recalibrate read base quality scores on individual sample alignment files [131–133]. DPGP SNP calls [108] were used as known variant sites in base quality score recalibration.

SNPs were called using the GATK's HaplotypeCaller in GVCF mode with default settings except sample ploidy was set to 1 and heterozygosity was set to 0.00752, the genome average value for African lines determined by ref. [108]. Samples were then jointly genotyped with GATK's GenotypeGVCFs with default parameters except heterozygosity and ploidy listed above, then hard filtered by the following criteria: $|\text{BaseQRankSum}| > 2.0$ to ensure sites were not supported only by low-mapping-quality reads, $|\text{ClippingRankSum}| > 2.0$ to ensure supporting reads were not being excessively trimmed, MQRankSum $<$ -2.0 to ensure mapping quality was not extremely low, ReadPosRankSum $<$ -2.0 to ensure call support came from all read regions instead of just the low-quality beginning and ends, and a per-sample genotype phred score $>$30. All SNP sites overlapping repeat-masked regions (UCSC Genome Browser Simple Repeats and RepeatMasker tracks, downloaded 21 June 2015) or putatively admixed sites were removed [83].

Multi-Species Alignments and SNP Polarization

I aligned *D. simulans* (Dsim) release 2.01, *D. sechellia* (Dsec) release 1.3, *D. yakuba* (Dyak) release 1.04, and *D. erecta* (Dere) release 1.04 genome assemblies to the release 6 *D. melanogaster* (Dmel) sequence following the UCSC Genome Browser pipeline for reference-guided multi-species alignments using roast v3 (part of the multiz 11.2 package) [134, 135]. Analyses requiring polarized SNP states only used SNPs where one of the Dmel SNP states unambiguously matched the state in species from both the Dyak/Dere and Dsim/Dsec clades.

d_N/d_S and M-K Calculations

We calculated the ratio of nonsynonymous substitutions per nonsynonymous site to synonymous substitutions per synonymous site (d_N/d_S or ω) for Dmel-specific duplicate pairs using

PAML 4.7a (Table A.8) [100]. Ortholog coding sequences (CDS) from Dsim, Dsec, Dyak, Dere, and *D. ananassae* (Dana) were collected from FlyBase release 2015_05. If a parent gene had multiple orthologs in a single species we only used the ortholog sequence with the highest score from pairwise BLASTn. The ortholog assignments were the same if we required highest similarity to the new gene instead of the parent gene. We used only the longest CDS for each gene in analyses. We used the Dmel reference genome CDS for the Dmel-specific gene and its parent in these analyses. Alignments were generated using TranslatorX v1.1 with MUSCLE v3.8.31 [136, 137]. PAML analyses used the guide tree ((((Dmel new gene, Dmel parent gene), (Dsim, Dsec)), (Dyak, Dere)), Dana). We estimated ω for each branch under a free-ratio model, and also specifically tested whether the Dmel-specific gene has evolved at a different rate from its orthologs using 2-ratio models. To specifically test if a gene has evolved with a significantly elevated ω, we compared a 2-ratio model allowing the gene's terminal branch to evolve at any rate to a 2-ratio model constraining the gene's branch to evolve with $\omega = 1$. Models were compared with a standard likelihood ratio test with 1 d.f.

Polarized McDonald-Kreitman tests, excluding singletons and sites called in fewer than 9/17 RG samples, were conducted for each duplicate gene pair using polymorphism data generated above [101]. Orthologs from Dsim and Dyak and the 17 Dmel ingroup sequences were aligned using TranslatorX v1.1 and MUSCLE v3.8.31.

Population Genetic Summary Statistics and Analyses

D, *H*, and Hudson-Kreitman-Aguadé-like statistics were calculated using perl scripts which take into account the fact that segregating sites may not be called in all samples [107, 108]. We calculated *D*, *H*, and *HKAl* in sliding windows of informative sites across the major chromosome arms 2L, 2R, 3L, 3R, and X. An informative site is a segregating site that is called in at least 9/17 RG genomes (D) and is confidently polarized (H and $HKAl$). We used windows of informative sites so that each window has equal information for D, H, and $HKAl$ estimates. D and H have complementary power to detect the effects of strong selection that acted on loci at different times after a selective sweep has occurred. H is sensitive to an excess of derived variants that hitchhiked to intermediate and high frequencies with a selected variant while D is sensitive to the excess of low-frequency derived variants that accumulate in the sweep region after sweep completion. Thus, H is most powerful to detect a sweep that has just completed while D relies on some amount of variation recovery and the two methods complement each other. It is also worth noting that, even at their best, these statistics have 60% power to detect even the strongest selective sweeps [106].

The expected numbers of divergent and segregating sites for each chromosome arm were calculated following ref. [107]. Low-recombination regions (as defined in ref. [108]) were excluded from the calculations and only confidently polarized sites were included. Sites that were both polymorphic and divergent were classified as segregating sites. These expected values were used in $HKAl$ calculations.

We followed Sattath et al. (2011) [138] to generate plots of D surrounding *D. melanogaster*-specific duplications and 1,000 random intergenic fixed sites or 1,000 random genes with $0 < K_s < 0.01$, where K_s is the fraction of synonymous sites that have mutated and fixed specifically in Dmel since the Dmel-Dsim split. The idea is that these genes have fixed multiple, putatively neutral variants in roughly the same period of time that Dmel-specific duplications arose and fixed and may reflect the reduction of D or π surrounding gene regions expected when they accumulate neutral mutations.

Expression Pattern Analyses

We used modENCODE gene expression level (FPKM) estimates for all annotated genes from FlyBase release 2015_05 [94, 95]. Figure 2.3 was generated using average FPKMs for redundant or similar datasets. Statistical analyses were performed with R 3.0.1 [139]. We constructed a gene co-expression network for all protein-coding genes in all modENCODE tissue and development datasets using pairwise Spearman Rank Correlations [140]. We excluded genes with no detected expression in any dataset. We used average FPKM values for male whole body, female whole body, carcass, digestive system, and head samples because they are redundant (Figure A.4). Gene pairs with correlations greater than the highest correlation value recovered by bootstrapping the datasets (10,000 replicates, maximum $\rho = 0.78$) were used as edges in network construction. We calculated network statistics using igraph 1.0.1 [141].

CHAPTER 3

A PAIR OF DUPLICATE *DROSOPHILA MELANOGASTER*-SPECIFIC GENES ARE BOTH ESSENTIAL, BUT FOR OPPOSITE SEXES

3.0.1 Abstract

New duplicate genes can rapidly become critical components of the biological networks controlling important processes such as development and reproduction. Thus, it is crucial to understand the process by which duplicate genes diverge at the levels of sequence, structure, and expression pattern to become functionally non-redundant. Here we investigate the evolution of a pair of duplicate genes, *CG32164* and *CG32165*, that formed, fixed, and functionally diverged specifically in *Drosophila melanogaster* (i.e. within that last ∼2 million years). We find that knockout of *CG32165*, but not knockout of *CG32164*, causes a significant (∼20%) reduction in fly survival. However, both duplicates are essential for fly reproduction and are sexually antagonistic:*CG32164* knockout causes complete female sterility but significantly increased male reproductive output, while *CG32165* knockout causes complete male sterility but significantly increased female reproductive output. The sterility effects are caused by defects in oogenesis and spermatogenesis, as *CG32164*^{null} flies produce round eggs and *CG32165*^{null} flies fail to produce mature sperm. Asymmetric expression pattern divergence between the duplicates and their *D. simulans* ortholog suggest that *CG32165* specifically has gained testis expression since the duplication event. Altogether, our results suggest that *CG32164*/*CG32165* evolution was driven by meiotic drive or, more likely, the resolution of sexual antagonism. We thus provide empirical evidence for a general framework for understanding the rapid evolution of essential new genes.

3.0.2 Introduction

New gene origination is a central process in the evolution of the genome and novel phenotypes [16, 27, 56]. New genes are often formed by duplication of an existing gene. Duplicate genes are functionally redundant until they accumulate mutations that cause them to have distinct expression patterns, gene sequences, and/or gene structures [18, 21, 23, 55, 76]. Duplicate gene pairs have been experimentally shown to become functionally non-redundant within just several millions of years (Chapter 2; see ref. [27] for additional examples). However, it is rarely clear exactly how these duplicates initially became non-redundant because they are relatively old and have accumulated many mutations since they formed. The mutations that initially caused the functional differentiation between copies have thus been obscured by continued mutation accumulation, and it is usually difficult to discern the cause of the functional differentiation and if selection has acted on gene copies.

In particular, gene pairs may become different because one copy gains a truly novel function while the the other maintains the ancestral function (neofunctionalization), or non-redundant subsets of ancestral gene functions may be partitioned into the two copies (subfunctionalization) [18, 23, 76]. Thus, it is important to study the evolution and fitness effects of genes at the earliest stages of their evolution to understand the initial stages of functional differentiation between duplicate genes and the forces that drive it [54].

New duplicate genes in Drosophila can quickly become essential components of the genetic networks controlling development within ~2 million years (Chapter 2) [63, 70]. We previously showed that genes found specifically in *D. melanogaster* often reside in selective sweeps and that they have significantly divergent expression patterns and gene structures (Chapter 2). Furthermore, three of these species-specific genes were essential for fly development. These results suggested that 1) new duplications are frequently strongly beneficial and 2) functional

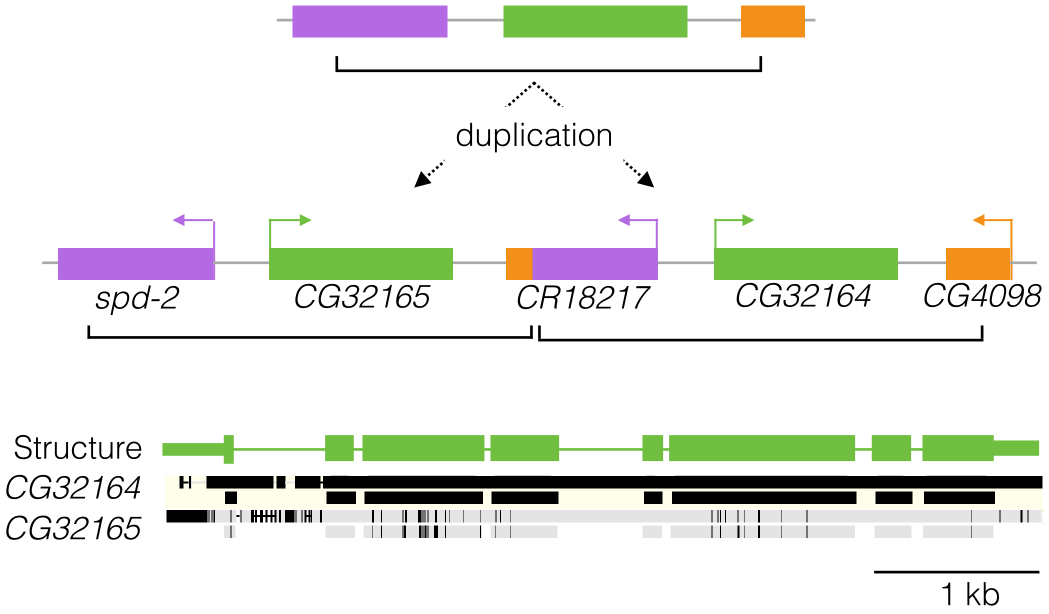


Figure 3.1: *CG32165* and *CG32164* were formed by 7.8 kb tandem duplication after *D. melanogaster* and *D. simulans* diverged ∼2 mya [78]. Top: The tandem duplication produced *CG32165, CG32164*, and a novel chimeric pseudogene *CR18217*. Bottom: Pairwise alignment of *CG32165* and *CG32164* consensus gene regions from 17 DPGP2 'RG' genomes [83]. The top bar shows the gene structure. Lines are introns, thick boxes are coding exons, and thin boxes are UTRs. For each gene, the nucleotide sequence (upper bar) and the corresponding amino acid translation (lower bar) are shown. Sites that differ in *CG32165* relative to *CG32164* are shown as vertical black bars. Thin lines are gaps.

differentiation must occur concurrent with or shortly after the duplication event. However, that work did not indicate what particular gene functions may have been selected to cause the selective sweeps, nor precisely why such new genes were essential but their parents were not. To better understand how new genes quickly become essential, we studied the evolution of one essential *D. melanogaster*-specific gene, *CG32165*, and its duplicate copy *CG32164*.

CG32165 and *CG32164* were formed by a tandem duplication on chromosome 3L in the last 2 million years, since *D. simulans* and *D. melanogaster* diverged (Figure 3.1), yet constitutive RNA interference (RNAi) of *CG32165* but not *CG32164* results in a ∼20% reduction in egg-to-adult survival. RNAi of either gene had no measured effects on male or female sterility (Chapter 2). Furthermore, *CG32165* and *CG32164* reside in the middle

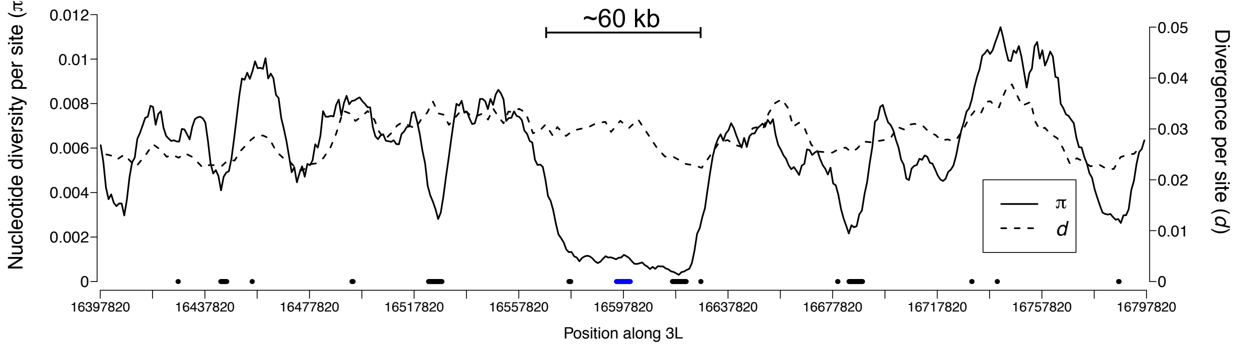


Figure 3.2: *CG32165* and *CG32164* reside in a recent, strong selective sweep. The blue bar along the X-axis delineates the tandem duplicates, while black bars denote repeat regions > 500 bp. π is calculated in sliding windows of 10 kb with 1 kb step. Divergence is polarized divergence to *D. simulans* and *D. yakuba* in sliding windows of 500 divergent sites (100 site step).

of a large selective sweep that was estimated to have completed 20,000 years ago and caused by a variant with a selection coefficient of 0.006, suggesting that the duplication, or a variant to which it is tightly linked, has a large beneficial effect (Figure 3.2) [91]. Thus, the young age, strong phenotype, clear non-redundancy, and potentially strong benefit of the *CG32164*/*CG32165* duplication make it a prime candidate for understanding the initial stages of duplicate gene differentiation and the role of selection in the process.

Here we use *CG32165* and *CG32164* CRISPR/Cas9-induced knockouts to show that, in fact, both duplicates are essential for fly reproduction and sexually antagonistic. Knocking out *CG32165* causes complete male sterility, but significantly increased female fertility, while knocking out *CG32164* causes complete female sterility, but significantly increased male reproductive output. Sterility is caused by defects late in oogenesis and spermatogenesis. We discuss these results in the framework of meiotic drive and sexual antagonism and suggest a model by which new genes like *CG32165* can quickly become essential.

3.0.3 Results

CG32165 and *CG32164* Are Both Essential For Reproduction

We recapitulated the results of constitutive *CG32164* or *CG32165* RNAi by knocking out these genes using the CRISPR/Cas9 system (Figure 3.3). Consistent with our RNAi results, ∼20% fewer *CG32165*^{null} homozygotes survive to adulthood than expected, while no lethal effect was observed in *CG32164*^{null} homozygotes (Figure 3.3). There was no difference in lethality between *CG32165*^{null} males and females (Welch's t test $p > 0.05$ in all cases). Furthermore, we did not observe any reduction in survival in heterozygotes, suggesting that *CG32165*^{null} mutations are recessive (not shown). The lethal effect of *CG32165* knockout is rescued by introgression into one *CG32165*^{null} line of a wild-type copy of *CG32165* inserted into attP40 on chromosome 2L, showing that the deletion induced in *CG32165* is the cause of the lethality (Figures 3.3, A.5, and A.6). Thus, *CG32165* is essential for proper fly development (Chapter 2).

Clearly, most or all homozygous *CG32165*^{null} or *CG32164*^{null} flies survive to adulthood. We next tested whether loss of *CG32164* or *CG32165* function affected male or female reproductive output by crossing individual mutant flies to two flies from of common wild-type strain. Surprisingly, all homozygous *CG32165*^{null} males are sterile (Figure 3.3; Table A.10). There was no significant increase in female sterility in these same lines (FET $p = 0.37$), and in fact fertile *CG32165*^{null} females produced significantly more offspring relative to controls (Figure 3.3).

Conversely, almost all homozygous *CG32164*^{null} females are sterile. *CG32164*^{null} males exhibit no increase in sterility rates, but instead sire significantly more offspring than controls (Figure 3.3; Table A.10). Like the *CG32165*^{null} phenotypes, *CG32164*^{null} phenotypes are rescued by insertion of a wild-type *CG32164* copy into chromosome 2L, showing that the

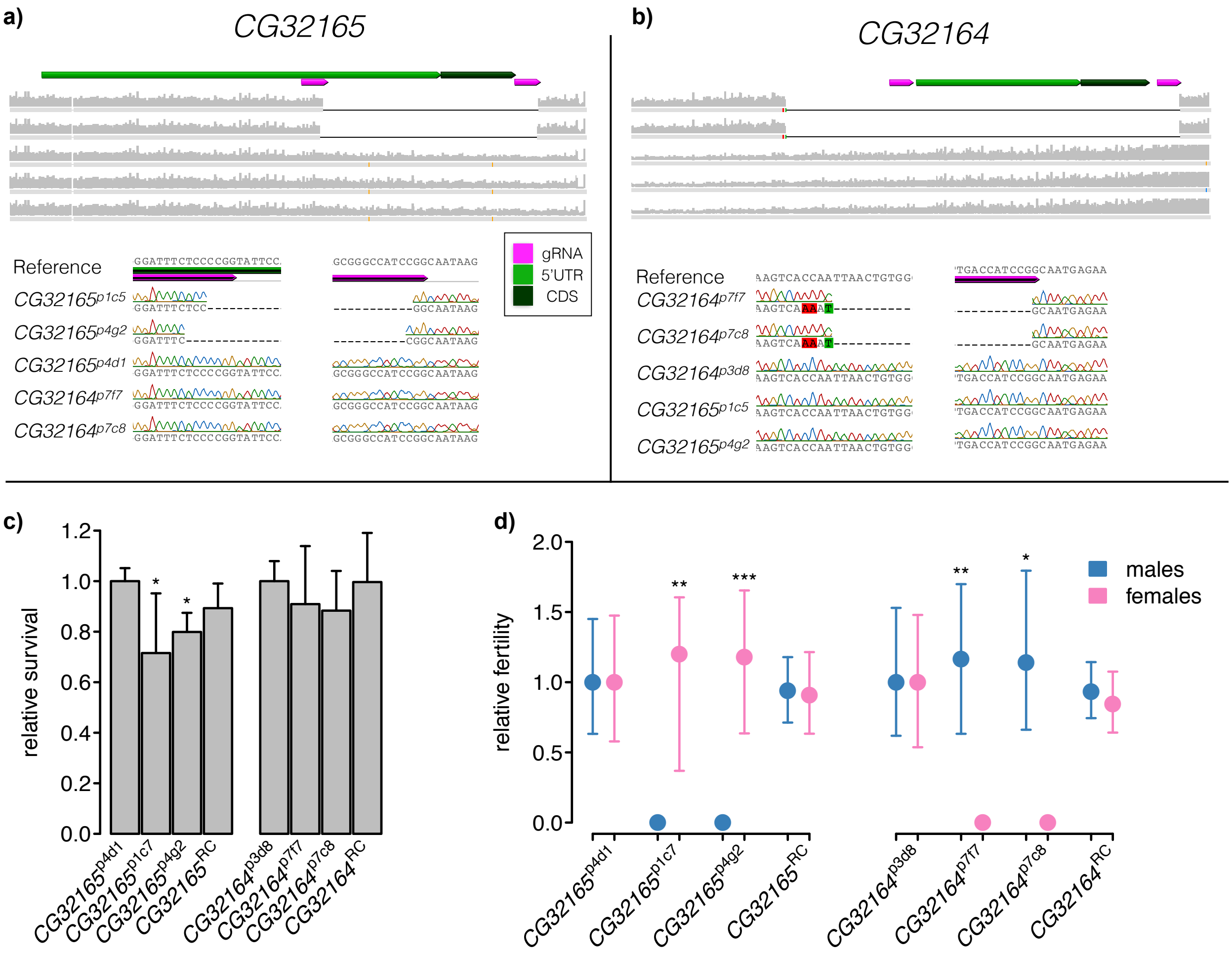


Figure 3.3: CRISPR/Cas9 knockouts of *CG32164* and *CG32165* and their effects on survival and fertility. **a)** *CG32165* deletions encompass exon 1 and most of the 5' UTR. Superscripts indicate independent lines. **b)** *CG32164* deletions encompass the entire 5' UTR and exon 1. **c)** *CG32165* deletions cause lethality while *CG32164* deletions do not. RC: rescue construct. **t*-test $p < 0.05$. **d)** Male and female fertility in deletion lines and controls. Raw offspring counts are shown relative to non-deletion lines (*CG32165*^{p4d1} and *CG32164*^{p3d8}, respectively). Means and 95% CI are shown. Each point and bar summarizes the number of progeny produced by individual test flies crossed to two BDSC 54590 flies in at least 30 replicate crosses. *t*-test $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

disruption of *CG32164* is the cause of the defect (Figure 3.3). Furthermore, the female sterility effect exists regardless of which strain *CG32164*^{null} females are mated to, suggesting it is not a paternal effect [142].

Thus, *CG32164* and *CG32165* are both essential for fly reproduction and both genes are sexually antagonistic. That is, *CG32164* is beneficial for females but detrimental to males, while *CG32165* is beneficial to males but detrimental to females.

CG32165 Knockout Disrupts Late Spermatogenesis

Drosophila spermatogenesis proceeds in a linear order from the germ cells at the apical tip of the testis to mature sperm at the base [143]. After an initial self-renewing germ cell division, the cell destined to produce mature sperm (gonialblast) undergoes four mitotic divisions and two meiotic divisions with incomplete cytokinesis to produce a cyst of 64 interconnected spermatids. Crucially, the large, bulky histones bound to DNA are then replaced with small, acidic protamines, substantially reducing the volume occupied by chromatin and allowing tight packing of the male DNA into sperm heads [143]. After this histone-to-protamine transition (HPT), spermatids are invested in their own membranes and shed excess cytoplasm and waste, producing mature sperm which are coiled and transferred to the seminal vesicle for storage.

CG32165^{null} mutants fail to produce mature sperm (Figure 3.4). We never observed any sperm in mutant seminal vesicles ($n = 87$), which were small and withered, even in old unmated males that typically accumulate unused sperm in the seminal vesicle and the base of the testis. The overall morphology of mutant testes was similar to controls, except spermatogenesis appears to fail prior to individualization (Figure 3.4) [144]. That is, we never observed normal late canoe-stage nuclei. This defect was not observed in *CG32165*^{RC} or *CG32164*^{null} males. *CG32165*^{null} mutations apparently affect all developing spermatids around the HPT. Interestingly, several new *D. melanogaster*-specific duplicate genes, including *tHMG1* and *ProtB*, are also involved specifically in the HPT [122].

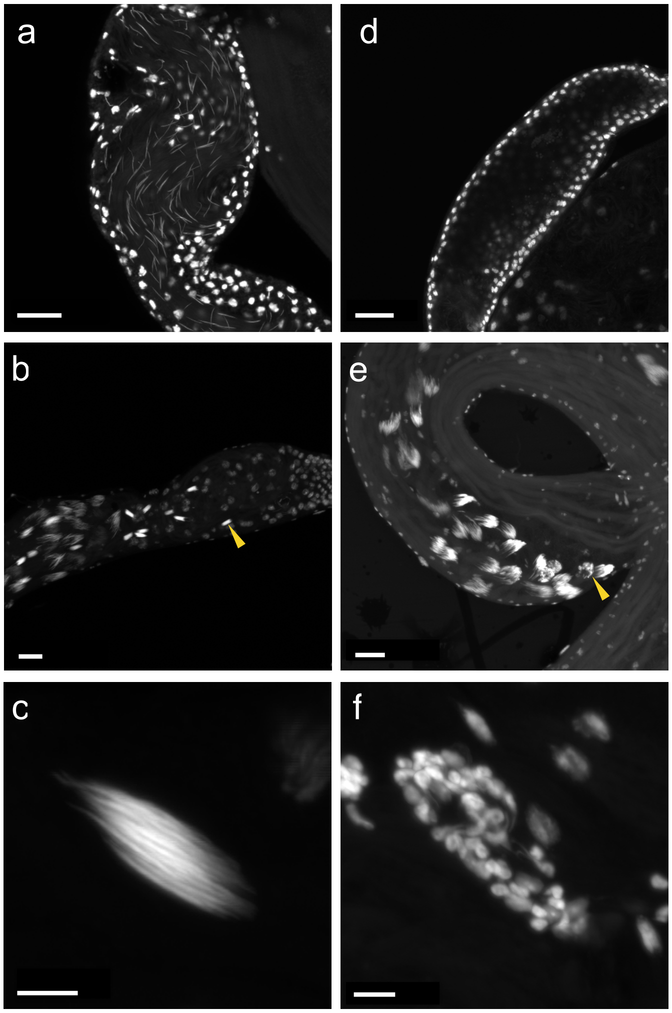


Figure 3.4: *CG32165* knockout causes spermatogenesis to fail near the histone-to-protamine transition. All images are DNA (DAPI) stains of testis or seminal vesicles from 10-day old mated males. **a)** A seminal vesicle from *CG32165*^{RC} male containing mature sperm. Large nuclei are the seminal vesicle somatic cells. **b)** The proximal third of a *CG32165*^{RC} testis, showing development proceeding from left to right. Maturing cysts are clearly visible, and proceed through the histone-to-protamine transition to produce dense bundles of spermatids just prior to individualization (e.g. the cyst indicated by the arrow). **c)** A single *CG32165*^{RC} cyst just prior to individualization, similar to the cyst indicated by an arrow in **b**. **d-e)** The same features as **a - c**, but in *CG32165*^{p4g2} mutant males. **d)** *CG32165*^{p4g2} mutants fail to produce mature sperm. **e)** Spermatid nuclei appear to progress normally through spermatogenesis until around the HPT, then revert to a globular form (arrow, similar to the cyst shown in **f**). Scale bars: **a,b,d,e** 25 μm; **c,f** 5 μm.

Oogenesis In *CG32164*^{null} Flies Is Abnormal

The cause of sterility in *CG32164*^{null} mutants is less clear. *CG32164*^{null} eggs remain rounded and have shortened, thick anterior filaments, suggesting failure earlier in oogenesis (Figure

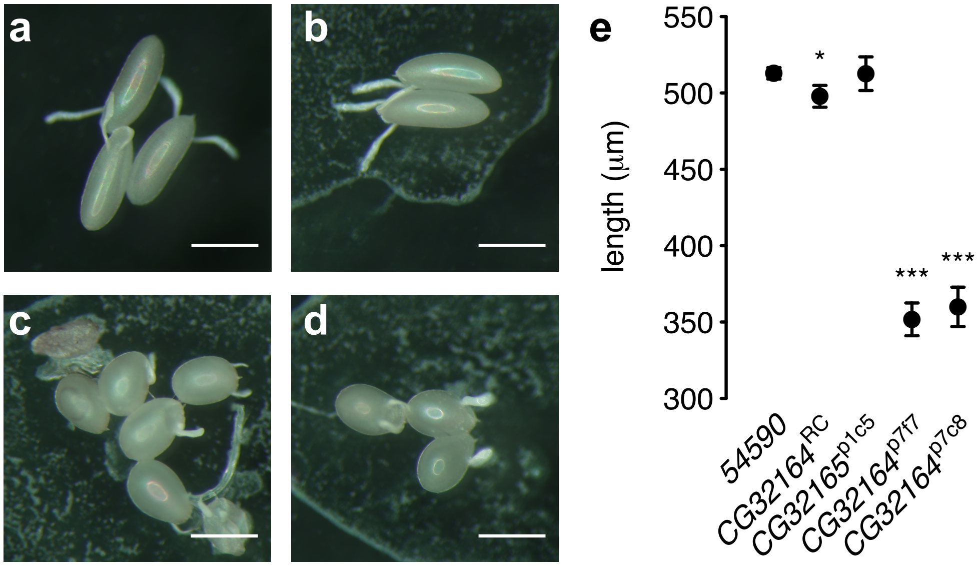


Figure 3.5: *CG32164* mutant eggs are rounded. **a)** Stock 54590 (wild-type), **b)** $CG32164^{RC}$, **c)** $CG32164^{p7f7}$ mutant, and **d)** $CG32164^{p7c8}$ mutant embryos. Scale: 300 μm. **e)** Quantification of embryo anterior-posterior length. Means ± 2 SEM are shown. Wilcoxon Rank Sum test results comparing 54590 vs. mutants shown above bars. $*p < 10^{-3}$; $**p < 10^{-9}$

3.5). Interestingly, the estimated volumes of mutant eggs, calculated as $\frac{\pi}{6}$*width^2*length, are not different from wild-type eggs (Welch's t-test $p > 0.05$ for all comparisons), suggesting that *CG32164* may exert its effect in the somatic cells that influence egg development (e.g. follicle cells; S. Horne-Badovinac personal communication). This egg length defect is not seen in *CG32165* mutants or in *CG32164* rescue lines. We do not know if these embryos were successfully fertilized, so the egg shape defect may not be the cause of the sterility. However, sterility occurs regardless of the genotype of the males that $CG32164^{null}$ flies are mated to, suggesting that it is an oogenesis defect rather than a paternal effect.

CG32164 and *CG32165* Expression Divergence

Overall, *CG32164* and *CG32165* fertility defects are consistent with their expression patterns across tissues and development (Figure 3.6). *CG32165* is moderately expressed in L3 larval imaginal discs and testis and only lowly in ovary, while *CG32164* is moderately expressed in L3 imaginal discs and ovary but lowly in testis. Furthermore, the *D. simulans* ortholog appears to be broadly expressed in female bodies, and especially female head, so there has likely been significant expression pattern divergence both since the speciation event and since the duplication event that likely contributed to the functional divergence we see between the duplicates now (Figure 3.6). Even though the gene pair was generated within the last 2 million years by tandem duplication, the Spearman Rank correlations in their expression patterns in modENCODE data are only 0.64 ($p = 0.0017$) across tissues and 0.42 across development ($p = 0.059$) [95].

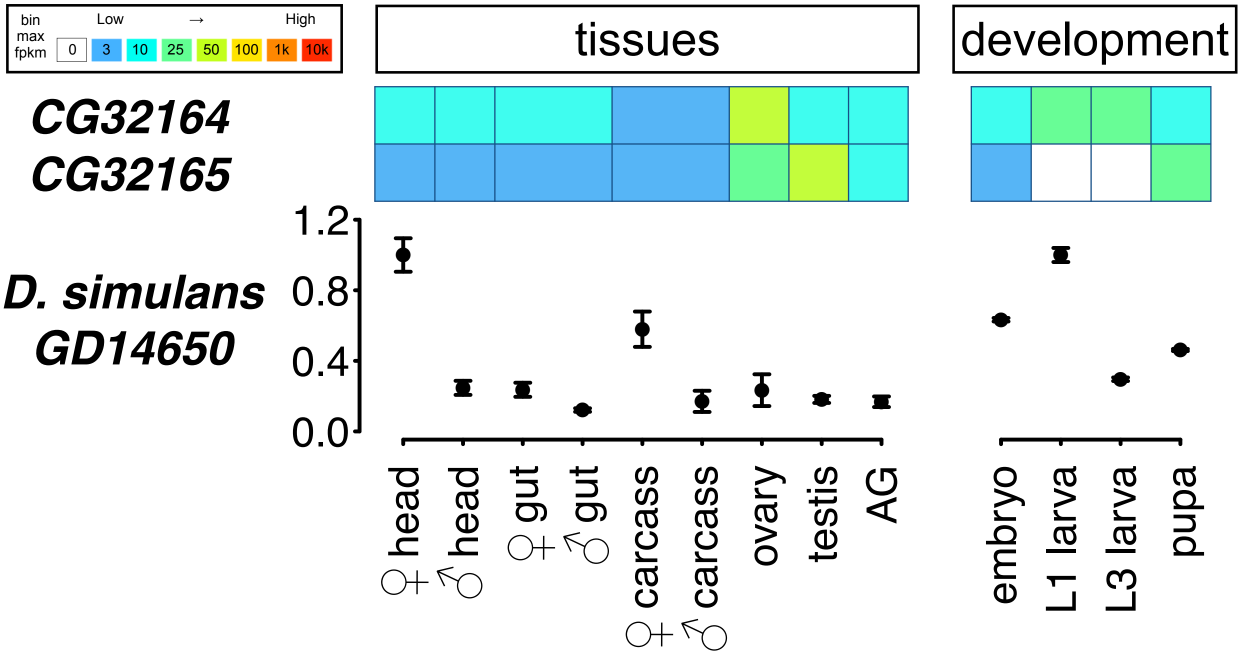


Figure 3.6: Expression pattern divergence between *CG32164*, *CG32165*, and their *D. simulans* ortholog, *GD14650*. *GD14650* expression levels are calculated relative to *RpL32* as $2^{-(C_{T,G}-C_{T,R})}$, where $C_{T,G}$ and $C_{T,R}$ are the threshold cycles of the target gene and *RpL32*, respectively. They are then divided by the maximum expression detected for that gene among tissues or development to highlight expression relative to maximum detected expression. Means and 2 SEM from three biological replicates are shown. *CG32164* and *CG32165* expression levels are taken from FlyBase 2015_05 [95]. Only similar datasets are shown.

This expression pattern divergence will likely have led to a significantly different set of proteins that the two genes are able to interact with. In fact, the genes share no direct interaction partners in a gene co-expression network constructed using modENCODE data (Chapter 2). The fact that the duplicates are expressed in the same tissues or developmental stages, but at different levels, may explain why they cannot compensate for each other's loss. However, it is still unknown which particular tissue or set of tissues the two genes actually exert their effects in.

CG32165 and *CG32164* sequences evolved relatively rapidly since duplication

CG32164 and *CG32165* may have identical molecular functions but act in different developmental stages or tissues, or they may have distinct molecular functions. To begin to understand if the *CG32164* and *CG32165* molecular functions have functionally diverged, we investigated patterns of amino acid substitutions in the proteins through time.

Neither gene has accumulated amino acid substitutions significantly faster or slower than expected for neutrally evolving genes according to maximum likelihood analyses of nonsynonymous and synonymous substitution rates (d_N and d_S) [100]. *CG32165* and *CG32164* d_N/d_S estimates are similar (likelihood ratio (LR) = 0.06, $p = 1$) and less than one, but both values are significantly higher than the background rate (*CG32165* $d_N/d_S = 0.750$, LR = 4.77, 1 d.f., $p = 0.002$; *CG32164* $d_N/d_S = 0.685$, LR = 2.18, $p = 0.019$). There is thus no evidence for asymmetric rates of sequence divergence between the two copies. Furthermore, neither gene has accumulated a significant excess of nonsynonymous substitutions according to McDonald-Kreitman tests (FET *CG32165* $p = 0.19$, *CG32164* $p = 0.39$) [101]. Finally, d_N/d_S of the pre-duplication, ancestral gene is estimated to be 0.226, suggesting it did not evolve rapidly in the period leading up to the duplication event, which is expected in some selection-driven subfunctionalization models [20, 21, 23]. Altogether, these results suggest that the rate of amino acid substitution has increased in both *CG32165* and *CG32164* since duplication, but whether this apparent increase is due to relaxed constraint or positive selection is unknown.

As mentioned previously, *CG32164* and *CG32165* reside in the middle of a large selective sweep in a region of average recombination (∼2.5 cM/Mb; [145]), suggesting that at least one of the duplicates, or a variant very tightly linked to them, was the target of selection. Constitutive knockdown of *CR18217*, a chimeric pseudogene formed by the *CG32164/CG32165* tandem duplication, had no effect on fly survival or fertility (not shown), suggesting it is likely not the target of selection.

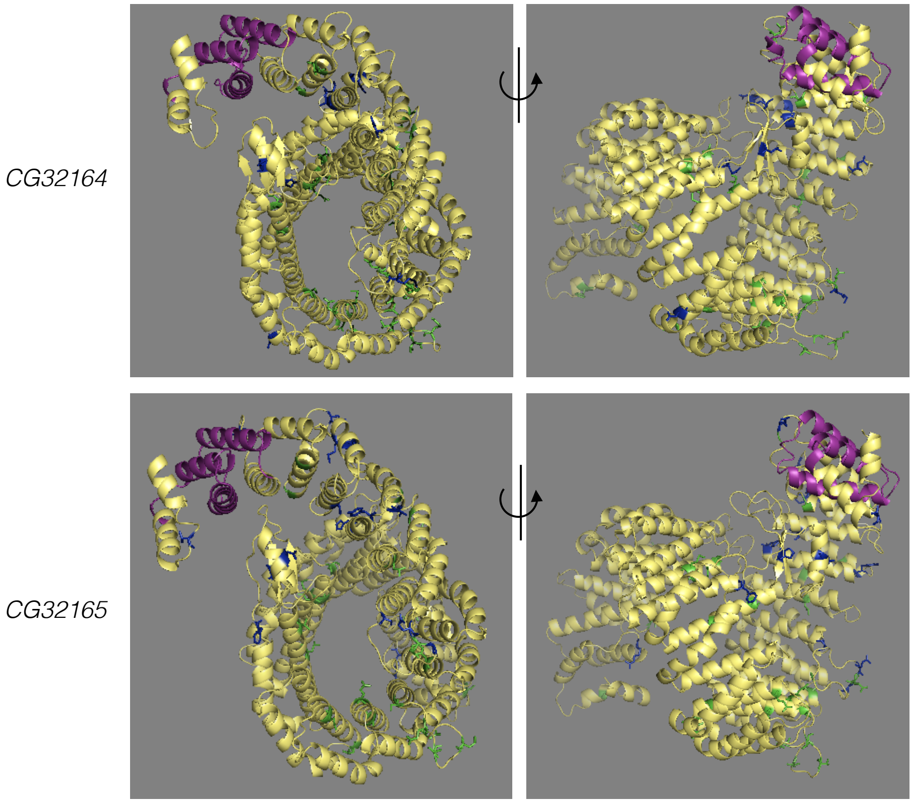


Figure 3.7: Predicted protein structures of *CG32164* and *CG32165*. Structures were predicted using the RaptorX web server and the consensus protein sequences for *CG32164* and *CG32165* (Methods) [146]. The two genes, like all importin-βs, form α helix-rich right-handed coils. Green residues are those that substitutions that are shared between the two copies and therefore arose in the pre-duplication ancestor. Blue substitutions are those that are specific to that gene copy. The purple region is the canonical importin-β motif. Structures were analyzed in PyMol [147].

CG32164 and *CG32165* are importin-β proteins. Importin-βs have diverse roles within the cell, including the regulation of mitotic spindle assembly, directional import of proteins into the nucleus, and nuclear envelope assembly, and each of these roles are regulated by the nucleus-to-cytoplasm gradient of the Ran GTPase bound to GTP (Figure 3.8) [148]. Thus, the ability to bind Ran is therefore crucial for importin-β function [149, 150]. The importin-β domain that binds Ran is highly conserved among eukaryotes, and we reasoned that mutations that disrupt the importin-β domain may interfere with Ran binding. The pre-duplication ancestor accumulated 44 amino acid substitutions between the *D.simulans/D. melanogaster* split and the duplication event, *CG32165* specifically accumulated 16, and *CG32164* specifically accumulated 8, yet none of these substitutions has occurred in the importin-β domain (Figure 3.7). Instead, substitutions are scattered throughout the proteins and tend to maintain the site polarity and hydrophobicity (Figure 3.7). Interestingly, while only 14% (6/44) of amino acid changes shared by the duplicates were drastic (i.e. changed the charge or hydrophobicity of the site), 44% *CG32165*-specific (FET $p = 0.0287$) and 38% *CG32164*-specific (FET $p = 0.13$) substitutions are drastic. Whether this accumulation of potentially deleterious substitutions is due to relaxation of selection after duplication or a sign of functional differentiation is unclear. In any case, there are no obvious differences in *CG32164* and *CG32165* protein sequences that suggest they perform different molecular functions.

Altogether these results suggest that functional divergence between the two copies was primarily driven through expression pattern divergence. The 5' ends and first introns have a multitude of structural differences that may influence this expression pattern divergence (Figure 3.1). However, further experiments are needed to determine if this hypothesis is true (see Experiment #1 below).

3.0.4 Discussion

Here we have shown that both copies of a species-specific duplicate gene pair are essential for fly fertility. RNAi and specific gene knockouts provide strong evidence that, in addition to *CG32165*'s effect on fly survival, *CG32165* and *CG32164* are required for male and female fertility due to their action in testis and ovary, respectively. Intriguingly, the gene pair appear to be sexually antagonistic, as *CG32164* knockout also results in significantly increased male fertility while knocking out *CG32165* significantly increases female fertility. The pair likely perform a similar molecular function, but at different developmental stages and tissues, considering the functional regions of the proteins have been preserved.

CG32165 Is An Essential Species-Specific Gene

CG32165 has greater sequence and structural divergence from its *D. simulans* and *D. yakuba* orthologs than *CG32164* and, consistent with any definition of young genes, can be called the new copy of the duplicate pair (Chapter 2) [35, 45, 63, 80]. Furthermore, *CG32165* has qualitatively higher expression in the testis relative to the *D. simulans* ortholog and *CG32164*, a phenomenon observed for young genes in organisms from mammals to flies (see refs [34, 56] for reviews). This novel expression pattern explains how *CG32165* is able to affect spermatogenesis, but not the cause of the sterile phenotype. Why is *CG32165* essential for fly development, while *CG32164* is not? It is not possible with our data to distinguish between neofunctionalization and subfunctionalization models, but it is difficult to explain how a novel essential function could evolve so quickly. It seems more plausible that *CG32165* is essential because it retained an ancestral essential function, probably expression in a particular developmental stage and/or tissue, while *CG32164* did not. However, functional tests of the ancestral, pre-duplication gene copy are needed to distinguish between these types of models (more below).

CG32165 Knockout Causes Sterility by Interfering With the Middle Stages of Spermatogenesis

Many examples of young, male-beneficial genes have been identified in *Drosophila*. Ding and her colleagues found that knocking out *nsr*, a ∼10 million year old duplicate gene in *D. melanogaster* and its close relatives, causes male sterility because spermiogenesis fails [62]. The *Sdic* chimeric gene family formed by tandem duplication specifically in *D. melanogaster*, yet significantly improves the competitiveness of sperm by ∼1.5% [118, 119]. And, like the other *D. melanogaster*-specific genes *CG32165*, *ProtB*, and *tHMG1 Sdic* resides in a recent and strong selective sweep. Interestingly, *ProtB*, *tHMG1*, and *CG32165* all appear to function during the histone-to-protamine transition (HPT) [121, 122]. In fact, *ProtB* is a protamine and *tHMG1* (another species-specific essential gene, Chapter 2) is a chromatin remodeling factor [121, 122]. Thus, it is possible that these *D. melanogaster*-specific genes function together in a species-specific spermatogenesis network, as well as development (Chapter 2). *Mojoless* and *nsr*, for example, ∼35 and 6 million year old genes, are both essential for normal spermatogenesis and male fertility [62, 151]. Loppin et al. (2005) showed that *K81* is a gene with paternal effect: eggs fertilized with sperm from *K81* mutant males fail to complete the first round of nuclear division [142].

CG32164 Knockout Causes Female Sterility by Disrupting Oogenesis

CG32164^{null} females produce rounded eggs with normal volumes, suggesting that the defect in oogenesis is caused by a defect in egg chamber elongation rather than a defect in oocyte maturation and loading.

Null mutations in *Ketel*, a well-studied importin-β protein, cause females to produce eggs with thin chorions [152], while dominant mutations that interfere with Ran binding cause pronuclear fusion and nuclear divisions to fail extremely early in embryogenesis [150, 153].

Three Reasonable Hypotheses Can Explain *CG32164*'s and *CG32165*'s Evolution

CG32164 and *CG32165* are both sexually antagonistic. Why is *CG32164* detrimental to male fitness and *CG32165* detrimental to female fitness? Both copies are expressed in the gonads of both sexes, which probably explains how they affect the fitness of both sexes. Our results suggest at least three plausible alternatives to explain the effects on reproduction of *CG32164* and *CG32165*. All four hypotheses require a test of the function of the ancestral, pre-duplication gene to distinguish between them, and cannot be definitively picked between here. We will highlight the main data that support each hypothesis and the data that are still needed to accept them. Two experiments would particularly help to distinguish between the following models.

Experiment #1: To test if expression pattern differences are the cause of functional differentiation between *CG32164* and *CG32165* the coding sequences of the two genes could be swapped to place *CG32165* expression under the control of *CG32164*'s regulatory sequences and vice versa. If expression pattern divergence drives functional divergence, then we expect that *CG32164* knockout can be rescued by *CG32165* expression under the control of *CG32164* regulatory sequences. No rescue should be observed if the functional difference is caused by protein sequence divergence.

Experiment #2: To test the essentiality of the ancestral, pre-duplication gene would definitively show whether the essential functions of *CG32164* and *CG32165* are novel or whether they are subsets of the ancestral functions. This experiment can be performed by knocking out *GD14650* in *D. simulans* only if we assume that the ortholog represents the ancestral state. Alternatively, ancestral state reconstruction may be used to reproduce the *CG32164/CG32165* ancestor and test its ability to rescue extant gene knockouts. However, the pitfall of this experiment is that the expression pattern likely matters and it is unclear how to handle this. In either experiment, if the ancestral gene is essential for the fertility of both sexes, then we would conclude that these genes have evolved according to a subfunctionalization model. If the ancestral gene shows no such effects, or an effect on one sex, then we would have evidence of neofunctionalization.

Model #1: Neofunctionalization

The first hypothesis that can explain *CG32164* and *CG32165* evolution is therefore that *CG32165* underwent classic neofunctionalization, which is predicted to be common in large populations like *D. melanogaster*'s [74]. *CG32165* appears to have gained high expression in testis, but mainly lost expression in other tissues and developmental stages relative to *CG32164* and *GD14650*. Experiment #2 above will help determine if this is a viable model.

Model #2: Meiotic Drive

CG32165 may have been rapidly driven to fixation by meiotic drive, the non-adaptive but biased transmission of particular alleles. Overall, our results bear a striking resemblance to the *D. melanogaster Segregation Distorter* (*SD*) system, probably the best-characterized meiotic drive system (see ref. [154] for review). Heterozygous males transmit *SD* chromosomes 65% - 95% of the time, instead of Mendel's expected 50% [155]. Distortion is caused by a partial tandem duplication of *RanGAP* (*Sd-RanGAP*). *Sd-RanGAP* produces a functional enzyme lacking a site that typically enables tethering to the outside of the nuclear envelope and maintenance of the RanGTP gradient because RanGAP (and Sd-RanGAP) stimulate RanGTP to hydrolyze GTP to GDP (Figure 3.8) [156, 157]. Without its tether, Sd-RanGAP enters the nucleus of the cell, distorts the RanGTP gradient, and by some unknown mechanism, causes preferential transmission of *SD* chromosomes [154, 156, 157]. In fact, *SD* chromosomes cause non-*SD* chromosomes to fail to proceed through the HPT, similar to *CG32165* knockout [158].

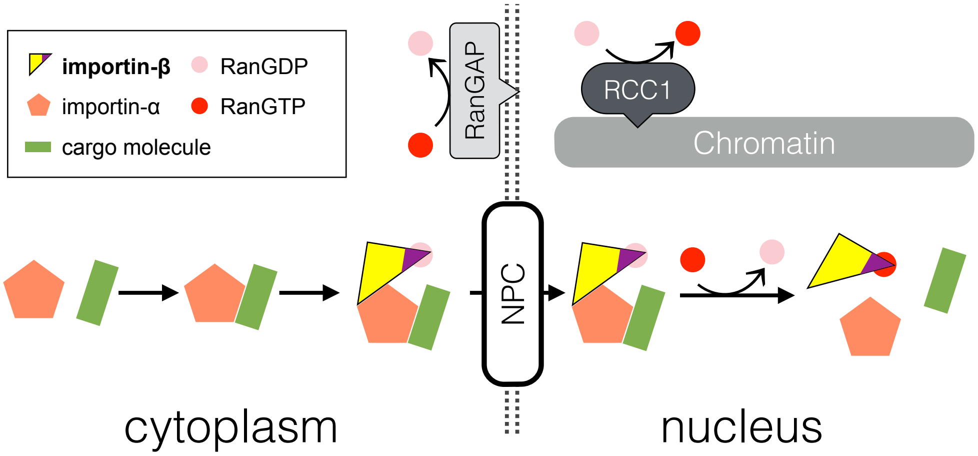


Figure 3.8: The RanGTP gradient determines importin-β function. The Ran GTPase activating protein (RanGAP) and Ran guanine nucleotide exchange factor (RCC1) act to maintain a RanGTP gradient between the nucleus and cytoplasm which drives directional, importin-mediated nuclear import through the nuclear pore complex (NPC). Importin-βs have an N-terminal Ran-binding domain (purple), a middle domain that interacts with the NPC, and a C-terminal importin-α/cargo binding domain. Some importin-βs also bind directly to nuclear localization signal sequences in cargo proteins without importin-α mediation.

The nuclear import machinery, including nucleoporins [159], nuclear transport factors, and even Ran duplicates [160, 161], frequently exhibits signs of recent and strong positive selection. These observations have led to the hypothesis that the nuclear import pathway, including many proteins that importin-βs directly interact with, is susceptible to genomic conflicts caused by meiotic drive [162]. The diverse roles of importin-βs in nuclear import and cell division provide plenty of potentially essential functions that may explain the lethal and sterile effects of *CG32165* and *CG32164*. Furthermore, strong meiotic drive will produce patterns of diversity reduction identical to selective sweeps. However, alleles fixed by male meiotic drive should not increase male fertility, whereas alleles fixed by natural or sexual selection might be expected to do so. Work in mice has led to the suggestion that such meiotic drive systems may potentiate more extensive divergence and even speciation by rapidly driving genetic divergence between populations, and that non-adaptive yet rapid evolution of recombination hotspots in mice can lead to the evolution of reproductive incompatibilities between populations in short periods of time (<1 million years) [163, 164].

Model #3: Male-Specific Selection

CG32165 is required for male reproduction, and, as we discussed above, many young Drosophila genes have similarly strong male-specific effects [61, 62, 119, 142, 151]. These results and more from mammals suggest that new gene origination is frequently driven by the male-specific benefit of new genes [56].

The difficulty with all of these studies and ours is that they are all loss-of-function assays. Therefore, we do not know if the gene itself is actually beneficial, only that its removal causes a phenotype. Though it is complicated by the fact that the expression pattern may not be comparable, one test of the male benefit of the new gene would be to insert it into an outgroup species like *D. simulans*. We would expect transgenic males to have higher fertility.

Model #4: Resolution of Sexual Conflict

A closely-related and final model to describe *CG32164* and *CG32165* evolution is that the duplication was favored because it allows the resolution of intralocus sexual conflict (SC) [165–170]. Males and females of sexually dimorphic species share the majority of their genomes but have different fitness optima. Thus, alleles beneficial to one sex may be detrimental to the other, and there is a conflict between male-specific and female-specific selective pressures. Detrimental effects are typically resolved by sex-biased gene expression, but resolution may often be prohibited by pleiotropic effects of changing gene expression patterns [171]. Theory predicts that gene duplication can relieve intralocus sexual conflict by allowing each copy to increase expression in the benefitted sex and reduce expression in the harmed sex while avoiding the pleiotropy problem [169, 172]. Such resolution is not immediate, but rapid divergence in expression pattern and therefore relief of strong SC should occur shortly after duplication occurs [172].

Several observations make us believe that SC resolution has governed *CG32164* and *CG32165* evolution. Obviously, *CG32164* and *CG32165* are sexually antagonistic. Furthermore, they exhibit symmetric amino acid divergence rates, but apparently asymmetric expression pattern divergence in which *CG32165* specifically increased testis expression while *CG32164*'s expression remained more constant. After duplication, the copies would be free to optimize their male and female effects separately and, by chance partitioning of expression patterns, *CG32165* kept the gene-gene interactions necessary for the ancestral gene's essential function during development. The chance partitioning of essential functions into one copy or other would also explain why there is a relatively constant fraction of essential genes of different ages [63]. We hypothesize that *CG32164/CG32165* SC is simply not yet fully resolved. This hypothesis can be tested if we assume that the *D. simulans* ortholog's expression pattern is similar to the expression pattern of pre-duplication *CG32164/CG32165* gene. We predict then that *D. simulans GD14650*, which is single-copy, is essential for fertility of both sexes.

This model is closely related to model #3 because the male-specific selective pressure, and therefore sexual conflict, may have existed before the duplication and provided the initial benefit to duplicate the *CG32164/CG32165* ancestor.

Detailed investigations of the fitness effects and molecular evolution of the remaining essential species-specific genes and their parents may help distinguish between whether new duplicates sweep to fixation because 1) they have a novel beneficial function, 2) there is strong meiotic drive, 3) they specifically improve male fitness, and/or 4) they allow resolution of sexual conflict.

3.0.5 Methods

CRISPR/Cas9-Induced *CG32165* and *CG32164* Knockout

We induced large deletions (180 bp and 700 bp, respectively) encompassing the *CG32165* or *CG32164* translation start sites using the CRISPR/Cas9 system generally following Bassett and Liu 2014. We designed gRNAs using the FlyCRISPR Optimal Target Finder with high stringency, choosing only gRNAs with no predicted off-target cut sites [174]. We synthesized gRNAs following Bassett and Liu 2014, and injected gRNAs (500 ng/μL each) into Bloomington Drosophila Stock Center strain 54590 [175]. Strain 54590 expresses Cas9 ubiquitously under control of the *Act5C* promoter. Mutants were isolated by crossing to BDSC strain 4534 (w*; $\frac{Sb^1}{TM3,P\{w^{+mC},Act-GFP\},Ser^1}$). Both gene copies were PCR amplified and sequenced to ensure that only the target copy contained a deletion. F1 flies without deletions in *CG32164* or *CG32165* were used as controls in fertility and lethality experiments.

Rescue Line Construction

CG32164 or *CG32165* gene regions plus ~750 bp up- and downstream were PCR amplified and inserted into a backbone carrying a *vermilion* marker (pVerm) before injection into BDSC strain 25709. Transformants were isolated by crossing to y^1v^1; $\frac{sna^{Sco}}{CyO}$ balancer strain (provided by the Harvard Transgenic RNAi Project's R. Binari). The appropriate rescue chromosome was introgressed into $CG32165^{p4g2}$ or $CG32164^{p7f7}$ with the aid of BDSC stock 7198 ($\frac{Kr^{lf-1}}{CyO}$; $\frac{D^1}{TM3,Ser^1}$). These rescue lines were used in subsequent fertility and lethality assays. See Figures A.5 and A.6 for additional details.

CG32165 and *CG32164* Knockout Fly Lethality and Fertility Effects

Twenty balanced flies of each sex were crossed in bottles in at least triplicate. The proportion of F1s without a balancer (i.e. homozygous for the deletion) for knockout lines was compared to the proportion from controls to estimate egg-to-adult lethality. We tested the effect of *CG32165* or *CG32164* knockout on fly fertility by crossing individual 3 - 5 day old mutant, rescue, or control flies to two 54590 flies. Flies were allowed to mate for 9 days in a 25°C incubator with 12h:12h light:dark cycle. Parents were then removed and total F1 progeny counted 18 days after cross setup. At least 30 crosses were used for each line (Table A.10).

Testis and Egg Staining and Microscopy

Testis were fixed with methanol/acetone and stained with DAPI following Bonaccorsi et al. (2011,2012) before visualization using a Zeiss LSM 710 confocal microscope [176, 177]. Embryo lengths were quantified in ImageJ using the maximal distance from anterior-to-posterior tips [178].

Population Genomic Data and Analyses

McDonald-Kreitman tests and maximum likelihood analyses were carried out using methods and polymorphism and divergence data described in Chapter 2 (page 30). We used PAML4.7a [100] to estimate relative rate ratios. We used the longest isoforms of *CG32164* orthologs defined in FlyBase release 2015_05 from *D. ananassae*, *D. simulans*, *D. sechellia*, *D. yakuba*, and *D. erecta* as outgroup sequences. We used the majority consensus sequences of *CG32165* and *CG32164* coding sequences from 17 DPGP2 RG genomes as representatives for those genes [83]. We estimated ω under a free-ratio model and also specifically tested whether *CG32165* or *CG32164* have evolved at different rates than their orthologs by comparing two-ratio models allowing *CG32165* (or *CG32164*) to evolve with a different ω than its orthologs to a one-ratio model constraining all branches to evolve under the same ω value. Likelihoods were compared using a likelihood ratio test with one degree of freedom.

CHAPTER 4

AN INVESTIGATION OF SELECTION ON INDIVIDUAL TANDEM DUPLICATIONS SEGREGATING IN *DROSOPHILA MELANOGASTER*

4.0.1 Abstract

All genes in a species' genome originated as a mutation in a single germ cell which then spread to all members of the species. Understanding the evolutionary forces governing this phase of new gene origination is crucial for understanding precisely when new genes become beneficial and therefore destined to be fixed and maintained in that species' genome over some period of evolutionary time. Theoretical models describe idealized ways in which new genes fix, while empirical studies show that the vast majority of genes segregating in populations are under strong purifying selection. To begin to reconcile these two ways of thinking about the polymorphic phase of new gene origination, I investigated the presence, structures, expression, and evolutionary forces acting on new duplicate genes segregating in *D. melanogaster*. Consistent with previous studies of *D. melanogaster* gene copy number variation (CNV), I find thousands of partially or completely duplicated genes segregating in a large African population of *D. melanogaster*. Ten percent of duplications completely duplicate at least one gene, but an additional 23% partially duplicate genes to form potentially novel gene structures, many of which are expressed. Analyses of nucleotide diversity and haplotype length around CNVs suggest that a small fraction have risen in frequency, probably due to positive selection. I discuss the implications of these results, future experiments, and ways to improve the power of these methods.

4.0.2 Introduction

New genes are those that were formed recently in evolutionary time and are therefore found in one group of closely related organisms but absent in all others. The initial stage of the process of new gene origination is a population genetic process in which a mutation in a germ cell in a single individual spreads through the population to fixation under the influence of evolutionary forces such as natural selection, genetic drift, and recombination [16, 75]. Most new genes are formed by duplication, and duplicate genes have long been hypothesized to be a significant source of novel gene functions that contribute to adaptation and evolutionary novelty. There has thus been intense theoretical investigation of when and how new beneficial gene functions arise [9, 18, 74, 75, 111, 179, 180].

Classic models of duplicate gene evolution assume that duplicates are functionally redundant long after the duplication event and therefore ignore the fixation phase of new gene origination (e.g. ref. [73]). However, more recent models show that the probability of duplicate retention and even neofunctionalization can be high, particularly in large populations, and the classic assumption of redundancy and neutrality before fixation may rarely be true [74]. In addition, several experimentally-supported models exist describing how new duplicates may evolve under continual positive selection, either because gene duplication is favored to increase gene dosage or because duplication allows copies to independently optimize distinct beneficial functions [20–24].

Conversely, massive amounts of population genomic data have shown that new genes segregating in populations frequently have strong deleterious fitness effects [117, 181–185]. Applications of powerful array and next-generation sequencing technologies in humans, flies, plants, and other organisms have shown that DNA copy number variants (CNVs), which include gene duplications, may account for more genomic differences between individuals than single nucleotide polymorphisms (SNPs) [117, 123, 185–187]. Empirical investigations in humans and flies consistently show that the frequency spectrum of CNVs, including complete gene duplications, is strongly skewed towards low frequencies, suggesting that most new duplicates are quickly purged from the population by strong purifying selection [117, 185, 187]. Furthermore, there are many strong associations between duplications and disease in humans (e.g. see ref. [188] for review and ref. [189]). Thus, there is abundant indirect and direct evidence that new duplicates are often deleterious.

The fact that new genes originate at high rates in diverse lineages, from human (∼25 per million years) to Drosophila (∼10 per million years), shows that some small fraction of newly-formed genes do fix [35, 43, 44, 46, 190]. Furthermore, 62% *D. melanogaster*-specific genes frequently exhibit signs of having recently been strongly selected (Chapter 2). Outstanding questions, then, are which new genes segregating in populations are likely to fix and what characteristics of those genes distinguish them from the vast majority segregating in the population? Comparing genes at the polymorphic stage to those that have fixed may shed light onto these questions.

This chapter was motivated by the observation that 62% of fixed *D. melanogaster*-specific genes (less than ∼2 million years old) reside in recent and strong selective sweeps, suggesting that natural selection often strongly influences the early stages of the evolution of genes that eventually fix in flies (Chapter 2). Whether or not selection acted on these new gene loci before fixation is still not certain, however, because the mean time to fixation for a neutral variant is $4N_e$ generations [191]; for flies, this equates to ∼400,000 years [116]. Thus, these new loci may have drifted to fixation and subsequently accumulated beneficial mutations which caused the loci to sweep. It is therefore still unclear if and how frequently natural selection acts on new genes when they still segregate in the population.

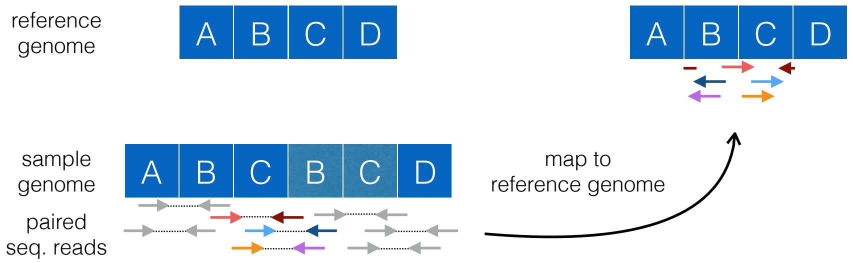


Figure 4.1: Divergently-mapped read pairs and split reads indicate tandem duplications in sample genomes relative to the reference genome. Arrows are sequencing reads and read pairs are connected by dotted lines (which are not sequenced). The reference genome region B-C is tandemly duplicated in the sample genome. Random fragments from the sample genome are sequenced with paired-end next generation sequencing and mapped back to the reference genome. Divergent pairs (←→) indicate the tandemly-duplicated region and split reads (dark red arrow) reveal the precise nucleotide breakpoints.

The goal of the work described in this Chapter was to, first, detect individual duplicates segregating at high frequencies within *D. melanogaster* that have been selected and, second, produce a high quality set of candidate genes with putatively beneficial functions that can be experimentally studied in the lab in order to better understand new gene origination and functional evolution. I investigated the presence, expression, and evolution of tandem duplications (TDs) segregating in *D. melanogaster*. At least 80% of CNVs [117, 187] and new genes that recently fixed in *D. melanogaster* were clearly formed by tandem duplication (Chapter 2) [35, 192]. I therefore identified TDs segregating in the *Drosophila* Population Genomics Project phase 2 (DPGP2) dataset, a collection of 109 deeply-sequenced haploid embryo sequences from individual flies collected around Africa [83]. This sequencing dataset has the significant advantage of being haploid sequences, which allows investigation of variation specifically among chromosomes in which the duplication is present or absent.

4.0.3 Results

TD Caller Specificity, Sensitivity, and Accuracy

Several algorithms have been developed to detect TDs in high-throughput, whole-genome re-sequencing data relative to a reference genome assembly (Figure 4.1) [193–195]. These algorithms can identify duplications with nucleotide resolution, offering a significant advantage in detecting the precise fusion points in chimeric genes over previous methods (Figure 4.1) [117]. I tested the sensitivity, specificity, and accuracy of two algorithms, pindel [193] and delly [195], using simulations based on DPGP2 data characteristics (Methods) [83, 196]. I specifically tested the effects of read depth, minimum supporting read mapping quality, repeat sequences, and the number of required supporting reads. Both algorithms have reasonably high sensitivity, and extremely high accuracy, even with low coverage sequencing (Figure 4.2).

Impressively, repeat regions and mapping quality cutoffs have negligible effects on the true positive ($< 1.7\%$ decrease) and false positive rates ($< 2.3\%$ increase). Previous studies used the intersection of calls from multiple algorithms to generate a 'high-confidence' set of variants [185, 187]. However, including pindel calls here only increases the false negative rate: with 30× mean coverage, delly correctly called 85% of duplications and pindel called 62% of duplications, but only 32% of calls were made by both programs. I analyze precise delly calls for the remainder of this chapter.

Thousands of Tandem Duplications Segregate in the DPGP2

I identified TDs on the major chromosome arms of the release 6 *D. melanogaster* genome in the 109 core, haploid African fly embryo genomes sequenced by the DPGP2 using delly v0.6.5 precise calls [83, 195]. It is critical to note that all TDs are duplications relative to the

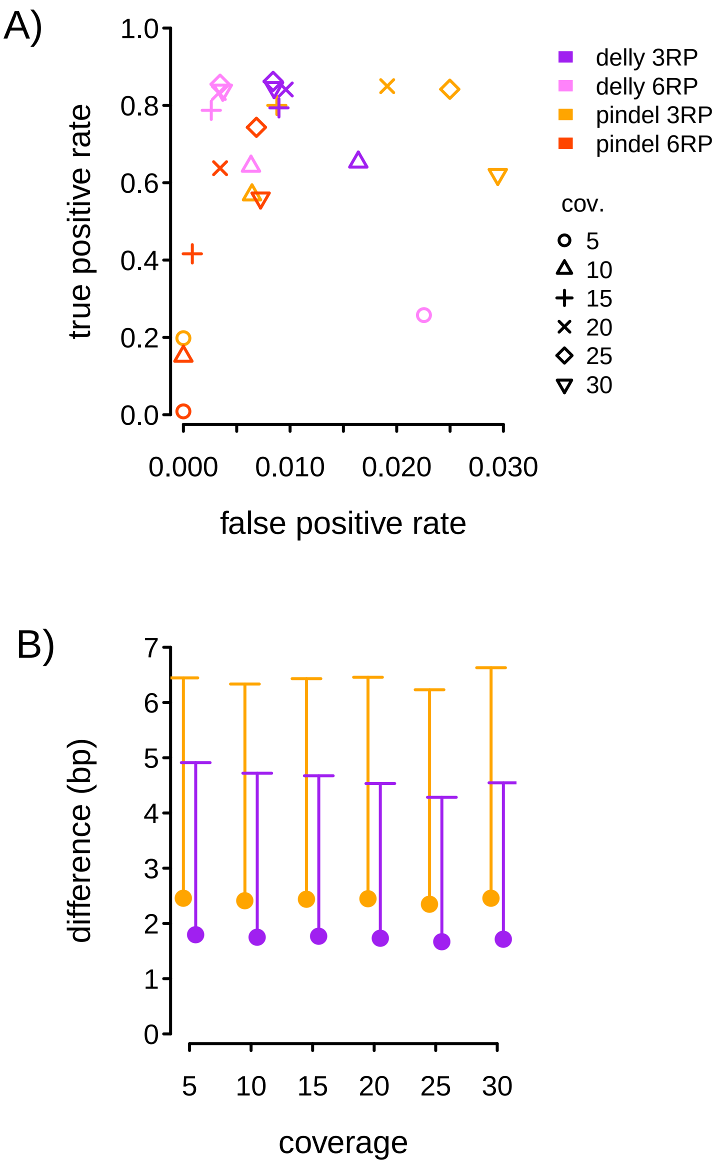


Figure 4.2: Tandem duplication calling algorithm specificity and sensitivity with minimum mapping quality 30. Only calls with at least one supporting split read (i.e. precise, nucleotide resolution calls) were used. A) True and false positive rates for delly and pindel at a range of mean read coverage depths. Results are shown for a minimum of 3 or 6 supporting read pairs (RP). The delly TPR and FPR at 5-fold coverage ○ are 0.31 and 0.10. B) Call accuracy for the two programs at various mean read coverage depths. This is the total absolute distance in base pairs between the call coordinates and the true duplication coordinates. Means plus two standard deviations are shown. Ninety-five percent of delly and pindel calls are within 5 and 7 bp of the true coordinates, respectively, for all coverages.

reference genome and therefore duplications present in the reference will not be identified (Chapter 2).

I find 7,068 TDs segregating in the 109 genomes at a mean (median) frequency of 0.055 (0.009). Two thirds (67.2%) of TDs are singletons and 3.6% are found at frequencies $\geq$0.5, significantly fewer than expected for neutral variants (FET $p < 0.001$; vs. SNPs in short introns; [197]) consistent with previous studies using smaller datasets and different methods (75% singletons [117] or 57% singletons [187]). While 11.6% (8.8%) of calls reciprocally overlap $\geq$ 50% ($\geq$80%) with calls made by Emerson et al. (2008), 46% (36.6%) overlap with calls made by Zichner et al. (2013), who used similar methods to identify TDs in a North American population of flies.

Tandem duplication can immediately generate genes with novel structures by partially duplicating existing genes [81, 123, 198, 199], and at least 18% of fixed *D. melanogaster*-specific genes are chimeric genes formed by partial tandem duplication and fusion of two genes (Chapter 2). However, theoretically any duplication encompassing a gene's promoter may allow transcription of new genome regions and the production of potentially novel coding or non-coding RNAs, and a large diversity of novel gene structures can potentially be formed by TD [123, 198].

I determined the context of DPGP2 TDs using the *D. melanogaster* release 6.05 reference gene annotations. Precise fusion points can be determined because of delly's accurate, nucleotide-resolution calls (Figure 4.2). Overall, 9.7% of TDs duplicate at least one entire gene, while 23% form at least one novel gene structure by fusing two existing genes or partially duplicating a gene, including its promoter. However, significantly more complete gene duplications (35) are found at frequencies $\geq$0.50 than chimeras or partial duplicates (28), suggesting that duplicates are less deleterious than partial duplicates or chimeras, as found

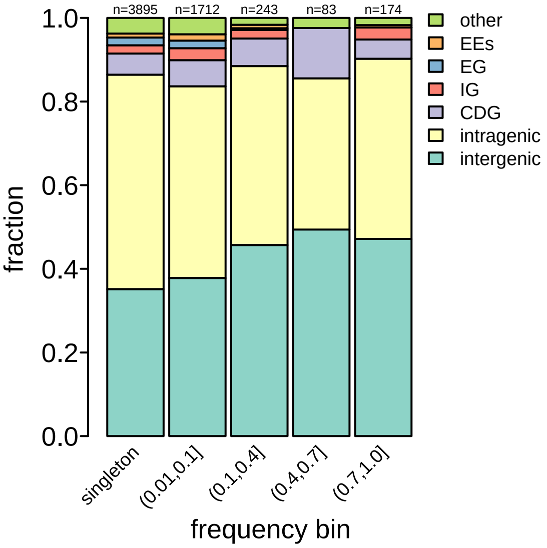


Figure 4.3: Tandem duplication type by frequency. The fraction tandem duplications in different contexts differs depending on the frequency bin. Only tandem duplications with a single predicted context were used. The five most abundant types are shown. EEs: exon-exon fusion of genes on the same strand; EG: exon-intergenic DNA fusion; IG: intron - intergenic DNA fusion; CDG: completely duplicated gene(s).

previously (FET $p < 0.0001$; Figure 4.3) [117, 187]. Interestingly, 51% of TDs are intragenic and cause expansion of introns (56%) or exons (12%). Furthermore, intragenic and intergenic TDs appear to be least deleterious, as they make up the vast majority of TDs that survive to intermediate and high frequencies, but a significantly lower proportion are found at these higher frequencies than expected for neutral variants (FET $p < 0.001$ vs. short intron SNPs; Figure 4.3).

There are thus myriad new gene loci segregating within the DPGP2 genomes. Many have novel gene structures and are frequently predicted to result in transcription fusing two genes or into intergenic DNA, creating many opportunities for novel coding or non-coding RNAs to be formed.

ID	freq.	CO	GA	KE	KK	KM	LA	MD	ML	MW	NK	NM	NR	OK	ZH	ZS
DUP162	0.837												▨			
DUP1900	0.128	▨	▨				▨				▨					▨
DUP2134	0.676														▨	
DUP2887	0.512						▨	▨	▨		▨	▨	▨			▨
DUP5193	0.028	▨	▨	▨	▨				▨	▨	▨	▨	▨	▨	▨	▨
DUP5200	0.056		▨	▨						▨			▨			
DUP5291	0.065		▨		▨			▨								
DUP5415	0.028							▨					▨	▨		
DUP5435	0.579										▨					▨
DUP5905	0.287	▨	▨	▨	▨	▨	▨			▨	▨	▨	▨	▨		
DUP6345	0.156													▨	▨	▨
DUP6954	0.560	▨			▨			▨			▨	▨		▨		
DUP6957	0.009						▨									

Figure 4.4: Tandem duplications were detected using PCR and Sanger sequencing in lines characterized by Emerson et al. 2008. Gray boxes indicate TD presence in that line. Frequency (freq.) is the frequency among the 109 DPGP2 lines. This is not a direct test of delly false positive rate because these are not DPGP2 lines.

TDs Are Present and Expressed in Lab Fly Lines

A major disadvantage of using the DPGP2 resource is that the fly lines from which embryos were collected were not inbred before collecting the embryos for sequencing [83, 196]. However, high-frequency duplications are often found in a variety of lab lines. I confirmed the presence of a set of 13 TDs in 15 inbred African-derived fly stocks studied by Emerson et al. (2008) with PCR and Sanger sequencing across the unique breakpoint (C-B junction in Figure 4.1; Figure 4.4). Current DPGP2 lines do not carry duplications that the haploid embryo sequences contain 54% of the time (7/13 confirmed, tested TDs; not shown).

It is difficult to specifically assay expression of completely duplicated genes because there is no way to amplify one copy versus the other, but chimeric genes have a unique breakpoint that can be spanned with PCR primers and amplified (Figure 4.1). I assumed in the previous section that any duplication that encompasses a transcription start site in the appropriate orientation can result in transcription through the unique breakpoint between the tandem duplicate copies (C-B in Figure 4.1) and potentially novel chimeric genes. I tested whether chimeric genes are expressed using RT-PCR and SOLiD whole-transcriptome RNA sequencing (RNAseq) from male and female whole flies. RT-PCR showed that all chimeric genes I confirmed in lab lines are expressed in all lines they are present in. More generally, RNAseq data from two lab lines contained multiple uniquely-mapped junction-spanning reads from 69% (51/74) and 75% (48/64) of polymorphic chimeric genes in lines Co and Zh1, respectively. These results show that partial gene duplications continue to be expressed through the breakpoint and potentially form novel fusion genes.

Reduced π and Extended Haplotypes Surrounding TDs

Classic genetic hitchhiking theory predicts that alleles linked to a selected allele will change frequency with the selected allele [102, 103, 200]. TDs that have been selected to intermediate and high frequencies should exhibit several characteristics, including: 1) a reduction in diversity flanking the duplicated regions specifically among chromosomes carrying the duplication; 2) an excess of rare variants among duplication chromosomes and an excess of intermediate and high frequency variants among ancestral chromosomes due to a recent effective population size reduction; 3) an excess of high-frequency variants in the whole set of chromosomes due to hitchhiking; and 4) long haplotypes surrounding duplications [102, 103, 200–206]. To test whether TDs are being positively selected I analyzed the reduction in nucleotide diversity (π, Figure 4.5 for example) and long haplotypes in TD flanking regions and compared these values to those surrounding putatively neutral SNPs. I will discuss alternatives and pitfalls below.

I first calculated nS_L, a powerful yet simple statistic to quantify haplotype lengths surrounding TDs and intergenic SNPs found in greater than 5 DPGP2 genomes, excluding SNPs in duplicated regions [206].

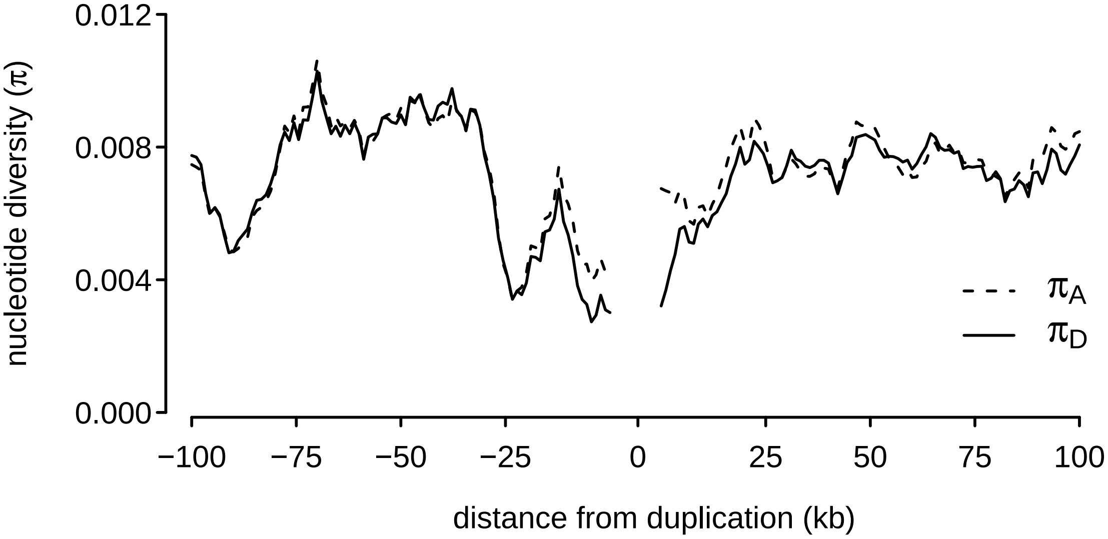


Figure 4.5: Nucleotide diversity (π) per site among chromsomes with (π_D) or without (π_A) DUP6345 in sliding 5 kb windows and 500 bp step.

$$\text{unstandardized } nS_L = log\left(\frac{SL_{A(k)}}{SL_{D(k)}}\right)$$

where $SL_{A(k)}$ is the quotient of the average physical length over which chromosomes carrying an ancestral allele at site k are identical by descent and the total number of segregating sites within those physical coordinates. S_L values were then standardized using S_L values for intergenic SNPs on the same chromosome arm in frequency bins of 0.05. Like other extended haplotype homozygosity statistics, the log ratio controls for differences in recombination rate between the two chromosome classes [205, 206]. Significantly long haplotypes surrounding derived alleles are indicated by extreme negative nS_L values. See Ferrer-Admetlla et al. (2014) for additional details [206].

Furthermore, I calculated π among chromosomes carrying (π_D, derived or duplicated) or not carrying (π_A, ancestral or absent) the duplication and compared the relative reduction in π_D to π_D around putatively neutral alleles at matched frequencies. I excluded the duplicated regions themselves to avoid the confounding effects of gene conversion, which may be a powerful homogenizing force in the early stages of duplication evolution [207, 208]. I quantified the reduction in π_D as the standardized difference between π_A and π_D ($\Delta\pi$) in regions of defined size w flanking the duplications.

$$\Delta\pi = \frac{\pi_A - \pi_D}{\pi_A} = \sum_{i=s-w}^{s} \frac{p_{A_i}(1-p_{A_i}) - p_{D_i}(1-p_{D_i})}{p_{A_i}(1-p_{A_i})} + \sum_{j=e}^{e+w} \frac{p_{A_j}(1-p_{A_j}) - p_{D_j}(1-p_{D_j})}{p_{A_j}(1-p_{A_j})}$$

where s and e are the start and end coordinates of the duplication and p_{A_i} is the frequency of the major allele at site i among ancestral (absence) chromosomes. Only biallelic sites were included. Analogous to nS_L analysis, I standardized $\Delta\pi$ values to those calculated for intergenic SNPs on the same chromosome at matched frequencies in bins of 0.05.

Table 4.1: nS_L and $\Delta\pi$ top hits.

Duplication	**Coordinates**	**Freq.**[a]	**Type**[b]	$\Delta\pi$ p_{5kb}	nS_L ***p***
DUP2887	3L:1310981-1315894	0.52	CDG;EEs	0.9946	0.0002
DUP3447	3L:13674879-13678963	0.19	intragenic_II	0.96	0.0014
DUP2137	2R:12207300-12207855	0.64	intergenic	0.9975	0.0016
DUP1930	2R:8960432-8964174	0.21	intergenic	0.9998	0.0023
DUP3755	3L:19643894-19644186	0.07	intergenic	0.9813	0.0047
DUP1900	2R:8563666-8566020	0.13	CDG	0.9954	0.0053
DUP461	2L:8997259-8997676	0.07	EIo;IEo	0.9166	0.0062
DUP5798	3R:26278219-26280524	0.12	intergenic	0.9866	0.0066
DUP781	2L:15935606-15937294	0.06	intragenic_II	0.9937	0.0072
DUP2426	2R:18050085-18053644	0.06	3Es	0.9918	0.0079
DUP1097	2L:21094034-21096539	0.08	intergenic	0.9937	0.0085
DUP6852	X:18230519-18235255	0.11	intergenic	0.9974	0.0101
DUP2261	2R:14853062-14857288	0.07	CDG;EEs	0.9885	0.0102
DUP5328	3R:17272377-17274695	0.07	intergenic	0.9841	0.0135
DUP4287	3R:444410-446824	0.07	intergenic	0.9547	0.0165
DUP5541	3R:21584673-21586910	0.11	intragenic_II	0.9907	0.0181
DUP3287	3L:10372726-10374665	0.09	intergenic	0.9716	0.0271
DUP144	2L:2653057-2654733	0.08	E5s	0.9355	0.0383
DUP4010	3L:23227775-23228068	0.15	intergenic	0.9985	0.0398
DUP2007	2R:10142104-10144624	0.06	EG	0.9492	0.0459
DUP3271	3L:9966782-9970344	0.06	E3o;E3s;3Eo	0.9614	0.0461

a: frequency among 109 DPGP2 genomes.
b: E - exon, I - intron, 3 - 3' UTR, G - intergenic DNA, CDG - completely duplicated gene, s (o) - fusion occurs between genes on the same (opposite) strand. TDs that form multiple
new genes separate the different types with ';'

I considered TDs with nS_L ($\Delta\pi$) values in the most negative (positive) 5% of TD and SNP values genome-wide as having good evidence of being positively selected (Table 4.1). This analysis detects 21 TDs with extremely negative nS_L values, all of which also exhibit significantly reduced $\Delta\pi$ in at least one window size (Table 4.1). Conversely, 104 (0.5 kb window) to 160 (5 kb window) TDs have extremely high $\Delta\pi$ values. The discrepancy is due to the fact that 1) $\Delta\pi$ has an absolute maximum value of 1 (i.e. when $\pi_D = 0$) and 2) many of these duplications have large repeat regions nearby that are masked, reduce the number of informative sites for π calculations, and therefore amplify differences between π_A and π_D (discussed below). About half of these duplications are at frequencies $\leq$0.10. Surprisingly, 8/21 are completely intergenic duplications, and could suggest that they affect the regulation of nearby genes (Discussion).

Case Studies

DUP2887 is a Prime Candidate for Further Study: DUP2887 has the most extreme $\Delta\pi$ and nS_L values of any TD (Figure 4.6). This duplication completely duplicates *CG9186* and also forms a novel chimeric gene comprised of the RCC1 domain of *Sherpa* and a domain of unknown function from *CG2469*. I hypothesized that the chimera had a novel function that was being selected, but I found that 70% of chimeric gene alleles segregating in lab lines contain frameshifts and premature stop codons (Figure 4.6). It is possible that these frameshifts were accumulated during inbreeding to generate these stocks. Cardoso-Moreira et al. (2016) also recently found DUP2887 to be being selected, and the *CG9186* duplication increased dosage of the gene, so it could be that the duplicate gene is the target of selection [199].

Gene - Intergenic DNA Fusions Can Immediately Generate Novel Transcripts: One ex-

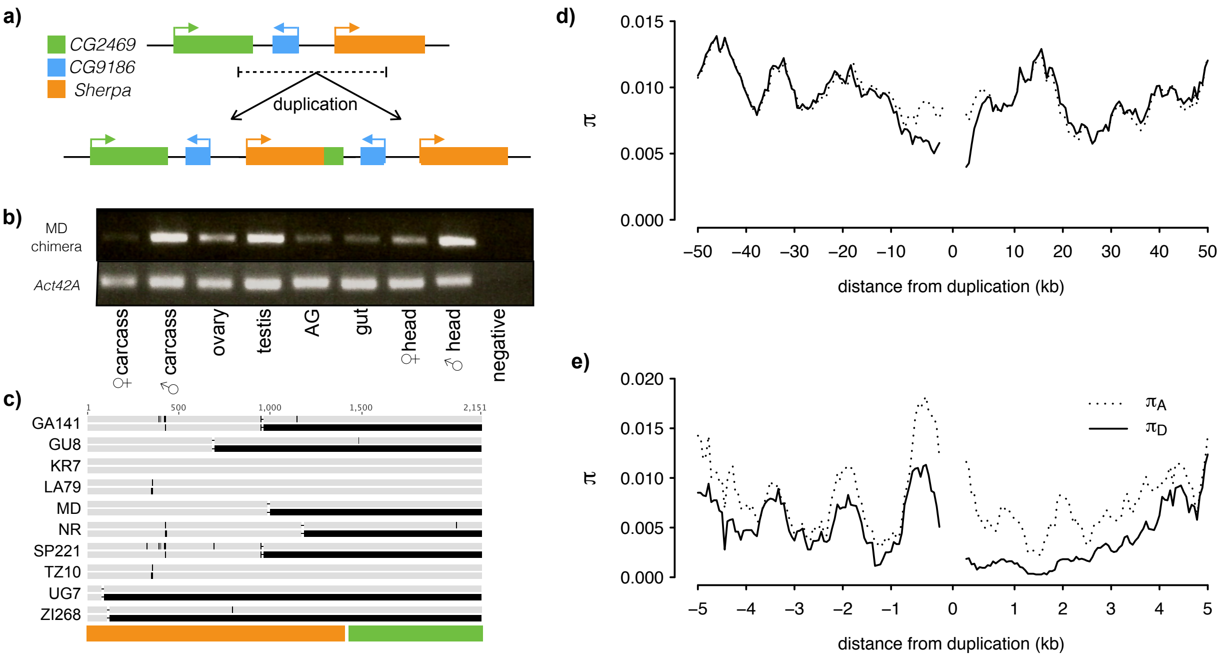


Figure 4.6: DUP2887 generates a novel chimeric gene and appears to be sweeping. The duplication is found in 51% of DPGP2 genomes and about half of lab strains I tested (Figure 4.4). **a)** A tandem duplication on chromosome 3L completely duplicates *CG9186* and partially duplicates and fuses *Sherpa* and *CG2469*. **b)** Semi-quantitative expression pattern of the chimera in tissues from one strain, MD. *Sherpa* is moderately to moderate-highly expressed in all modENCODE tissue and development datasets [95]. **c)** Cloning and sequencing the chimeric gene region in 10 lab strains showed that 70% of the chimeric genes contain indels that cause frameshifts and premature stop codons. For each sample (pair of horizontal bars), the top bar is the nucleotide sequence and the bottom bar is the translation. Vertical black bars indicate positions that differ from the predicted chimeric CDS (protein) sequence based on the reference genome. Regions that are out-of-frame with respect to the reference are solid black. **d)** π per site in 50 kb flanking DUP2887, calculated in sliding 5 kb windows (0.5 kb step). **e)** π in the 5 kb flanking regions, calculated in sliding windows of 0.5 kb and 50 bp step. π_A: π among chromosomes without the duplication; π_D: π among chromosomes carrying the duplication.

ample will highlight the large potential that TDs have to generate novel genes beyond strict gene duplication. A duplication of 3R:30,888,920..30,893,496 found in 15% of lab lines [117] copies only the promoter and first codon of *Sap-r* and produces novel transcription into intergenic DNA (Figure 4.7). There are three implications. First, without knowing that a line contains this duplication, it would appear that there is random, high transcription of intergenic DNA and potentially cause it to be annotated as a non-coding RNA. Second, if this duplication did eventually fix the locus would appear to have arisen *de novo* from previously non-coding DNA. I suspect this is one major way that *de novo* gene origination can be jump-started, though transcription of intergenic regions appears to be abundant in humans, yeast, and flies [95, 209–211]. Whether this is simply caused by a permissive chromatin environment or, in some cases, unknown duplications, is worth investigating. Third, this gene has multiple isoforms, indicating that splicing can occur at the earliest stages of *de novo* gene origination. Further work is needed to understand the contribution of this type of duplication to genome evolution, which accounts for 6% of duplications.

4.0.4 Discussion

Altogether, I find thousands of tandem duplications segregating in the DPGP2 dataset. While 10% of TDs completely duplicate genes, many more (23%) duplications partially duplicate pieces of one or more genes to generate novel fusions between all possible combinations of genetic elements. While the vast majority of TDs are known to be rapidly lost due to drift or purifying selection [117], at least 21 TDs in this dataset have extremely reduced diversity and long haplotypes, suggesting that they are rising in frequency due to positive selection.

Is Selection Acting on Tandem Duplications Segregating in the DPGP2?

A major unresolved question about new gene evolution is whether or not selection acts on new gene loci before fixation. Detecting positive selection on new loci would suggest that new loci are not redundant even immediately after they are formed. This would rule out classical models of new gene evolution in many cases, which assume two copies are redundant for very long periods of time [18, 73], and instead favor more recent models that posit selection before or shortly after new gene formation [21, 23, 213]. It certainly appears that most duplications are not redundant and in fact deleterious, because they are kept at low frequencies or purged

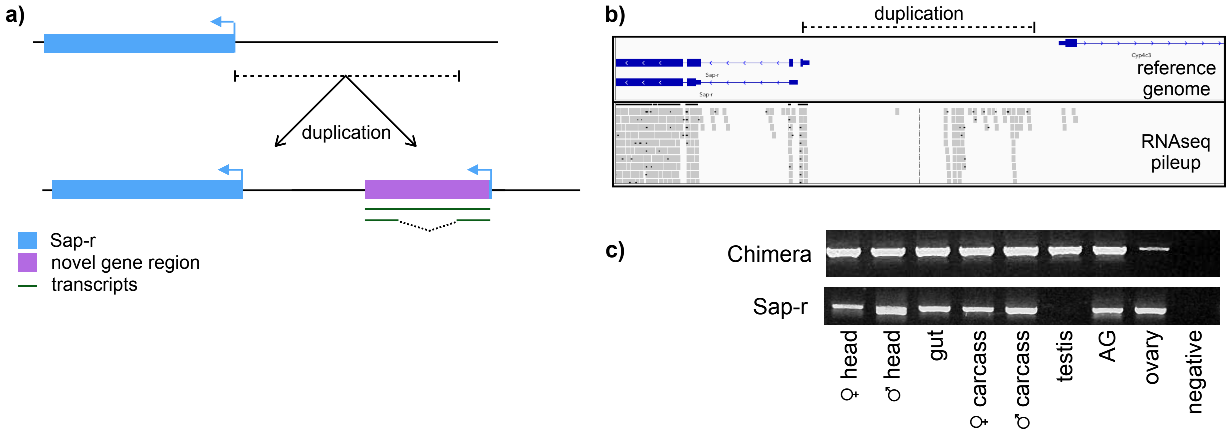


Figure 4.7: a) A duplication on chromosome 3R copies the promoter and first codon of *Sap-r*, resulting in transcription into previously intergenic DNA. Interestingly, this gene already has multiple isoforms, determined using 3' RACE: a 2 kb transcript and an 800 bp transcript produced by splicing out a 1.2 kb region. **b)** Whole transcriptome sequencing (RNAseq) shows transcription from this region in lines carrying the duplication (Zh1 is shown). A snapshot from the Integrated Genomics Viewer is shown [212]. *Cyp4c3* is on the opposite strand. **c)** Qualitative expression pattern using reverse-transcriptase PCR of *Sap-r* and the novel chimera shows that the novel gene gained expression in testis.

from the population. In addition, recent work has even shown that tandem duplication itself may often immediately change expression levels in different ways than expected [214].

Cardoso-Moreira et al. (2016) also recently published a manuscript showing 1) that duplications frequently affect gene dosage and 2) some duplications appear to be being positively selected [199]. She and her colleagues used a similar approach to the one take here (increased linkage and decreased diversity levels) in a different *D. melanogaster* population genomic dataset and found DUP2887 to be under selection and increased dosage from *CG9186* [199]. Certainly an interesting future direction would be to analyze genes nearby intergenic duplications to determine if the duplication is affecting transcript levels. This could be one explanation for the signatures of selection surrounding the eight intergenic TDs (Table 4.1).

An accurate assessment of the proportion and types of new loci that evolve non-neutrally, as Cardoso-Moreira et al. (2016) and I have attempted, will also help to begin to distinguish between classes of new genes that are being selected. It is well-documented that genes involved in acute selective pressures (pesticides, heavy metals, pathogen resistance, etc.) are frequently and recurrently duplicated and can rise quickly in frequency, probably due to their effect of increasing gene dosage [199, 215, 216]. However, fixed new genes are not enriched for with these types (Chapter 2) [63], suggesting that duplicated resistance genes are most often eventually lost or perhaps maintained as balanced polymorphisms. Therefore the distinction will need to be made between new genes with continually beneficial functions and new genes with conditionally beneficial functions. New experimental techniques such as CRISPR/Cas9 should make the insertion/deletion of duplications into/out of lines relatively straightforward and provide the ability to directly test these types of hypotheses.

A final, serious issue with this analysis is that deleterious variants may also rise (or fall) in frequency just as quickly as beneficial variants and cause identical changes in the patterns of linkage and diversity in their flanking regions [217, 218]. Distinguishing these cases from truly positively-selected and beneficial cases is a problem that can only be solved using detailed molecular studies of duplication fitness effects.

$\Delta\pi$ and nS_L Caveats

I compared TD $\Delta\pi$ and nS_L values to those surrounding intergenic SNPs at matched frequencies because this is a large class of sites that allows reasonable estimates of the distribution of expected values surrounding relatively neutral variants at different frequencies. Furthermore, there are at most eight intergenic TDs in bins >0.15, precluding their use as a null set. In contrast, there are a minimum of 7,769 SNPs in every bin. Are SNPs and TDs comparable? I would argue that using SNPs is an extremely conservative approach to determining which TDs have extreme statistics. The results were identical using synonymous SNPs (not shown).

A second difficulty we face analyzing π and haplotype lengths around TDs is that linkage disequilibrium decays by 50% over ∼50 bp in *D. melanogaster* [108, 219]. This also means that $\Delta\pi$ values calculated in the regions flanking SNPs should be more severe than those flanking TDs. Haplotypes also have a greater chance to decay before they reach the edge of the TDs, where I began the nS_L calculation. nS_L is a measure of the total physical distance over which a haplotype extends around a focal variant and is calculated as $\log((L_A/S_A)/(L_D/S_D))$, where S_A and S_D are the average number of segregating sites in the longest haplotypes on chromosomes carrying the ancestral or derived alleles, and L_A and L_D are the average physical haplotype lengths, respectively. Longer physical distances containing fewer SNPs produce larger positive S_L values, and further reduce the power to detect long haplotypes surrounding TDs. For example, nS_L may calculate a particular 500 bp-long TD to have $L_A = L_D = 1000$ bp and $S_A = S_D = 10$ SNPs, and $S_L = 0$. However, the true L_D is 1500 bp and thus S_L is actually -0.405. So, excluding the duplicated region biases TD S_L values upward and reduces power to detect long haplotypes containing recent duplicates (Figure 4.8). Furthermore, there is evidence in the literature that duplication-containing chromosomes have similar, if not slightly elevated, recombination rates and should not have inflated haplotype lengths [220–222].

One possible future direction is to use coalescent simulations to determine significance, but an appropriate demographic model needs to be used and it is not yet clear precisely what that is for African *D. melanogaster*, although there are good estimates [223, 224].

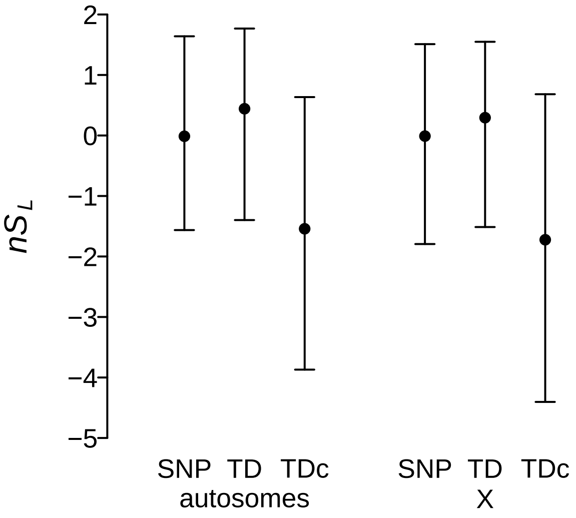


Figure 4.8: nS_L statistics for SNPs and tandem duplications segregating in the DPGP2. Means and 95% confidence intervals are shown. TDc values are nS_L values corrected for the fact that the physical distance on duplication chromosomes is increased by the length of the duplication. S_L values were normalized to the distribution of intergenic SNP S_L values.

4.0.5 Methods

Tandem Duplication Calling Algorithms

I tested the specificity and sensitivity of Pindel v0.2.5b1 [193] and delly v0.6.5 [195] using simulations modeled after the *Drosophila* Population Genomics Project Phase 2 (DPGP2) haploid embryo sequences [83, 196]. I avoided TD calling methods based on read depth because the haploid embryo sequencing protocol results in high variance in read depth across the genome [196]. I simulated TDs in the *D. melanogaster* release 6 chromosome 2L using RSVSim [225], randomly drawing TD lengths from an empirical distribution of 2L TDs [117], then simulated next-generation sequencing reads from this pseudo-reference genome with ART (version Chocolate Cherries 03-09-2015) [226] using default settings and the Illumina GAIIx error profile included in the package. I mapped reads to the dm6 reference 2L with bwa mem v0.7.5a [128] with default parameters. To approximate the coverage distribution of the DPGP2 genomes, I simulated 3 sets of reads with mean coverages of 10, 25, and 50, then downsampled to a target mean coverage. TDs were then called with pindel and delly. I simulated 100 rearranged 2L sequences with an average of 30 tandem duplications each. I tested the sensitivity and specificity of delly and pindel to varying mean read coverage (5, 10, 15, 20, 25, or 30) and minimum read mapping quality (10, 20, or 30). I considered a caller to have correctly called a TD if both the start and end coordinates were within 50 bp of the true TD coordinates.

TDs in the DPGP2 Genomes

I called TDs in the 109 core DPGP2 genomes using delly v0.6.5 with default settings, then filtered calls using custom perl scripts. A TD was considered to be present in an individual genome if the genome had at least one supporting read pair and the TD call had at least three total supporting read pairs from all sample genomes. TD context was determined using FlyBase release 2015_05 gene models [96], BEDtools [227], and custom perl scripts.

TD Confirmation and Expression in Lab Lines

PCR primers were designed using primer-BLAST and PCRs performed using Standard Taq Polymerase (New England BioLabs, USA) on a BioRad C100 Thermal Cycler. All Sanger sequencing was performed at the University of Chicago Comprehensive Cancer Center Sequencing Core.

Male and female tissues were dissected in 1× PBS at room temperature and immediately placed in a 1.5 ml tube with RNA*later* (Ambion, USA). Samples were frozen until extraction. Prior to extraction, RNA*later* was removed by filling the tube with fresh PBS, centrifuging at maximum speed at $4°C$ for 5 minutes, pipetting away the supernatant, and repeating the wash once. I extracted RNA using TRIzol (Ambion, USA) then treated the RNA with RNase-free DNase (QIAgen, USA), all according the manufacturer's instructions. I synthesized cDNA from 2μg treated RNA using Superscript III Reverse Transcriptase (Invitrogen, USA) and oligo-dT$_2$0 priming. cDNA was diluted 1:10 in water before using 1μL in PCRs.

Whole transcriptome sequencing using SOLiD 5500xl (Applied Biosystems, USA) technology was performed on two biological replicates of each sex for lines Co and Zh1. RNA was extracted from 15 five day old males or females per replicate using TRIzol (Ambion, USA). To detect expression of putative chimeric genes, we added pseudo-TD junctions to the release 6 reference genome file and then mapped raw colorspace reads to this pseudo reference using BFAST 0.7.a [228, 229]. I considered a chimeric gene to be expressed if I found $\geq$3 uniquely and fully-mapped reads spanning the pseudo TD junction.

nS_L Calculations

I used nS_L version 0.47 to calculate the nS_L statistic for each duplication [206]. Regions that were duplicated in any genome were masked in all genomes prior to analysis. SNPs encompassed by the duplication were replaced with a single entry at the duplication midpoint for each TD in the VCF file generated in Chapter 2 (page 33). nS_L input files were generated using custom perl scripts. Ancestral SNP states were assigned following Chapter 2 (page 34).

π Calculations

Only TDs with frequencies between 0.05 and 0.95 were analyzed. π was calculated following ref. [107] using SNP calls generated in Chapter 2 (page 33) and custom perl scripts.

CHAPTER 5

FUTURE DIRECTIONS

5.0.1 Recap

I introduced this thesis in Chapter 1 as an exploration of two questions regarding genes that were formed specifically in *Drosophila melanogaster*:

1. What are the fitness effects of *D. melanogaster*-specific genes?

2. What are the evolutionary forces that governed or are governing *D. melanogaster*-specific gene evolution?

I used genetic manipulation and empirical population genetic analyses to provide at least partial answers to these questions. First, I showed in Chapters 2 and 3 that at least 27% of *species-specific* genes, less than 2 million years old, are essential for fly development, and at least one is essential for fly fertility. These results make it clear that new gene copies can be critical components of the pathways and networks controlling important processes such as development and reproduction. Furthermore, I provide evidence in Chapter 3 that sexual selection and resolution of sexual antagonism may drive new gene evolution. Finally, I provide some evidence suggesting that a handful of the thousands of duplications segregating in *D. melanogaster* populations appear to be rising in frequency due to positive selection, suggesting that whatever new gene function that selection acted on in fixed new genes probably operated from the moment they were formed or very shortly thereafter.

The main point of Chapter 2, that a gene's sequence conservation or presence in a wide range of species is not predictive of its functional importance, has now been well-established (Chapter 2 and ref. [63]). Thus, strong arguments can be made for careful molecular studies of genes of all ages, not just ancient genes, in a variety of organisms to understand the evolution of phenotype, disease, and processes like development and reproduction. The reason for this is that, as I have emphasized several times throughout this work, no gene acts in isolation. Recent advances in functional genomics techniques have only just begun to really highlight the complexity of how genes interact with each other to produce functional and fit organisms.

The big question that needs to be addressed then changes from "In what ways and why are processes such as development and reproduction *conserved* between organisms?" to "In what ways and why do these existing processes *differ* between organisms?". The answers to these questions bear on a broad range of fields, from theoretical population genetics descriptions of the limits of selection to the robustness of gene regulation to the causes of the evolution of development.

5.0.2 Future Work

The Fitness Effect of Gaining a New Gene

All experimental studies of new gene evolution use loss-of-function methods to ask what a new gene's function is. This approach has clearly been successful, but only allows *inference* of the new gene's fitness effect. What is the initial benefit, if any, of gaining a new gene? I see two approaches to answering this question.

First, what is the effect of simulating a *D. melanogaster*-specific duplication in *D. simulans*? It is likely that *D. melanogaster*-specific duplications had some fitness benefit that caused them to become fixed and maintained in that species. They are thus good candidates for further study of the initial fitness effects and evolution of new genes. Insertion of *D. melanogaster*-specific genes or duplications into *D. simulans* may provide clues as to why these duplications became fixed in *D. melanogaster* and what those initial fitness effects were. This experiment can be executed with the CRISPR/Cas9 system to knock in *D. melanogaster*-specific genes or simulate a duplication of the orthologous region in *D. simulans* (i.e. insert a second copy of the *D. simulans* gene into the *D. simulans* strain). The fitness effect could easily be measured using protocols like those in Chapter 3. One could also imagine marking the insertion and tracking its frequency in experimental fly populations over time, simulating what might happen to the new duplication in natural populations.

Second, what is the effect of gaining a new gene duplicate in a population? Thousands of new genes segregate within *D. melanogaster* populations. What are their fitness effects, if any? I think this question can best be answered by inserting new duplicates into *D. melanogaster* strains that do not carry the duplicates and assaying their fitness. Insertions can be made using either the ϕC31 system or the CRISPR/Cas9 system and take advantage of the large number of different *D. melanogaster* strains available to the community (e.g. the DGRP). To maximize the probability of observing a fitness effect, one should focus on those duplicates found to be potentially under positive selection in Chapter 3, for example. Fitness effects can be assayed relative to no-insertion lines, again following protocols like those in Chapter 3.

By far the biggest difficulty with these experiments is providing the new gene's expression pattern. In the simplest case, and to maximize the probability of observing a fitness effect, one could place the new gene under control of a constitutive promoter. However, this approach may bias the results, probably causing new genes to appear more detrimental than they really are because such a broad expression pattern will cause many pleiotropic effects. Alternatively, one could use the promoter region of the single-copy gene whose duplication is being tested.

Altogether, though, the power of CRISPR/Cas9 and the fact that there are some very good candidate new genes should encourage experiments like these, which should help elucidate the fitness effects, if any, that cause new genes to become fixed and maintained in the genome.

New Gene Male Bias: Causes and Effects

I see two outstanding problems in the area of new gene research in general. First, it is still unclear what evolutionary forces actually drive the acquisition of new genes by populations. My work suggests that new genes may be being selected specifically for their role in male reproduction or resolution of sexual antagonism. In the future, some careful work will need to be done investigating the fitness effects of the pre-duplication ancestral gene, either by ancestral state reconstruction and testing, or by studying the single-copy ortholog in outgroup species, as described above.

New genes tend to be highly expressed in testis in flies and mammals. Is this a cause of new gene origination or is it an effect? That is, are testis-biased genes the most frequently beneficial type of gene, or is it just that the testis is a permissive tissue that ‘allows’ new genes to survive because there is low pleiotropy for testis-biased genes?

Understanding New Gene Evolution Using Systems Biology

I emphasized and argued several times throughout this thesis that no gene acts in isolation and that it is the interactions between genes that dictate a particular gene’s phenotypic effects. High-throughput technologies have allowed the rapid evolution of the comparative and functional genomics fields in the last decade or so. These technologies are fantastic for describing broad-scale patterns of genome evolution and gene-gene or gene-DNA interactions. However, we still have few solid experimental links between those patterns and actual (evolutionarily important) phenotypes.

We are now on the verge of true functional genomics, though, with technologies like the CRISPR/Cas9 system and all its variants. Some interesting questions to ask about new gene evolution and new gene phenotypic effects using functional genomics techniques are:

1. How does the addition of a new gene affect existing gene-gene interaction networks? This type of question will be important to understand the contribution of models like developmental systems drift [230] to evolution.
2. How do new gene expression patterns rapidly diverge? That is, what is the role of divergence in *cis*-regulatory elements or *trans*-acting factors that cause duplicate expression patterns to diverge?
3. What is the relative contribution (and order) of expression pattern and protein sequence divergence to duplicate gene pair evolution?

Conclusion

Finally, there are many exciting avenues to take to understand the contributions of new genes and other young genetic variants to evolution. The future will depend on the continuing rapid development of precise genome editing techniques and experimental characterization of genome-wide patterns of sequence, expression pattern, and gene-gene interaction evolution, but looks extremely promising. This also means that combinations of computational, experimental, and ecological approaches to understanding evolution will be important to achieve a fuller understanding of the past and continuing evolution of life.

BIBLIOGRAPHY

1. Darwin, C. *On the Origin of Species By Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life* 1st ed., 502 (John Murray, 1859).
2. Desmond, A. & Moore, J. *Darwin* (Penguin Group, 2009).
3. Darwin, C. *The Variation of Animals and Plants Under Domestication* 1st ed., 441 (John Murray, 1868).
4. Darwin, C. *The Descent of Man and Selection in Relation to Sex* 1st ed. (John Murray, 1871).
5. Provine, W. B. *The Origins of Theoretical Population Genetics* 211 (The University of Chicago Press, 2001).
6. Morgan, T. H., Sturtevant, A. H., Muller, H. J. & Bridges, C. B. *The Mechanism of Mendelian Heredity* 262 (Henry Holt and Co., 1915).
7. Kohler, R. E. *Lords of the Fly* 344 (University of Chicago Press, 1994).
8. Sturtevant, A. H. The effects of unequal crossing over at the bar locus in Drosophila. *Genetics* **10,** 117–147 (1925).
9. Muller, H. J., Prokofyeva-Belgovskaya, A. A. & Kossikov, K. V. Unequal Crossing Over in the Bar Mutant as a Result of Duplication of a Minute Chromosome Section. *Comptes Rendus de L'Academie des Sciences de L'URRS* **2,** 78 (1936).
10. Bridges, C. B. Bar as a Duplication. *Science* **83,** 210–211 (1936).
11. Muller, H. J. Bar Duplication. *Science* **83,** 528–530 (1936).
12. Fisher, R. A. *The Genetical Theory of Natural Selection* (Clarendon Press, 1930).
13. Wright, S. The Genetical Theory of Natural Selection. A Review. *Journal of Heredity* **21,** 349–356 (1930).
14. Haldane, J. The Time of Action of Genes, and Its Bearing on some Evolutionary Problems. *The American naturalist* **66,** 5–24 (1932).
15. Lewin, B., Krebs, J. E., Kilpatrick, S. T. & Goldstein, E. S. *Lewin's Genes X* 10th ed. (Jones and Bartlett Publishers, LLC, 2011).
16. Long, M., VanKuren, N. W., Chen, S. & Vibranovski, M. D. New gene evolution: little did we know. *Annual Review of Genetics* **47,** 307–333 (2013).
17. Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* **287,** 2204–2215 (2000).
18. Ohno, S. *Evolution by gene duplication* 160 (Springer-Verlag, 1970).
19. Ohno, S. Gene duplication, mutation load, and mammalian genetic regulatory systems. *Journal of Medical Genetics* **9,** 254–263 (1972).
20. Francino, M. P. An adaptive radiation model for the origin of new gene functions. *Nature Genetics* **37,** 573–577 (2005).

21. Bergthorsson, U., Andersson, D. I. & Roth, J. R. Ohno's dilemma: Evolution of new genes under continuous selection. *Proceedings of the National Academy of Sciences of the United States of America* **104,** 17004–17009 (2007).

22. Piatigorsky, J. & Wistow, G. The Recruitment of Crystallins : New Functions Precede Gene Duplication. *Science* **252,** 1078–1079 (1991).

23. Hughes, A. L. The Evolution of Functionally Novel Proteins after Gene Duplication. *Proceedings of the Royal Society of London B* **256,** 119–124 (1994).

24. Hittinger, C. T. & Carroll, S. B. Gene duplication and the adaptive evolution of a classic genetic switch. *Nature* **449,** 677–681 (2007).

25. Des Marais, D. L. & Rausher, M. D. Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature* **454,** 762–765 (2008).

26. Deng, C., Cheng, C.-H. C., Ye, H., He, X. & Chen, L. Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 21593–21598 (2010).

27. Chen, S., Krinsky, B. H. & Long, M. New genes as drivers of phenotypic evolution. *Nature Reviews Genetics* **14,** 645–60 (2013).

28. Betrán, E., Thornton, K. & Long, M. Retroposed new genes out of the X in Drosophila. *Genome research* **12,** 1854–1859 (2002).

29. Wang, W. *et al.* High Rate of Chimeric Gene Origination by Retroposition in Plant Genomes. *The Plant Cell* **18,** 1791–1802 (2006).

30. Gilbert, W. Why genes in pieces? *Nature* **271,** 501 (1978).

31. Patthy, L. Exons - Original building blocks of proteins? *BioEssays* **13,** 187–192 (1991).

32. Levine, M. T., Jones, C. D., Kern, A. D., Lindfors, H. A. & Begun, D. J. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proceedings of the National Academy of Sciences of the United States of America* **103,** 9935–9939 (2006).

33. Long, M., Betrán, E., Thornton, K. & Wang, W. The origin of new genes: glimpses from the young and old. *Nature Reviews Genetics* **4,** 865–75 (2003).

34. Kaessmann, H., Vinckenbosch, N. & Long, M. RNA-based gene duplication: mechanistic and evolutionary insights. *Nature Reviews Genetics* **10,** 19–31 (2009).

35. Zhou, Q. *et al.* On the origin of new genes in Drosophila. *Genome Research* **18,** 1446–55 (2008).

36. Marques-Bonet, T., Girirajan, S. & Eichler, E. E. The origins and impact of primate segmental duplications. *Trends in Genetics* **25,** 443–454 (2009).

37. Roelofs, J. & Van Haastert, P. Genes lost during evolution. *Nature* **411,** 1013–1014 (2001).

38. Fortna, A. *et al.* Lineage-specific gene duplication and loss in human and great ape evolution. *PLoS Biology* **2,** 937–954 (2004).

39. Demuth, J. P., De Bie, T., Stajich, J. E., Cristianini, N. & Hahn, M. W. The Evolution of Mammalian Gene Families. *PloS One* **1,** e85 (2006).

40. Hahn, M. W., Han, M. V. & Han, S. G. Gene family evolution across 12 Drosophila genomes. *PLoS Genetics* **3,** 2135–2146 (2007).

41. Demuth, J. P. & Hahn, M. W. The life and death of gene families. *BioEssays* **31,** 29–39 (2009).

42. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290,** 1151–1155 (2000).

43. Zhang, Y. E., Vibranovski, M. D., Krinsky, B. H. & Long, M. Age-dependent chromosomal distribution of male-biased genes in Drosophila. *Genome Research* **20,** 1526–1533 (2010).

44. Zhang, Y. E., Landback, P., Vibranovski, M. D. & Long, M. Accelerated Recruitment of New Brain Development Genes into the Human Genome. *PLoS Biology* **9,** e1001179 (2011).

45. Zhang, Y. E., Landback, P., Vibranovski, M. & Long, M. New genes expressed in human brains: Implications for annotating evolving genomes. *BioEssays* **34,** 982–991 (2012).

46. Emerson, J. J., Kaessmann, H., Betrán, E. & Long, M. Extensive gene traffic on the mammalian X chromosome. *Science* **303,** 537–540 (2004).

47. Vibranovski, M. D., Zhang, Y. & Long, M. Out of the X chromosomal gene movement in the Drosophila genus. *Genome Research* **19,** 897–903 (2009).

48. Vibranovski, M. D., Zhang, Y. & Long, M. General gene movement off the X chromosome in the Drosophila genus. *Genome Research* **19,** 897–903 (2009).

49. Wang, J., Long, M. & Vibranovski, M. D. Retrogenes Moved Out of the Z Chromosome in the Silkworm. *Journal of Molecular Evolution* **74,** 113–126 (2012).

50. Parisi, M. *et al.* Paucity of Genes on the Drosophila X Chromosome Showing Male-Biased Expression. *Science* **299,** 697–700 (2003).

51. Ranz, J. M., Castillo-Davis, C. I., Meiklejohn, C. D. & Hartl, D. L. Sex-Dependent Gene Expression and Evolution of the Drosophila Transcriptome. *Science* **300,** 1742–1745 (2003).

52. Khil, P. P., Smirnova, N. A., Romanienko, P. J. & Camerini-Otero, R. D. The mouse X chromosome is enriched for sex-biased genes not subject to selection by meiotic sex chromosome inactivation. *Nature Genetics* **36,** 642–646 (2004).

53. Reinke, V., Gil, I. S., Ward, S. & Kazmer, K. Genome-wide germline-enriched and sex-biased expression profiles in *Caenorhabditis elegans*. *Development* **131,** 311–323 (2004).

54. Long, M. & Langley, C. H. Natural Selection and the Origin of *jingwei*, a Chimeric Processed Functional Gene in Drosophila. *Science* **260,** 91–95 (1993).

55. Conant, G. C. & Wolfe, K. H. Turning a hobby into a job: How duplicated genes find new functions. *Nature Reviews Genetics* **9,** 938–950 (2008).

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

57. Ding, Y., Zhou, Q. & Wang, W. Origins of New Genes and Evolution of Their Novel Functions. *Annual Review of Ecology, Evolution, and Systematics* **43,** 345–363 (2012).

58. Chen, L., DeVries, A. L. & Cheng, C. H. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. *Proceedings of the National Academy of Sciences of the United States of America* **94,** 3811–6 (1997).

59. Chen, L., DeVries, A. L. & Cheng, C. H. Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod. *Proceedings of the National Academy of Sciences of the United States of America* **94,** 3817–3822 (1997).

60. Cheng, C.-H. C. & Chen, L. Evolution of an antifreeze glycoprotein. *Nature* **401,** 443–444 (1999).

61. Dai, H. *et al.* The evolution of courtship behaviors through the origination of a new gene in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America* **105,** 7478–83 (2008).

62. Ding, Y. *et al.* A Young Drosophila Duplicate Gene Plays Essential Roles in Spermatogenesis by Regulating Several Y-linked Male Fertility Genes. *PLoS Genetics* **6,** e1001255 (2010).

63. Chen, S., Zhang, Y. E. & Long, M. New genes in Drosophila quickly become essential. *Science* **330,** 1682–1685 (2010).

64. Matsuno, M. *et al.* Evolution of a Novel Phenolic Pathway for Pollen Development. *Science* **325,** 1688–1692 (2009).

65. Chen, S. *et al.* Reshaping of global gene expression networks and sex-biased gene expression by integration of a young gene. *The EMBO journal* **31,** 2798–2809 (2012).

66. Chen, S. *et al.* Frequent Recent Origination of Brain Genes Shaped the Evolution of foraging behavior in Drosophila. *Cell Reports* **1,** 118–132 (2012).

67. Wu, X. & Sharp, P. A. Divergent transcription: a driving force for new gene origination? *Cell* **155,** 990–996 (2013).

68. Smock, R. G., Yadid, I., Dym, O., Clarke, J. & Tawfik, D. S. De Novo Evolutionary Emergence of a Symmetrical Protein Is Shaped by Folding Constraints. *Cell* **164,** 476–486 (2016).

69. Ragsdale, E. J., Mu, M. R., Ro, C. & Sommer, R. J. A Developmental Switch Coupled to the Evolution of Plasticity Acts through a Sulfatase. *Cell* **155,** 922–933 (2013).

70. Ross, B. D. *et al.* Stepwise Evolution of Essential Centromere Function in a Drosophila Neogene. *Science* **340,** 1211–1214 (2013).

71. Jacob, F. Evolution and Tinkering. *Science* **196,** 1161–1166 (1977).

72. Ashburner, M. *et al.* An Exploration of the Sequence of a 2.9-Mb Region of the Genome of *Drosophila melanogaster*: The *Adh* Region. *Genetics* **153,** 179–219 (1999).

73. Walsh, J. B. How Often Do Duplicated Genes Evolve New Functions? *Genetics* **139,** 421–428 (1995).

74. Lynch, M., O'Hely, M., Walsh, B. & Force, A. The probability of preservation of a newly arisen gene duplicate. *Genetics* **159,** 1789–1804 (2001).

75. Walsh, B. Population-genetic models of the fates of duplicate genes. *Genetica* **118,** 279–294 (2003).

76. Force, A. *et al.* Preservation of Duplicate Genes by Complementary, Degenerative Mutations. *Genetics* **151,** 1531–1545 (1999).

77. Assis, R. & Bachtrog, D. Neofunctionalization of young duplicate genes in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America* **110,** 17409–17414 (2013).

78. Obbard, D. J. *et al.* Estimating divergence dates and substitution rates in the drosophila phylogeny. *Molecular Biology and Evolution* **29,** 3459–3473 (2012).

79. Clark, A. G. *et al.* Evolution of genes and genomes on the Drosophila phylogeny. *Nature* **450,** 203–218 (2007).

80. Yates, A. *et al.* Ensembl 2016. *Nucleic Acids Research* **44,** D710–D716 (2016).

81. Rogers, R. L., Bedford, T., Lyons, A. M. & Hartl, D. L. Adaptive impact of the chimeric gene Quetzalcoatl in Drosophila melanogaster. *Proceedings of the National Academy of Sciences of the United States of America* **107,** 10943–8 (2010).

82. Zhan, Z. *et al.* Rapid functional divergence of a newly evolved polyubiquitin gene in Drosophila and its role in the trade-off between male fecundity and lifespan. *Molecular Biology and Evolution* **29,** 1407–1416 (2011).

83. Pool, J. *et al.* Population Genomics of sub-saharan Drosophila melanogaster: African diversity and non-African admixture. *PLoS Genetics* **8,** e1003080 (2012).

84. Dietzl, G. *et al.* A genome-wide transgenic RNAi library for conditional gene inactivation in Drosophila. *Nature* **448,** 151–156 (2007).

85. Ni, J. *et al.* Vector and parameters for targeted transgenic RNA interference in *Drosophila melanogaster. Nature Methods* **5,** 49–51 (2008).

86. Miklos, G. L. G. & Rubin, G. M. The Role of the Genome Project in Determining Gene Function : Insights from Model Organisms. *Cell* **86,** 521–529 (1996).

87. Perrimon, N., Lanjuin, A., Arnold, C. & Noll, E. Zygotic Lethal Mutations With Maternal Effect Phenotypes in Drosophila melanogaster. II. Loci on the Second and Third Chromosomes Identified by P-element-Induced Mutations. *Genetics* **144,** 1681–1692 (1996).

88. Gu, Z., Nicolae, D., Lu, H. H.-S. & Li, W.-H. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends in Genetics* **18,** 609–613 (2002).

89. Gu, Z., Rifkin, S. A., White, K. P. & Li, W.-H. Duplicate genes increase gene expression diversity within and between species. *Nature Genetics* **36,** 577–579 (2004).

90. Bai, Y., Casola, C., Feschotte, C. & Betrán, E. Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in Drosophila. *Genome Biology* **8,** R11.1–R11.9 (2007).

91. Rogers, R. L. & Hartl, D. L. Chimeric genes as a source of rapid evolution in Drosophila melanogaster. *Molecular Biology and Evolution* **29,** 517–529 (2012).

92. Xu, G., Guo, C., Shan, H. & Kong, H. Divergence of duplicate genes in exon-intron structure. *Proceedings of the National Academy of Sciences of the United States of America* **109,** 1187–1192 (2012).

93. Soria, P. S., Mcgary, K. L. & Rokas, A. Functional Divergence for Every Paralog. *Molecular Biology and Evolution* **31,** 984–992 (2014).

94. Celniker, S. E. *et al.* Unlocking the secrets of the genome. *Nature* **459,** 927–930 (2009).

95. Graveley, B. R. *et al.* The developmental transcriptome of *Drosophila melanogaster*. *Nature* **471,** 473–479 (2011).

96. Attrill, H. *et al.* Flybase: Establishing a gene group resource for *Drosophila melanogaster. Nucleic Acids Research* **44,** D786–D792 (2016).

97. Arthur, R. K. *et al.* Evolution of H3K27me3-marked chromatin is linked to gene expression evolution and to patterns of gene duplication and diversification. *Genome Research* **24,** 1115–1124 (2014).

98. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research* **37,** 1–13 (2009).

99. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4,** 44–57 (2008).

100. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24,** 1586–1591 (2007).

101. McDonald, J. H. & Kreitman, M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351,** 652–654 (1991).

102. Maynard Smith, J. & Haigh, J. The hitch-hiking effect of a favourable gene. *Genetical Research* **23,** 23–35 (1974).

103. Kaplan, N. L., Hudson, R. R. & Langley, C. H. The "hitchhiking effect" revisited. *Genetics* **123,** 887–899 (1989).

104. Hudson, R. R., Kreitman, M. & Aguadé, M. A Test of Neutral Molecular Evolution Based on Nucleotide Data. *Genetics* **116,** 153–159 (1987).

105. Tajima, F. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics* **123,** 585–595 (1989).

106. Fay, J. C. & Wu, C.-i. Hitchhiking Under Positive Darwinian Selection. *Genetics* **155,** 1405–1413 (2000).

107. Begun, D. J. *et al.* Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in Drosophila simulans. *PLoS Biology* **5,** 2534–2559 (2007).

108. Langley, C. H. *et al.* Genomic Variation in Natural Populations of *Drosophila melanogaster*. *Genetics* **192,** 533–598 (2012).

109. Svetec, N., Pavlidis, P. & Stephan, W. Recent Strong Positive Selection on *Drosophila melanogaster HDAC6*, a Gene Encoding a Stress Surveillance Factor, as Revealed by Population Genomic Analysis. *Molecular Biology and Evolution* **26,** 1549–1556 (2009).

110. Hu, T. T., Eisen, M. B., Thornton, K. R. & Andolfatto, P. A second-generation assembly of the Drosophila simulans genome provides new insights into patterns of lineage-specific divergence. *Genome Research* **23,** 89–98 (2013).

111. Haldane, J. The part played by recurrent mutation in evolution. *The American Naturalist* **67,** 435–469 (1933).

112. Pavlicev, M. & Wagner, G. P. A model of developmental evolution: Selection, pleiotropy and compensation. *Trends in Ecology and Evolution* **27,** 316–322 (2012).

113. Klomp, J. *et al.* A cysteine-clamp gene drives embryo polarity in the midge Chironomus. *Science* **348,** 1040–1042 (2015).

114. Jiang, P., Ludwig, M. Z., Kreitman, M. & Reinitz, J. Natural variation of the expression pattern of the segmentation gene even-skipped in melanogaster. *Developmental Biology* **405,** 173–181 (2016).

115. Kimura, M. Model of effectively neutral mutations in which selective constraint is incorporated. *Proceedings of the National Academy of Sciences of the United States of America* **76,** 3440–3444 (1979).

116. Kreitman, M. Nucleotide Polymorphism at the *alcohol dehydrogenase* locus of *Drosophila melanogaster*. *Nature* **304,** 412–417 (1983).

117. Emerson, J. J., Cardoso-Moreira, M., Borevitz, J. O. & Long, M. Natural selection shapes genome-wide patterns of copy-number polymorphism in *Drosophila melanogaster*. *Science* **320,** 1629–1631 (2008).

118. Nurminsky, D., Nurminskaya, M., De Aguiar, D. & Hartl, D. Selective sweep of a newly evolved sperm-specific gene in Drosophila. *Nature* **396,** 572–575 (1998).

119. Yeh, S.-D. *et al.* Functional evidence that a recently evolved Drosophila sperm-specific gene boosts sperm competition. *Proceedings of the National Academy of Sciences of the United States of America* **109,** 2043–2048 (2012).

120. Potrzebowski, L. *et al.* Chromosomal Gene Movements Reflect the Recent Origin and Biology of Therian Sex Chromosomes. *PLoS Biology* **6,** e80 (2008).

121. Rathke, C. *et al.* Distinct functions of Mst77F and protamines in nuclear shaping and chromatin condensation during Drosophila spermiogenesis. *European Journal of Cell Biology* **89,** 326–338 (2010).

122. Gärtner, S. M. K. *et al.* The HMG-box-containing proteins tHMG-1 and tHMG-2 interact during the histone-to-protamine transition in Drosophila spermatogenesis. *European Journal of Cell Biology* **94,** 46–59 (2015).

123. Rogers, R. L. *et al.* Landscape of Standing Variation for Tandem Duplications in *Drosophila yakuba* and *Drosophila simulans*. *Molecular Biology and Evolution* **31,** 1750–1766 (2014).

124. Ma, Y., Creanga, A., Lum, L. & Beachy, P. A. Prevalence of off-target effects in Drosophila RNA interference screens. *Nature* **443,** 359–363 (2006).

125. Green, E. W., Fedele, G., Giorgini, F. & Kyriacou, C. P. A Drosophila RNAi collection is subject to dominant phenotypic effects. *Nature methods* **11,** 222–223 (2014).

126. Doench, J. G. & Sharp, P. A. Specificity of microRNA target selection in translational repression. *Genes & Development* **18,** 504–511 (2004).

127. Hu, Y. *et al.* FlyPrimerBank: an online database for Drosophila melanogaster gene expression analysis and knockdown evaluation of RNAi reagents. *G3* **3,** 1607–1616 (2013).

128. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–60 (2009).

129. *Picard Tools* <%7Bhttps://broadinstitute.github.io/picard/%7D>.

130. *GATK Best Practices Workflow for v3.4* 2015. <%7Bhttps://software.broadinstitute.org/gatk/best-practices/%7D>.

131. McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyuzing next-generation DNA sequencing data. *Genome Research* **20,** 1297–1303 (2010).

132. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics* **43,** 491–498 (2011).

133. Van der Auwera, G. A. *et al.* in *Current Protocols in Bioinformatics* (John Wiley & Sons, Inc., 2013).

134. Blanchette, M. *et al.* Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. *Genome Research* **14,** 708–715 (2004).

135. Harris, R. S. *Improved Pairwise Alignment of Genomic DNA* PhD thesis (Pennsylvania State University, 2007), 84.

136. Edgar, R. C. MUSCLE : multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32,** 1792–1797 (2004).

137. Abascal, F., Zardoya, R. & Telford, M. J. TranslatorX : multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Research* **38,** 7–13 (2010).

138. Sattath, S., Elyashiv, E., Kolodny, O., Rinott, Y. & Sella, G. Pervasive Adaptive Protein Evolution Apparent in Diversity Patterns around Amino Acid Substitutions in Drosophila simulans. *PLoS Genetics* **7,** e1001302 (2011).

139. Team, R. C. *R: A language and environment for statistical computing* 2004.

140. Obayashi, T. & Kinoshita, K. Rank of Correlation Coefficient as a Comparable Measure for Biological Significance of Gene Coexpression. *DNA Research* **16,** 249–260 (2009).

141. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal* **Complex Sy,** 1605 (2006).

142. Loppin, B., Lepetit, D., Dorus, S., Couble, P. & Karr, T. L. Origin and neofunctionalization of a Drosophila paternal effect gene essential for zygote viability. *Current Biology* **15,** 87–93 (2005).

143. Ashburner, M. A., Golic, K. G. & Hawley, R. S. *Drosophila: A Laboratory Handbook* 2nd ed. (Cold Spring Harbor Press, 2004).

144. Wakimoto, B. T., Lindsley, D. L. & Herrera, C. Toward a comprehensive genetic analysis of male fertility in *Drosophila melanogaster. Genetics* **167,** 207–216 (2004).

145. Comeron, J. M., Ratnappan, R. & Bailin, S. The many landscapes of recombination in Drosophila melanogaster. *PLoS genetics* **8,** e1002905 (2012).

146. Källberg, M. *et al.* Template-based protein structure modeling using the RaptorX web server. *Nature Protocols* **7,** 1511–1522 (2012).

147. Schrödinger, LLC. *The PyMOL Molecular Graphics System, Version 1.8* Nov. 2015.

148. Forbes, D. J., Travesa, A., Nord, M. S. & Bernis, C. Nuclear transport factors: Global regulation of mitosis. *Current Opinion in Cell Biology* **35,** 78–90 (2015).

149. Vetter, I. R., Arndt, A., Kutay, U., Görlich, D. & Wittinghofer, A. Structural view of the Ran-importin-β interaction at 2.3 angstroms resolution. *Cell* **97,** 635–646 (1999).

150. Timinszky, G. *et al.* The importin-beta P446L dominant-negative mutant protein loses RanGTP binding ability and blocks the formation of intact nuclear envelope. *Journal of cell science* **115,** 1675–1687 (2002).

151. Kalamegham, R., Sturgill, D., Siegfried, E. & Oliver, B. Drosophila *mojoless*, a retroposed GSK-3, has functionally diverged to acquire an essential role in male fertility. *Molecular Biology and Evolution* **24,** 732–742 (2007).

152. Female Sterile Mutations on the Second Chromosome of *Drosophila melanogaster*. II. Mutations Blocking Oogenesis or Altering Egg Morphology. *Genetics* **129,** 1119–1136 (1991).

153. The *Ketel*^{D} Dominant-Negative Mutations Identify Maternal Function of the Drosophila Importin-*beta* Gene Required for Cleavage Nuclei Formation. *Genetics* **156,** 1901–1912 (2000).

154. Larracuente, A. M. & Presgraves, D. C. The selfish *Segregation Distorter* gene complex of *Drosophila melanogaster. Genetics* **192,** 33–53 (2012).

155. Sandler, L., Hiraizumi, Y. & Sandler, I. Meiotic Drive in Natural Populations of *Drosophila melanogaster*. I. the Cytogenetic Basis of Segregation-Distortion. *Genetics* **44,** 233–250 (1959).

156. Kusano, A., Staber, C. & Ganetzky, B. Nuclear Mislocalization of Enzymatically Active RanGAP Causes Segregation Distortion in Drosophila. *Developmental Cell* **1,** 351–361 (2001).

157. Kusano, A., Staber, C. & Ganetzky, B. Segregation distortion induced by wild-type RanGAP in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America* **99,** 6866–6870 (2002).

158. Hauschteck-Jungen, E. & Hartl, D. L. DNA distribution in spermatid nuclei of normal and segregation distorter males of *Drosophila melanogaster. Genetics* **89,** 15–35 (1978).

159. Nolte, V., Pandey, R. V., Kofler, R. & Schloẗterer, C. Genome-wide patterns of natural variation reveal strong selective sweeps and ongoing genomic conflict in *Drosophila mauritiana. Genome Research* **23,** 99–110 (2013).

160. Merrill, C., Bayraktaroglu, L., Kusano, A. & Ganetzky, B. Truncated RanGAP encoded by the Segregation Distorter locus of Drosophila. *Science* **283,** 1742–1745 (1999).

161. Tracy, C., Río, J., Motiwale, M., Christensen, S. M. & Betrán, E. Convergently recruited nuclear transport retrogenes are male biased in expression and evolving under positive selection in Drosophila. *Genetics* **184,** 1067–1076 (2010).

162. Presgraves, D. C. Does genetic conflict drive rapid molecular evolution of nuclear transport genes in Drosophila? *BioEssays* **29,** 386–391 (2007).

163. Baker, C. L. *et al.* PRDM9 Drives Evolutionary Erosion of Hotspots in *Mus musculus* through Haplotype-Specific Initiation of Meiotic Recombination. *PLoS Genetics* **11,** e1004916 (2015).

164. Davies, B. *et al.* Re-engineering the zinc fingers of PRDM9 reverses hybrid sterility in mice. *Nature* **530,** 171–176 (2016).

165. Rice, W. R. Sex chromosomes and the evolution of sexual dimorphism. *Evolution* **38,** 735–742 (1984).

166. Rice, W. R. Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. *Nature* **381,** 232–234 (1996).

167. Chippindale, A. K., Gibson, J. R. & Rice, W. R. Negative genetic correlation for adult fitness between sexes reveals ontogenetic conflict in Drosophila. *Proceedings of the National Academy of Sciences* **98,** 1671–1675 (2001).

168. Cox, R. M. & Calsbeek, R. Sexually antagonistic selection, sexual dimorphism, and the resolution of intralocus sexual conflict. *The American Naturalist* **173,** 176–87 (2009).

169. Bonduriansky, R. & Chenoweth, S. F. Intralocus sexual conflict. *Trends in Ecology & Evolution* **24,** 280–8 (2009).

170. Innocenti, P. & Morrow, E. H. The sexually antagonistic genes of *Drosophila melanogaster*. *PLoS Biology* **8,** e1000335 (2010).

171. Mank, J. E., Hultin-Rosenberg, L., Zwahlen, M. & Ellegren, H. Pleiotropic Constraint Hampers the Resolution of Sexual Antagonism in Vertebrate Gene Expression. *The American Naturalist* **171,** 35–43 (2008).

172. Connallon, T. & Clark, A. G. The resolution of sexual antagonism by gene duplication. *Genetics* **187,** 919–937 (2011).

173. Bassett, A. & Liu, J. L. CRISPR/Cas9 mediated genome engineering in Drosophila. *Methods* **69,** 128–136 (2014).

174. *flyCRISPR Optimal Target Finder* Accessed: 10-7-2016. <%7Bhttp://flycrispr.molbio.wisc.edu/%7D>.

175. Port, F., Chen, H.-M., Lee, T. & Bullock, S. L. Optimized CRISPR/Cas tools for efficient germline and somatic genome engineering in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America* **111,** E2967–76 (2014).

176. Bonaccorsi, S., Giansanti, M. G., Cenci, G. & Gatti, M. Immunostaining of Drosophila testes. *Cold Spring Harbor Protocols* **6,** 1273–1275 (2011).

177. Bonaccorsi, S., Giansanti, M. G., Cenci, G. & Gatti, M. Chromatin staining of Drosophila testes. *Cold Spring Harbor Protocols* **7,** 911–912 (2012).

178. Abràmoff, M. D., Magalhães, P. J. & Ram, S. J. Image processing with imageJ. *Biophotonics International* **11,** 36–41 (2004).

179. Spofford, J. Heterosis and the evolution of duplications. *The American Naturalist* **111,** 1169–1194 (1977).

180. Clark, A. G. Invasion and maintenance of a gene duplication. *Proceedings of the National Academy of Sciences of the United States of America* **91,** 2950–2954 (1994).

181. Sebat, J. *et al.* Large-Scale Copy Number Polymorphism in the Human Genome. *Science* **305,** 525–528 (2004).

182. Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nature Genetics* **36,** 949–951 (2004).

183. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444,** 444–454 (2006).

184. Sebat, J. *et al.* Strong association of de novo copy number mutations with autism. *Science* **316,** 445–449 (2007).

185. Conrad, D. F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464,** 704–712 (2010).

186. Cao, J. *et al.* Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nature Genetics* **43,** 956–963 (2011).

187. Zichner, T. *et al.* Impact of genomic structural variation in *Drosophila melanogaster* based on population-scale sequencing. *Genome Research* **23,** 568–579 (2013).

188. Usher, C. L. & McCarroll, S. A. Complex and multi-allelic copy number variation in human disease. *Briefings in Functional Genomics* **14,** 329–338 (2015).

189. Nuttle, X. *et al.* Emergence of a Homo sapiens-specific gene family and chromosome 16p11.2 CNV susceptibility. *Nature* **536,** 205–209 (2016).

190. Marques, A. C., Dupanloup, I., Vinckenbosch, N., Reymond, A. & Kaessmann, H. Emergence of Young Human Genes After a Burst of Retroposition in Primates. *PLoS Biology* **3,** e357 (2005).

191. Kimura, M. & Ohta, T. The average number of generations until fixation of a mutant gene in a finite population. *Genetics* **61,** 763–771 (1969).

192. Osada, N. & Innan, H. Duplication and gene conversion in the Drosophila melanogaster genome. *PLoS genetics* **4,** e1000305 (2008).

193. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25,** 2865–2871 (2009).

194. Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nature Reviews Genetics* **12,** 363–376 (2011).

195. Rausch, T. *et al.* DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28,** i333–i339 (2012).

196. Langley, C. H., Crepeau, M., Cardeno, C., Corbett-Detig, R. & Stevens, K. Circumventing heterozygosity: sequencing the amplified genome of a single haploid *Drosophila melanogaster* embryo. *Genetics* **188,** 239–246 (2011).

197. Halligan, D. L. & Keightley, P. D. Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison. *Genome Research* **16,** 875–884 (2006).

198. Rogers, R. L. *et al.* Tandem duplications and the limits of natural selection in *Drosophila yakuba* and *Drosophila simulans*. *PLoS ONE* **10,** 1–28 (2015).

199. Cardoso-Moreira, M. *et al.* Evidence for the fixation of gene duplications by positive selection in Drosophila. *Genome Research* **26,** 787–798 (2016).

200. Stephan, W., Wiehe, T. H. E. & Lenz, M. The Effect of Strongly Selected Substitutions on Neutral Polymorphism : Analytical Results Based on Diffusion Theory. *Theoretical Population Biology* **41,** 237–254 (1992).

201. Hudson, R. R., Bailey, K., Skarecky, D., Kwiatowski, J. & Ayala, F. J. Evidence for positive selection in the superoxide dismutase (Sod) region of *Drosophila melanogaster*. *Genetics* **136,** 1329–1340 (1994).

202. Braverman, J. M., Hudson, R. R., Kaplan, N. L., Langley, C. H. & Stephan, W. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* **140,** 783–796 (1995).

203. Llopart, A., Comeron, J. M., Brunet, G., Lachaise, D. & Long, M. Intron presence absence polymorphism in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America* **99,** 8121–8126 (2002).

204. Sabeti, P. C. *et al.* Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419,** 832–837 (2002).

205. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS biology* **4,** e72 (2006).

206. Ferrer-Admetlla, A., Liang, M., Korneliussen, T. & Nielsen, R. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Molecular Biology and Evolution* **31,** 1275–1291 (2014).

207. Thornton, K. R. The Neutral Coalescent Process for Recent Gene Duplications and Copy-Number Variants. *Genetics* **177,** 987–1000 (2007).

208. Teshima, K. M. & Innan, H. The coalescent with selection on copy number variants. *Genetics* **190,** 1077–1086 (2012).

209. Bertone, P. *et al.* Global Identification of Human Transcribed Sequences with Genome Tiling Arrays. *Science* **306,** 2242–2246 (2004).

210. Nagalakshmi, U. *et al.* The transcriptional landscape of the yeast genome defined by RNA sequencing. **320,** 1344–1349 (2008).

211. Brown, J. B. *et al.* Diversity and dynamics of the Drosophila transcriptome. *Nature* **512** (2014).

212. Robinson, J. T. *et al.* Integrative Genomics Viewer. *Nature Biotechnology* **29,** 24–26 (2011).

213. Näsvall, J., Sun, L., Roth, J. R. & Andersson, D. I. Real-Time Evolution of New Genes by Innovation, Amplification, and Divergence. *Science* **338,** 384–387 (2012).

214. Loehlin, D. W. & Carroll, S. B. Expression of tandem gene duplicates is often greater than twofold. *Proceedings of the National Academy of Sciences* **113,** 201605886 (2016).

215. Maroni, G., Wise, J., Young, J. E. & Otto, E. Metallothionein gene duplications and metal tolerance in natural populations of *Drosophila melanogaster*. *Genetics* **117,** 739–744 (1987).

216. Harrop, T. W. R. *et al.* Evolutionary changes in gene expression, coding sequence and copy-number at the Cyp6g1 locus contribute to resistance to multiple insecticides in Drosophila. *PloS one* **9,** e84879 (2014).

217. Maruyama, T. & Kimura, M. A Note on the Speed of Gene Frequency Changes in Reverse Directions in a Finite Population. *Evolution* **28,** 161–163 (1974).

218. Maruyama, T. The Age of a Rare Mutant Gene in a Large Population. *American Journal of Human Genetics* **26,** 669–673 (1974).

219. Mackay, T. F. C. *et al.* The Drosophila melanogaster Genetic Reference Panel. *Nature* **482,** 173–178 (2012).

220. Roberts, P. A. A tandem duplication that lowers recombination throughout a chromosome arm of Drosophila melanogaster. *Genetics* **54,** 969–979 (1966).

221. Roberts, P. A. & Broderick, D. J. Properties and evolution potential of newly induced tandem duplications in *Drosophila melanogaster*. *Genetics* **102,** 75–89 (1982).

222. Lowe, B., Mathern, J. & Hake, S. Active Mutator elements suppress the knotted phenotype and increase recombination at the Kn1-O tandem duplication. *Genetics* **132,** 813–822 (1992).

223. Stephan, W. & Li, H. The recent demographic and adaptive history of Drosophila melanogaster. *Heredity* **98,** 65–68 (2007).

224. Singh, N. D., Jensen, J. D., Clark, A. G. & Aquadro, C. F. Inferences of demography and selection in an african population of *Drosophila melanogaster*. *Genetics* **193,** 215–228 (2013).

225. Bartenhagen, C. & Dugas, M. Genome analysis RSVSim : an R / Bioconductor package for the simulation of structural variations. **29,** 1679–1681 (2013).

226. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: A next-generation sequencing read simulator. *Bioinformatics* **28,** 593–594 (2012).

227. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–2 (2010).

228. Homer, N., Merriman, B. & Nelson, S. F. BFAST: An alignment tool for large scale genome resequencing. *PLoS ONE* **4** (2009).

229. Homer, N., Merriman, B. & Nelson, S. F. Local alignment of two-base encoded DNA sequence. *BMC bioinformatics* **10,** 175 (2009).

230. True, J. R. & Haag, E. S. Developmental system drift and flexibility in evolutionary trajectories. *Evolution and Development* **3,** 109–119 (2001).

231. Bartoszewski, S., Luschnig, S., Desjeux, I., Grosshans, J. & Nusslein-Volhard, C. Drosophila p24 homologues eclair and baiser are necessary for the activity of the maternally expressed Tkv receptor during early embryogenesis. *Mechanisms of Development* **121,** 1259–1273 (2004).

232. Saleem, S. *et al.* Drosophila melanogaster p24 trafficking proteins have vital roles in development and reproduction. *Mechanisms of Development* **129,** 177–191 (2012).

233. Reverse Genetic Screening Reveals Poor Correlation Between Morpholino-Induced and Mutant Phenotypes in Zebrafish. *Developmental Cell* **32,** 97–108 (2015).

234. Rossi, A. *et al.* Genetic compensation induced by deleterious mutations but not gene knockdowns. *Nature* **524,** 230–233 (2015).

235. Bellen, H. J. *et al.* The BDGP Gene Disruption Project : Single Transposon Insertions Associated With 40 % of Drosophila Genes. *Genetics* **167,** 761–781 (2004).

236. Laviolette, M. J., Nunes, P., Peyre, J.-B., Aigaki, T. & Stewart, B. A. A Genetic Screen for Suppressors of Drosophila NSF2 Neuromuscular Junction Overgrowth. *Genetics* **170,** 779–792 (2005).

237. Tirmarche, S. *et al.* Drosophila Protamine-Like Mst35Ba and Mst35Bb Are Required for Proper Sperm Nuclear Morphology but Are Dispensable for Male Fertility. *G3* **4,** 2241–2245 (2014).

238. Weber, U., Gault, W. J., Olguin, P., Serysheva, E. & Mlodzik, M. Novel Regulators of Planar Cell Polarity : A Genetic Analysis in Drosophila. *Genetics* **191,** 145–162 (2012).

239. Neumuller, R. A. *et al.* Resource Genome-Wide Analysis of Self-Renewal in Drosophila Neural Stem Cells by Transgenic RNAi. *Cell Stem Cell* **8,** 580–593 (2011).

240. Hadar, N. *et al.* A screen identifying genes responsive to Dpp and Wg signaling in the Drosophila developing wing. *Gene* **494,** 65–72 (2012).

241. Orihara-Ono, M. *et al.* The slender lobes gene , identified by retarded mushroom body development , is required for proper nucleolar organization in Drosophila. *Developmental Biology* **281,** 121–133 (2005).

242. Tao, Y., Masly, J. P., Araripe, L., Ke, Y. & Hartl, D. L. A sex-ratio Meiotic Drive System in Drosophila simulans. I: An Autosomal Suppressor. *PLoS Biology* **5,** e292 (2007).

243. Tao, Y. *et al.* A sex-ratio Meiotic Drive System in Drosophila simulans. II: An X-linked Distorter. *PLoS Biology* **5,** e293 (2007).

244. Vicoso, B. & Charlesworth, B. The Deficit of Male-Biased genes on the D. melanogaster X Chromosome is Expression-Dependent: A Consequence of Dosage Compensation? *Journal of Molecular Evolution* **68,** 576–83 (2009).

245. Metta, M. & Schlötterer, C. Non-random genomic integration - an intrinsic property of retrogenes in Drosophila? *BMC Evolutionary Biology* **10,** 114 (2010).

246. Vibranovski, M. D. *et al.* Segmental dataset and whole body expression data do not support the hypothesis that non-random movement is an intrinsic property of Drosophila retrogenes. *BMC Evolutionary Biology* **12,** 169 (2012).

247. Ellegren, H. & Parsch, J. The evolution of sex-biased genes and sex-biased gene expression. *Nature Reviews Genetics* **8,** 689–698 (2007).

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

Appendices

APPENDIX A

MAIN SUPPLEMENT

A.0.1 Supplementary Note #1: eca and p24-2

eca and *p24-2* were formed by tandem duplication specifically in *D. melanogaster* and appear to be present only in the reference genome sequence (Table A.1). That is, I found no evidence that both copies are present in any of the 109 DPGP2 genomes [83].

However, several studies have provided evidence that both gene copies are essential [63, 231, 232]. *eca* is likely essential, as disruptive point mutations cause development to fail prior to hatching [231]. The only evidence that *p24-2* is essential comes from constitutive RNAi using a single specific RNAi line, KK109179 [63, 232]. However, we discovered that this RNAi line contains an insertion near *tiptop* [125]. Recombining out this bad insertion and re-testing constitutive knockdown produced no lethal effect (*Act5C::GAL4*: balancer F1s = 91, RNAi F1s = 102, $\chi^2 = 4.62$, $p = 0.032$; *αTub84B::GAL4*: balancer F1s = 70, RNAi F1s = 74, $\chi^2 = 0.13$, $p = 0.72$). I obtained similar results with a new RNAi line I constructed. I note that Chen et al. (2010) used RNAi line GD5843 to suggest that *p24-2* knockdown was lethal, but this line hits both *eca* and *p24-2* and probably causes lethality through *eca* knockdown.

Altogether, the data suggest that *eca* is essential while *p24-2* is not. *p24-2* has recruited an additional 43 amino acids into its 3' end that are not present in *eca* or is orthologs from *D. ananassae, D. yakuba, D. erecta, Dsechellia*, or *D. simulans*, and it should be defined as the 'new' gene copy.

A.0.2 Supplementary Note #2: tHMG1 Essentiality Contrasts with Gärtner et al. (2015) EJCB

I showed in Chapter 2 that constitutive RNAi of the *Drosophila melanogaster*-specific gene *tHMG1* caused lethality early in fly development. This experiment was repeatable with multiple independent RNAi lines and drivers (Figure 2.2; Tables A.3, A.4). Furthermore, *tHMG1* knockdown in the lines I tested using qPCR appeared to be specific and fairly strong (Figure 2.2b). I concluded in that chapter that *tHMG1* is an essential species-specific gene.

However, Gärtner et al. (2015) used P-element mobilization to generate a deletion that appears to encompass *tHMG1* and the 5' end of *tHMG2* and found that homozygous Δ*tHMG1/tHMG2* flies exhibited no defects in morphology or spermatogenesis [122]. These authors did not measure the effect of Δ*tHMG1/tHMG2* on male reproductive output.

I think that the discrepancy between these results is caused by compensation in the knockout mutants by other genes in the same gene network. It has been well documented in zebrafish that knocking out or knocking down a particular gene can produce a different phenotype [233, 234]. Gene knockouts are somehow recognized by the cell and compensated for by differential expression of other genes in the gene interaction network in which the knocked-out gene participates. Conversely, expression knockdowns do not appear to be recognized and compensated for and frequently result in stronger phenotypes than knockouts of the same gene [234]. Indeed, Gärtner et al. (2015) show that related proteins, like *hmgz*, are upregulated in Δ*tHMG1/tHMG2* mutant flies, which suggests that this related protein may act to compensate for *tHMG1/tHMG2* loss.

It is also possible that *tHMG1* and *tHMG2* antagonize each other and simultaneous loss of both copies relieves the antagonism and mitigates the phenotype.

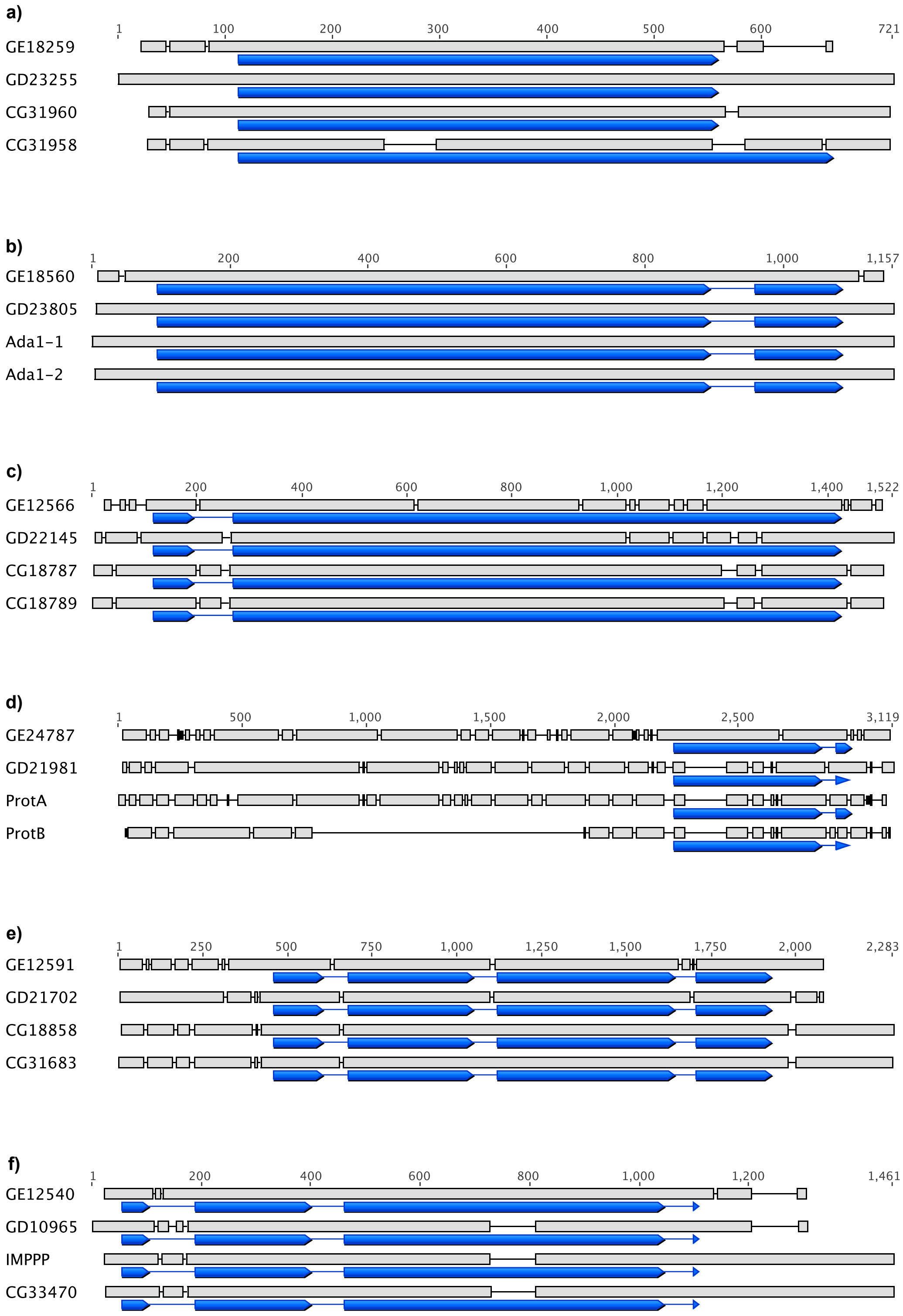


Figure A.1: *D. melanogaster*-specific duplicate gene structure divergence. *Continued on next page*

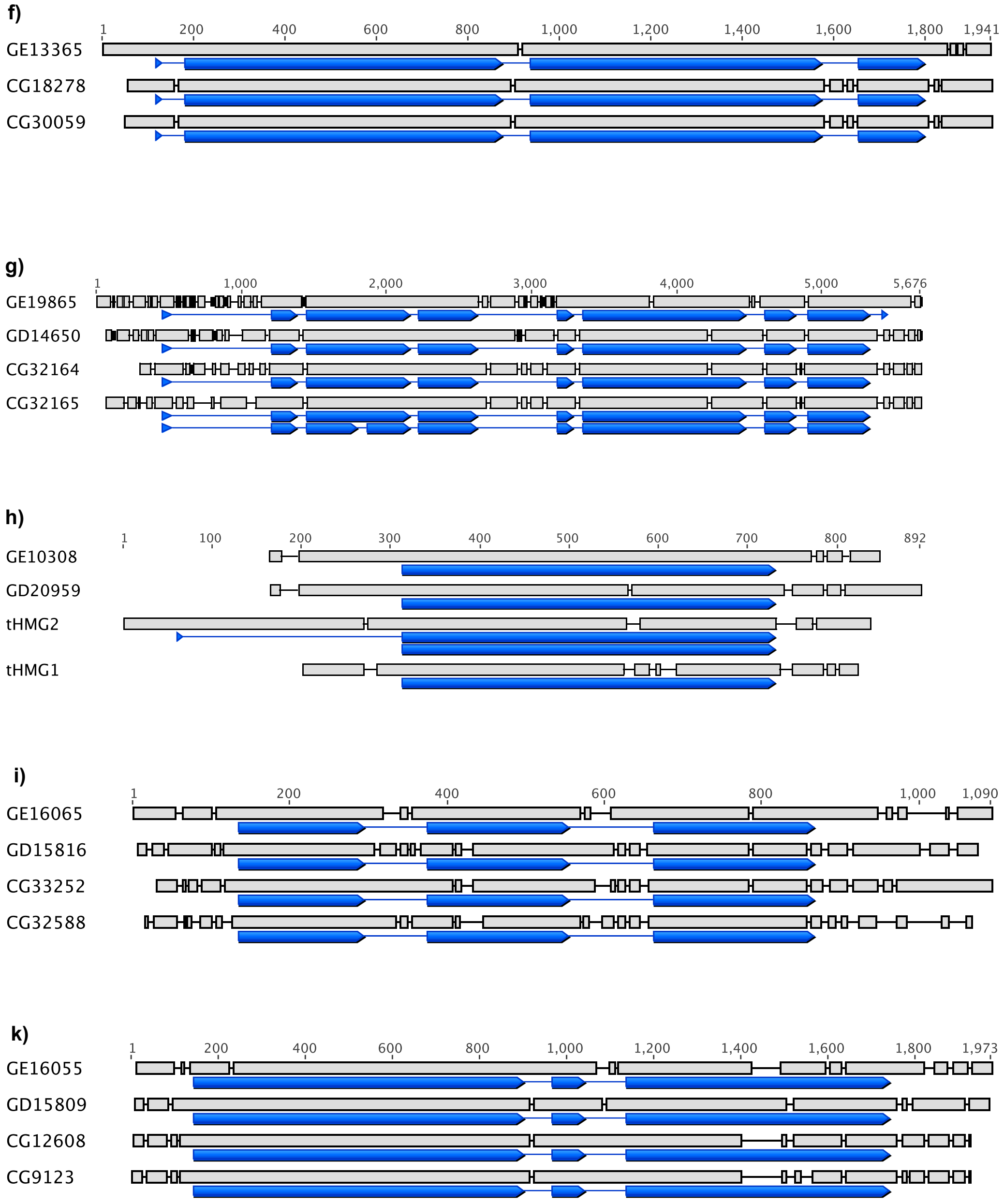


Figure A.1 continued. *D. melanogaster*-specific duplicate gene structure divergence. Multiple sequence alignments of new and parent genes and their *D. simulans* and *D. yakuba* orthologs. Blue bars underneath the gray (nucleotide sequence) bars denote the coding sequence. Alignment blocks are outlined in black. Chimeric genes (*Qtzl*, *CG31687*, and *CG12592*) are not shown. For each alignment, the sequences are top to bottom: *D. yakuba*, *D. simulans*, *D. melanogaster* parent, *D. melanogaster* new.

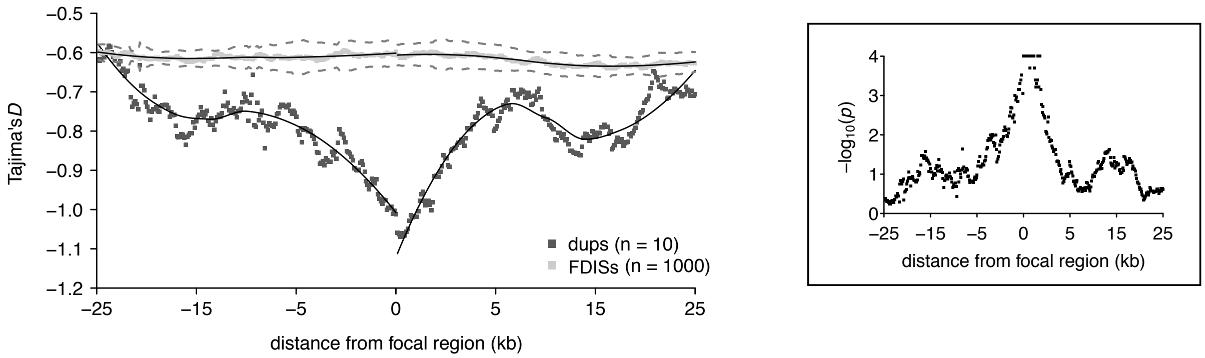


Figure A.2: Mean Tajima's *D* in regions flanking *D. melanogaster*-specific duplications versus intergenic sites that mutated and fixed specifically in *D. melanogaster*. *D* was calculated in 5 kb sliding windows (100 bp step) in the regions flanking duplications (dups) or fixed diverged intergenic sites (FDISs). Solid lines are smoothed LOESS fits using with a 0.75 span. Separate fits were calculated for the left and right flanking regions. Dotted lines represent 95% confidence intervals on the means for each window. Probability values for each window (right) were determined using permutations.

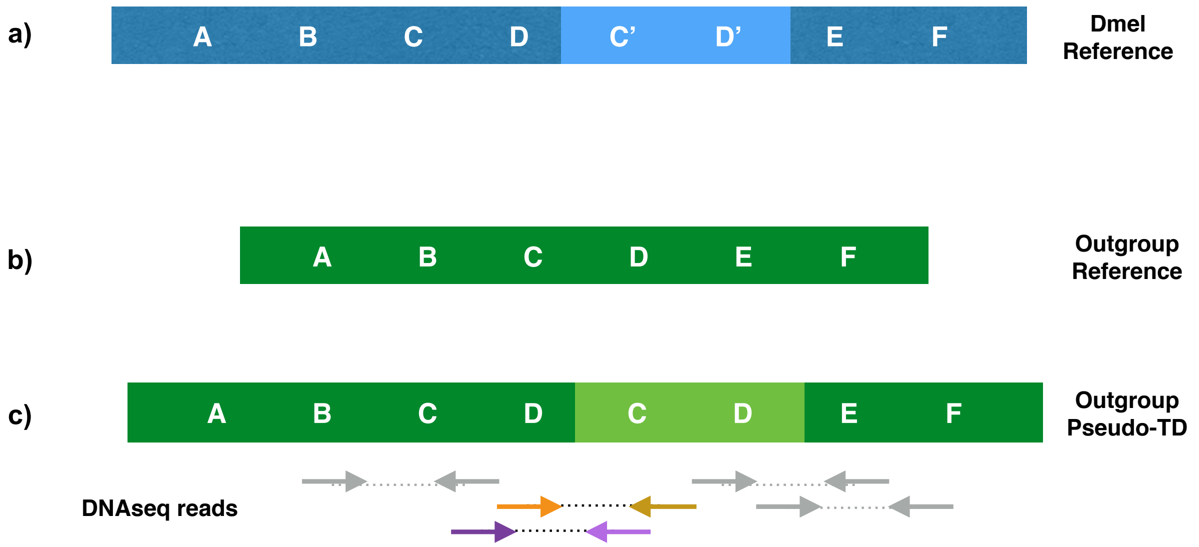
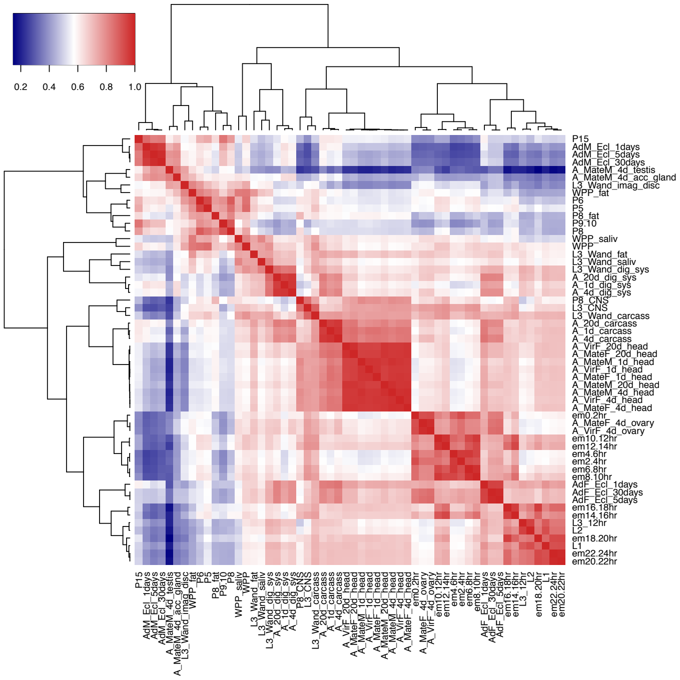


Figure A.3: Scheme to test if *D. melanogaster*-specific tandem duplications exist in *D. simulans* or *D. yakuba* population genomic data from Rogers et al. 2014. A - F are genome regions. **a)** The C - D and C' - D' segments are tandem duplicate copies in the Dmel reference genome (blue bar), but single-copy in **(b)** an outgroup genome (green bar). If the region is also duplicated in the outgroup, but missing in the outgroup reference genome due to misassembly (e.g. light green C - D in **c**), we expect that next-generation sequencing should produce paired-end reads (arrows) spanning the unique D → C breakpoint. We simulated the D → C breakpoint breakpoint by pasting two copies of the outgroup C - D region together, added this pseudo-TD sequence to the outgroup reference genome FASTA file, and re-mapped population genomic sequencing reads the pseudo-reference genome. We parsed the resulting BAM files for properly-mapped read pairs to determine if the D → C breakpoint junction existed in any of the 20 sequenced *D. simulans* or 20 *D. yakuba* genomes [123].



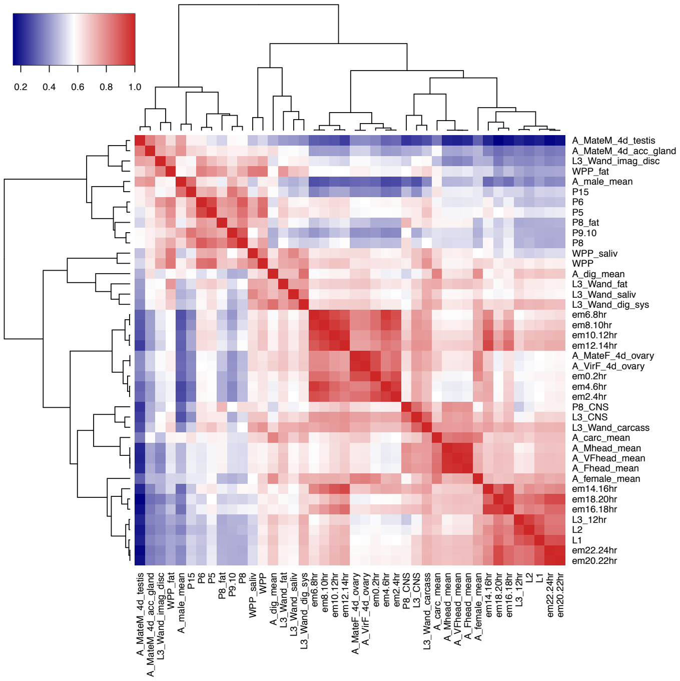


Figure A.4: Correlations between modENCODE expression data collected from FlyBase 2015_05. FPKM values were used to generate a correlation heatmap in R 3.0.1. Large blocks of highly-correlated datasets (e.g. all adult male datasets) provide little additional data and reduce power to detect true correlations (top). While not complete, averaging carcass, digestive system, mated female head, virgin female head, male head, male whole body, and female whole body samples reduces this redundancy (bottom).

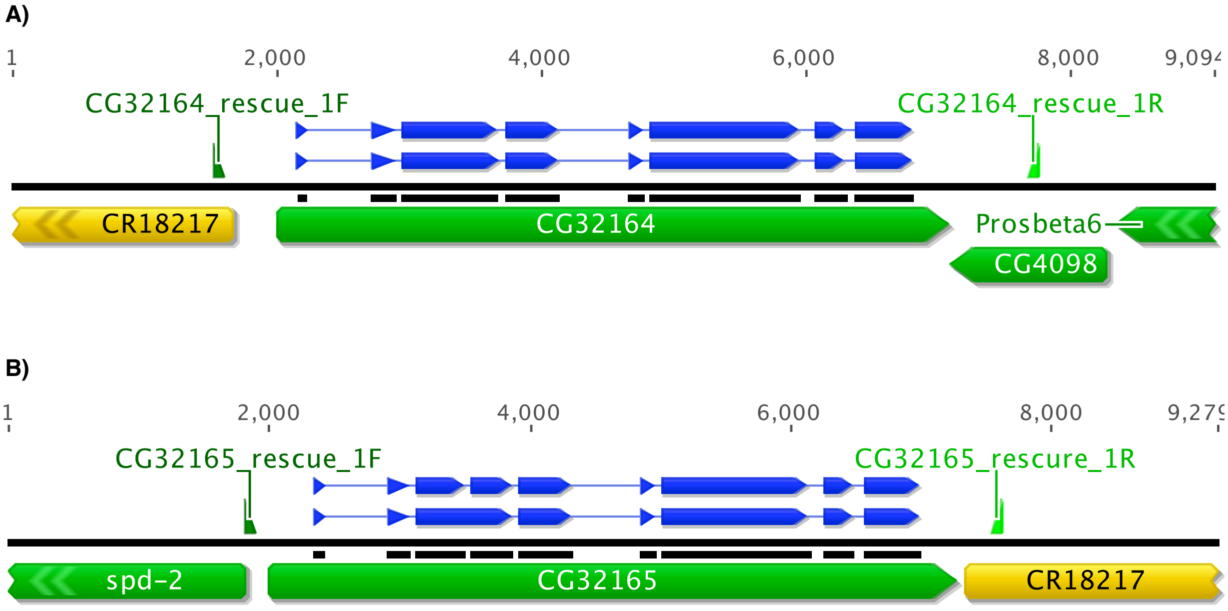


Figure A.5: Generation of *CG32164* and *CG32165* rescue constructs. *CG32164* (A) or *CG32165* (B) gene regions and 500 bp upstream and downstream were amplified with the primers below using the MasterAmp Extra-Long PCR Kit (Epicentre, USA). PCR products were then cloned into BamHI-digested pVerm using the HiFi DNA Assembly Master Mix (NEB, USA). Rescue constructs were then injected into stock 25709 $(y^1v^1P\{y^{+t7.7} = nos - \Phi C31 - int.NLS\}X; P\{y^{+t7.7} = CaryP\}attP40)$ and isolated by crossing to $y^1v^1; \frac{sna^{Sco}}{CyO}$ (TRiP Toolkit Stock). I introgressed the rescue constructs into either $CG32164^{p7f7}$ or $CG32165^{p4g2}$ using a crossing scheme involving stock 7198 $(w*; \frac{Kr^{If-1}}{CyO}; \frac{D^1}{TM3,Ser^1})$ (Figure A.6). Black bars are DNA sequence, green annotations are gene regions, blue annotations are coding exons, yellow annotations are pseudogenes, and green triangles are primer sites. Primers, where underlined regions are those that overlap with BamHI-digested pVerm: CG32165_rescue_1F 5'-CGCGAATGCATCTAGATATCGGATCCGCGTCGCTTCGATCA; CG32165_rescue_1R 5'-CTCTGCAGTCGACGGGCCCGGGATCCGGGAGGCGTTCAGGTATAC; CG32164_rescue_1F 5'-TCGGTACCTCGCGAATGCATCTAGATATCGTCCACTGCTTCTGTCCATTCC; CG32164_rescue_1R 5'-CTCTGCAGTCGACGGGCCCGGGGCAGATGGGGTCCAAATCA

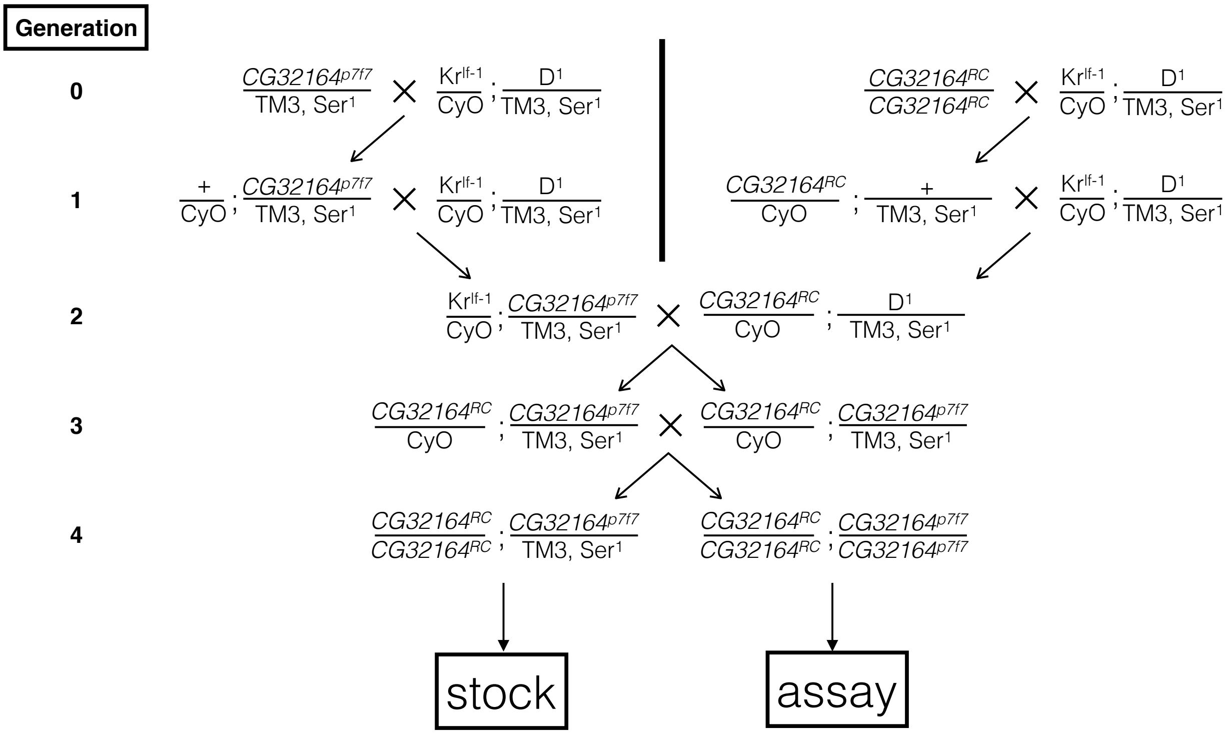

Generation
0
$\frac{CG32164^{p7f7}}{TM3, Ser^1}$ × $\frac{Kr^{If-1}}{CyO}$; $\frac{D^1}{TM3, Ser^1}$
$\frac{CG32164^{RC}}{CG32164^{RC}}$ × $\frac{Kr^{If-1}}{CyO}$; $\frac{D^1}{TM3, Ser^1}$
1
$\frac{+}{CyO}$; $\frac{CG32164^{p7f7}}{TM3, Ser^1}$ × $\frac{Kr^{If-1}}{CyO}$; $\frac{D^1}{TM3, Ser^1}$
$\frac{CG32164^{RC}}{CyO}$; $\frac{+}{TM3, Ser^1}$ × $\frac{Kr^{If-1}}{CyO}$; $\frac{D^1}{TM3, Ser^1}$
2
$\frac{Kr^{If-1}}{CyO}$; $\frac{CG32164^{p7f7}}{TM3, Ser^1}$ × $\frac{CG32164^{RC}}{CyO}$; $\frac{D^1}{TM3, Ser^1}$
3
$\frac{CG32164^{RC}}{CyO}$; $\frac{CG32164^{p7f7}}{TM3, Ser^1}$ × $\frac{CG32164^{RC}}{CyO}$; $\frac{CG32164^{p7f7}}{TM3, Ser^1}$
4
$\frac{CG32164^{RC}}{CG32164^{RC}}$; $\frac{CG32164^{p7f7}}{TM3, Ser^1}$
$\frac{CG32164^{RC}}{CG32164^{RC}}$; $\frac{CG32164^{p7f7}}{CG32164^{p7f7}}$
stock
assay

Figure A.6: Crossing scheme to introgress rescue constructs into deletion lines using stock 7198. The same scheme was used to introgress $CG32165^{RC}$ into $CG32165^{p4g2}$.

Table A.1: Putative *D. melanogaster*-specific tandem duplications before frequency filter.

5' Copy Coordinates[a]	**5' Copy Genes**	**3' Copy Coordinates**[a]	**3' Copy Genes**	***D. simulans* (r1.0) Coordinates**	***D. yakuba* (r1.3) Coordinates**	**Freq.**[c]
2L:3,784,873-3,785,655	*CG31960*	2L:3,785,665-3,786,386	*CG31958*	2L:3,741,618-3,742405	2L:3,786,936-3,787,604	15/17
2L:7,729,139-7,735,283	*CG31904*[b]	2L:7,735,283-7,741,436	-	2L:7,527,415-7,534,160	2L:5,161,273-5,167,589	0/17
2L:11,992,235-11,996,148	*CG18789,* *Ada1-2,* *Qtzl*[b]	2L:11,996,148-12,000,059	*CG18787,* *Ada1-1*	2L:11,800,218-11,804,142	2L:8,415,089-8,419,012	17/17
2L:14,878,773-14,881,853	*ProtA*	2L:14,881,853-14,883,944	*ProtB*	2L:14,628,304-14,631,431	2L:14,984,014-14,987,345	17/17
2L:20,442,296-20,451,418	*CG31683,* *CG31687*[a]	2L:20,451,418-20,460,527	*CG18853*	2L:20,012,997-20,022,064	2R:6,807,637-6,816,500	16/17
2R:6,995,222-7,001,513	*Tsp42Eb*	2R:7,829,850-7,832,960	*CG30160*	2R:2,434,118-2,437,303	2L:15,604,394-15,609,733	0/17
2R:7,826,739-7,829,850	*Prosalpha1,* *CG18853*[b]	2R:7,001,513-7,007,802	*CG30382*	2R:1,689,836-1,695,893	2L:16,380,581-16,383,696	0/17
2R:13,405,871-13,410,599	*CG18278,* *IMPPP*	2R:13,410,599-13,415,337	*CG30059,* *CG33470*	2R:7,751,868-7,756,507	2R:7,657,726-7,662,488	15/17
3L:1,245,256-1,246,250	*CG32318*[b]	3L:1,246,250-1,247,243	-	3L:843,999-845,000	3L:1,204,048-1,205,047	0/17
3L:16,593,391-16,601,091	*CG32165*	3L:16,601,091-16,608,783	*CG32164*	3L:15,930,954-15,938,734	3L:17,464,667-17,472,380	17/17
3R:9,684,713-9,692,034	*eca,* *CG18542,* *Unc-115b*	3R:9,692,034-9,699,920	*p24-2,* *CG32939,* *Unc-115a*	3R:15,817,049-15,824,776	3R:9,564,211-9,572,053	0/17
3R:10,335,749-10,338,733	*CG12592*[b]	3R:10,338,733-10,341,340	-	3R:15,220,327-15,224,562	3R:10,198,389-10,201,332	17/17
3R:19,766,978-19,774,378	*CG6300*	3R:19,774,378-19,778,336	*CG11659*	3R:5,884,488-5,891,988	3R:4,356,864-4,365,646	17/17
3R:22,491,586-22,492,928	*tHMG1*	3R:22,492,928-22,498,079	*tHMG2*	3R:18,126,527-18,131,707	3R:19,163,099-19,168,252	17/17
3R:27,958,777-27,961,425	*CG31131*	3R:27,961,425-27,964,073	*CG31253*	3R:23,490,881-23,493,547	3R:22,710,346-22,713,002	0/17
∼X:6,276,000-6280,000	*Ubi-p5E*	∼X:6,280,000-6,283,900	*CG11700*	∼X:4,881,000-4,883,000	∼X:2,175,000-2,179,000	17/17
X:15,339,059-15,342,364	*CG9123*	X:15,342,364-15,345,100	*CG12608*	X:11,756,037-11,759,659	X:9,433,332-9,436,220	17/17

a: Coordinates were determined using BLAT and are accurate to within 8 bp for all duplications, except for *CG11700/Ubi-p5E*, which is unclear

b: chimeric gene, which resides in both copies

c: Frequency in 17 core RG (Rwanda) fly genomes sequenced by Pool et al. 2012.

Table A.2: New RNAi lines constructed in this study.

Line	Target	Forward primer	Reverse primer	length	19- mers	on	off	genes
NV-CG11659	*CG11659*	TCAATTGGTCAATGGGCTA	GATTCGCCGTCATCATTTTT	51	33	33	0	NA
NV-CG11700	*CG11700*	TCGACCCTTCACTTGGTCCT	CACAAAGATCTGCATGCCAC	51	33	33	0	NA
NV-CG31958	*CG31958*	ACATCAATTTCGAGGAGTCCA	TGTAATTTAGAATGGGCGTGG	54	36	36	0	NA
NV-CG32165	*CG32165*	ACCGTGCATAACATGCTGAC	GCCTTAACCACAATCTGCTCA	86	68	68	0	NA
NV-CG32588	*CG32588*	TTGACTTGAACGAAGCACCT	CTCGAAGAATGTCATGCCG	84	66	66	0	NA
NV-CG4478	*ProtB*	CGCCTATTTGAATTTCGTGC	AATTAATTCCTGCGGCTTCA	67	49	48	1	ProtA
NV-CG7045	*tHMG1*	CGATCTGCAGGAAGTGCTTA	GTAAACGCTTAGTGAAGGCACC	52	34	34	0	NA
NV-CG9123	*CG9123*	AACACACTGAAATTCACGCCT	TCATATGGCCATCATCGCT	64	46	46	0	NA
NV-CG6300	*CG6300*	CGGCATTGATGTGTTATTGC	CGATTCGATTCGTAGGGAAA	162	144	144	0	NA
NV-CG32744	*Ubi-p5E*	AGTTAGTTGTGCCACAGGGC	ATTTTGTGCCTTTGAATAAACCA	77	59	59	0	NA
NV-CG31960	*CG31960*	CAATTTCGAGGAGTTCACCAA	AATTTAGAAAGCGTGTGCGTG	65	47	47	0	NA
NV-CG32164	*CG32164*	AAGATGAGCGATCTTTAAAAGGG	AGCCTTCAATGAAACAAGTAATCC	144	126	126	0	NA
NV-CG33252	*CG33252*	TGTGCGCACCAACCTAAATA	AATTTGAGAACGCTCGAAGG	131	113	113	0	NA
NV-CG4479	*ProtA*	CAACGCCTATTTGAATTTTGTG	CCGCGGTTTCAAGTTACAGT	61	43	42	1	ProtB
NV-CG7046	*tHMG2*	AAGGAGCACCAGACGGAGT	CGTACACAAAAAGGGTTGGC	97	79	79	0	NA
NV-CG12608	*CG12608*	TTGAATAGTCGTTGTCGTTGG	CAAGAATTTGTTTCTGCTGGC	52	34	34	0	NA

Table A.3: Full Chapter 2 RNAi screen results.

line	gene	source	control	driver[a]	rep.	balF	wtF	balM	wtM
25709	25709_CTRL	TRiP	25709	act	1	39	31	24	19
25709	25709_CTRL	TRiP	25709	act	2	23	29	18	29
25709	25709_CTRL	TRiP	25709	act	3	24	20	27	28
25709	25709_CTRL	TRiP	25709	act	4	24	20	22	24
25709	25709_CTRL	TRiP	25709	tub	5	28	33	21	16
25709	25709_CTRL	TRiP	25709	tub	4	29	22	22	24
25709	25709_CTRL	TRiP	25709	tub	2	18	29	20	20
25709	25709_CTRL	TRiP	25709	tub	1	14	11	10	13
25709	25709_CTRL	TRiP	25709	tub	3	21	10	18	22
25710	25710_CTRL	TRiP	25710	act	1	38	35	30	20
25710	25710_CTRL	TRiP	25710	act	2	21	18	14	20
25710	25710_CTRL	TRiP	25710	act	3	31	36	40	19
25710	25710_CTRL	TRiP	25710	act	4	29	19	24	15
25710	25710_CTRL	TRiP	25710	tub	1	28	36	35	30
25710	25710_CTRL	TRiP	25710	tub	2	22	30	16	24
25710	25710_CTRL	TRiP	25710	tub	3	22	30	25	22
25710	25710_CTRL	TRiP	25710	tub	4	22	25	22	30
60100	60100_CTRL	NEW	60100	act	1	14	24	32	26
60100	60100_CTRL	NEW	60100	act	2	21	27	21	27
60100	60100_CTRL	NEW	60100	act	3	29	34	39	31
60100	60100_CTRL	NEW	60100	act	4	16	24	16	15
60100	60100_CTRL	NEW	60100	act	5	14	19	14	9
60100	60100_CTRL	NEW	60100	tub	1	10	12	8	14
60100	60100_CTRL	NEW	60100	tub	2	17	13	18	12
60100	60100_CTRL	NEW	60100	tub	3	26	29	17	27
60100	60100_CTRL	NEW	60100	tub	4	10	12	8	11
60100	60100_CTRL	NEW	60100	tub	5	16	16	12	26
31735	*Cdc23*	TRiP	25710	act	1	19	15	14	15
31735	*Cdc23*	TRiP	25710	act	2	23	18	15	17
31735	*Cdc23*	TRiP	25710	tub	1	14	13	8	17
31735	*Cdc23*	TRiP	25710	tub	2	10	30	19	14
55268	*CG11659*	TRiP	25709	act	1	16	22	12	36
55268	*CG11659*	TRiP	25709	act	2	16	14	22	34
55268	*CG11659*	TRiP	25709	tub	2	12	15	12	11
55268	*CG11659*	TRiP	25709	tub	3	11	19	20	16
NV-CG11659-2	*CG11659*	NEW	60100	act	3	32	29	27	28
NV-CG11659-2	*CG11659*	NEW	60100	act	1	30	27	28	28
NV-CG11659-2	*CG11659*	NEW	60100	act	2	29	32	17	26
NV-CG11659-2	*CG11659*	NEW	60100	tub	1	26	26	27	25
NV-CG11659-2	*CG11659*	NEW	60100	tub	2	29	20	25	30
NV-CG11659-2	*CG11659*	NEW	60100	tub	3	23	28	35	32
32933	*CG11700*	TRiP	25710	act	1	34	14	13	20
32933	*CG11700*	TRiP	25710	act	2	19	35	16	12
32933	*CG11700*	TRiP	25710	act	3	13	17	31	30
32933	*CG11700*	TRiP	25710	tub	1	31	16	13	22
32933	*CG11700*	TRiP	25710	tub	2	19	29	16	32
32933	*CG11700*	TRiP	25710	tub	3	27	35	30	30
NV-CG11700-1	*CG11700*	NEW	60100	act	1	33	30	34	24
NV-CG11700-1	*CG11700*	NEW	60100	act	2	19	33	17	23
NV-CG11700-1	*CG11700*	NEW	60100	act	3	13	12	19	23
NV-CG11700-1	*CG11700*	NEW	60100	act	4	26	23	23	28
NV-CG11700-1	*CG11700*	NEW	60100	tub	1	15	24	23	15
NV-CG11700-1	*CG11700*	NEW	60100	tub	2	14	16	21	16
NV-CG11700-1	*CG11700*	NEW	60100	tub	3	28	24	21	32
NV-CG11700-1	*CG11700*	NEW	60100	tub	4	28	24	10	25
44037	*CG12264*	TRiP	25709	act	1	19	0	14	0
44037	*CG12264*	TRiP	25709	tub	2	14	1	13	0
36079	*CG12592*	TRiP	25710	act	1	33	25	20	16

a: act: *Act5C::GAL4* driver or tub: *αTub84B::GAL4* driver

Table A.3: *Continued from previous page*

line	gene	source	control	driver[a]	rep.	balF	wtF	balM	wtM
36079	*CG12592*	TRiP	25710	act	2	16	14	12	7
36079	*CG12592*	TRiP	25710	tub	1	18	13	18	16
36079	*CG12592*	TRiP	25710	tub	2	15	7	16	11
36079	*CG12592*	TRiP	25710	tub	3	15	15	15	18
36079	*CG12592*	TRiP	25710	tub	4	16	17	15	19
NV-CG12608-11-4	*CG12608*	NEW	60100	act	1	11	13	7	14
NV-CG12608-11-4	*CG12608*	NEW	60100	act	2	23	14	16	18
NV-CG12608-11-4	*CG12608*	NEW	60100	act	3	18	15	24	23
NV-CG12608-11-4	*CG12608*	NEW	60100	act	5	27	31	28	31
NV-CG12608-11-4	*CG12608*	NEW	60100	act	6	21	16	19	14
NV-CG12608-11-4	*CG12608*	NEW	60100	tub	1	29	20	18	17
NV-CG12608-11-4	*CG12608*	NEW	60100	tub	2	19	30	19	24
NV-CG12608-11-4	*CG12608*	NEW	60100	tub	3	11	26	15	26
NV-CG12608-11-4	*CG12608*	NEW	60100	tub	4	18	23	19	20
NV-CG12608-11-4	*CG12608*	NEW	60100	tub	5	31	31	26	23
NV-CG12608-11-4	*CG12608*	NEW	60100	tub	6	28	28	17	31
NV-CG12608-11-7	*CG12608*	NEW	60100	act	1	20	24	12	18
NV-CG12608-11-7	*CG12608*	NEW	60100	act	2	26	30	19	28
NV-CG12608-11-7	*CG12608*	NEW	60100	act	3	20	20	19	20
NV-CG12608-11-7	*CG12608*	NEW	60100	act	4	21	13	20	28
NV-CG12608-11-7	*CG12608*	NEW	60100	tub	1	11	20	11	17
NV-CG12608-11-7	*CG12608*	NEW	60100	tub	2	21	12	27	25
NV-CG12608-11-7	*CG12608*	NEW	60100	tub	3	15	22	21	30
NV-CG12608-11-7	*CG12608*	NEW	60100	tub	4	13	17	13	19
NV-CG12608-11-7	*CG12608*	NEW	60100	tub	5	25	34	22	22
NV-CG12608-11-7	*CG12608*	NEW	60100	tub	6	13	23	24	25
51878	*CG18278*	TRiP	25709	act	1	34	35	29	10
51878	*CG18278*	TRiP	25709	act	1	34	35	29	25
51878	*CG18278*	TRiP	25709	act	3	45	24	34	24
51878	*CG18278*	TRiP	25709	act	3	45	24	34	6
51878	*CG18278*	TRiP	25709	act	4	20	11	12	21
51878	*CG18278*	TRiP	25709	tub	1	19	21	14	24
51878	*CG18278*	TRiP	25709	tub	1	19	21	14	8
51878	*CG18278*	TRiP	25709	tub	2	22	24	24	31
51878	*CG18278*	TRiP	25709	tub	3	25	21	12	15
51878	*CG18278*	TRiP	25709	tub	4	10	8	23	25
55375	*CG18787*	TRiP	25709	act	1	47	14	16	17
55375	*CG18787*	TRiP	25709	act	2	25	17	25	23
55375	*CG18787*	TRiP	25709	act	3	21	14	17	8
55375	*CG18787*	TRiP	25709	act	4	35	26	22	23
55375	*CG18787*	TRiP	25709	act	5	27	33	33	20
55375	*CG18787*	TRiP	25709	tub	1	13	17	11	12
55375	*CG18787*	TRiP	25709	tub	2	8	13	19	12
55663	*CG18789*	TRiP	25709	act	1	37	0	21	0
55663	*CG18789*	TRiP	25709	act	2	20	0	22	0
55663	*CG18789*	TRiP	25709	act	3	15	0	23	0
55663	*CG18789*	TRiP	25709	act	4	33	0	19	0
55663	*CG18789*	TRiP	25709	tub	1	18	0	14	0
55663	*CG18789*	TRiP	25709	tub	2	24	0	22	0
55663	*CG18789*	TRiP	25709	tub	3	31	0	17	0
55663	*CG18789*	TRiP	25709	tub	4	19	0	21	0
28520	*CG30059*	TRiP	25710	act	1	26	32	22	16
28520	*CG30059*	TRiP	25710	act	2	29	21	23	28
28520	*CG30059*	TRiP	25710	act	3	27	26	17	13
28520	*CG30059*	TRiP	25710	tub	1	27	23	20	13
28520	*CG30059*	TRiP	25710	tub	2	20	20	12	19
28520	*CG30059*	TRiP	25710	tub	3	21	17	23	18
28520	*CG30059*	TRiP	25710	tub	4	21	21	31	25
28607	*CG30059*	TRiP	25710	act	1	29	28	21	23

a: act: *Act5C::GAL4* driver or tub: *αTub84B::GAL4* driver

Table A.3: *Continued from previous page*

line	gene	source	control	driver[a]	rep.	balF	wtF	balM	wtM
28607	*CG30059*	TRiP	25710	act	2	34	32	34	27
28607	*CG30059*	TRiP	25710	act	3	20	24	25	20
28607	*CG30059*	TRiP	25710	tub	1	24	20	16	31
28607	*CG30059*	TRiP	25710	tub	2	25	22	16	15
28607	*CG30059*	TRiP	25710	tub	3	21	26	22	21
28607	*CG30059*	TRiP	25710	tub	4	17	15	13	16
31569	*CG31958*	TRiP	25710	act	1	24	28	29	24
31569	*CG31958*	TRiP	25710	act	2	25	16	21	15
31569	*CG31958*	TRiP	25710	act	3	34	30	28	42
31569	*CG31958*	TRiP	25710	act	4	16	19	12	26
31569	*CG31958*	TRiP	25710	tub	2	20	26	19	20
31569	*CG31958*	TRiP	25710	tub	3	23	19	23	34
31569	*CG31958*	TRiP	25710	tub	4	18	11	21	25
NV-CG31958-1	*CG31958*	NEW	60100	act	1	24	19	18	20
NV-CG31958-1	*CG31958*	NEW	60100	act	2	31	27	24	29
NV-CG31958-1	*CG31958*	NEW	60100	act	3	22	24	17	18
NV-CG31958-1	*CG31958*	NEW	60100	act	4	27	19	25	30
NV-CG31958-1	*CG31958*	NEW	60100	act	5	27	20	24	32
NV-CG31958-1	*CG31958*	NEW	60100	tub	2	28	25	30	29
NV-CG31958-1	*CG31958*	NEW	60100	tub	3	20	14	18	17
NV-CG31958-1	*CG31958*	NEW	60100	tub	4	14	13	11	15
NV-CG31958-1	*CG31958*	NEW	60100	tub	5	18	23	19	13
NV-CG31958-2	*CG31958*	NEW	60100	act	1	18	15	30	17
NV-CG31958-2	*CG31958*	NEW	60100	act	2	21	37	20	22
NV-CG31958-2	*CG31958*	NEW	60100	tub	1	20	16	19	16
NV-CG31958-2	*CG31958*	NEW	60100	tub	2	7	16	14	19
NV-CG31958-2	*CG31958*	NEW	60100	tub	3	19	22	12	22
NV-CG31960-1	*CG31960*	NEW	60100	act	1	25	22	21	20
NV-CG31960-1	*CG31960*	NEW	60100	act	2	31	21	26	33
NV-CG31960-1	*CG31960*	NEW	60100	act	3	24	23	26	27
NV-CG31960-1	*CG31960*	NEW	60100	tub	1	20	21	25	27
NV-CG31960-1	*CG31960*	NEW	60100	tub	2	24	27	16	23
NV-CG31960-1	*CG31960*	NEW	60100	tub	3	20	30	29	27
28692	*CG32164*	TRiP	25710	act	1	32	32	20	30
28692	*CG32164*	TRiP	25710	act	2	27	27	19	29
28692	*CG32164*	TRiP	25710	act	3	21	27	14	17
28692	*CG32164*	TRiP	25710	act	4	12	20	32	27
28692	*CG32164*	TRiP	25710	tub	1	24	29	15	13
28692	*CG32164*	TRiP	25710	tub	2	11	14	14	15
28692	*CG32164*	TRiP	25710	tub	3	16	18	20	29
28692	*CG32164*	TRiP	25710	tub	4	23	31	19	24
NV-CG32164-7	*CG32164*	NEW	60100	act	1	25	18	24	23
NV-CG32164-7	*CG32164*	NEW	60100	act	2	15	16	13	20
NV-CG32164-7	*CG32164*	NEW	60100	tub	1	9	12	14	16
NV-CG32164-7	*CG32164*	NEW	60100	tub	2	17	22	14	21
NV-CG32164_1C	*CG32164*	NEW	60100	act	1	25	18	17	22
NV-CG32164_1C	*CG32164*	NEW	60100	act	2	24	25	23	13
NV-CG32164_1C	*CG32164*	NEW	60100	tub	1	13	13	17	11
NV-CG32164_1C	*CG32164*	NEW	60100	tub	3	20	21	16	16
NV-CG32164_1C	*CG32164*	NEW	60100	tub	4	12	11	12	11
NV-CG32164_1R3	*CG32164*	NEW	60100	act	2	28	24	16	16
NV-CG32164_1R3	*CG32164*	NEW	60100	tub	1	28	17	22	13
NV-CG32164_1R3	*CG32164*	NEW	60100	tub	2	28	24	16	27
GD49306	*CG32165*	VDRC	60100	act	1	17	28	11	21
GD49306	*CG32165*	VDRC	60100	act	2	29	27	26	14
GD49306	*CG32165*	VDRC	60100	act	3	20	15	21	14
GD49306	*CG32165*	VDRC	60100	act	4	31	27	25	22
GD49306	*CG32165*	VDRC	60100	tub	1	11	8	11	10
GD49306	*CG32165*	VDRC	60100	tub	2	14	27	20	30

a: act: *Act5C::GAL4* driver or tub: *αTub84B::GAL4* driver

Table A.3: *Continued from previous page*

line	gene	source	control	driver[a]	rep.	balF	wtF	balM	wtM
GD49306	*CG32165*	VDRC	60100	tub	3	12	18	15	13
GD49306	*CG32165*	VDRC	60100	tub	4	24	27	23	24
GD49307	*CG32165*	VDRC	60100	act	1	22	14	19	27
GD49307	*CG32165*	VDRC	60100	act	2	25	22	15	21
GD49307	*CG32165*	VDRC	60100	tub	1	18	19	17	28
GD49307	*CG32165*	VDRC	60100	tub	2	32	27	21	22
NV-CG32165	*CG32165*	NEW	60100	act	2	19	20	19	27
NV-CG32165	*CG32165*	NEW	60100	act	3	24	24	19	21
NV-CG32165	*CG32165*	NEW	60100	tub	2	22	19	13	22
NV-CG32165	*CG32165*	NEW	60100	tub	3	15	22	11	27
NV-CG32165-2	*CG32165*	NEW	60100	act	1	10	13	16	18
NV-CG32165-2	*CG32165*	NEW	60100	tub	1	22	7	23	17
NV-CG32165-2	*CG32165*	NEW	60100	tub	2	15	15	8	18
NV-CG32165-2-4	*CG32165*	NEW	60100	act	1	25	10	24	10
NV-CG32165-2-4	*CG32165*	NEW	60100	act	2	24	13	23	14
NV-CG32165-2-4	*CG32165*	NEW	60100	act	3	25	8	26	6
NV-CG32165-2-4	*CG32165*	NEW	60100	act	4	15	13	14	12
NV-CG32165-2-4	*CG32165*	NEW	60100	tub	1	20	14	22	18
NV-CG32165-2-4	*CG32165*	NEW	60100	tub	2	26	13	20	16
NV-CG32165-2-4	*CG32165*	NEW	60100	tub	3	16	6	26	14
NV-CG32165-2-6	*CG32165*	NEW	60100	act	1	26	11	17	5
NV-CG32165-2-6	*CG32165*	NEW	60100	act	2	17	17	17	10
NV-CG32165-2-6	*CG32165*	NEW	60100	act	3	18	15	23	12
NV-CG32165-2-6	*CG32165*	NEW	60100	act	4	14	20	23	13
NV-CG32165-2-6	*CG32165*	NEW	60100	tub	1	16	11	14	13
NV-CG32165-2-6	*CG32165*	NEW	60100	tub	2	18	11	14	22
NV-CG32165-2-6	*CG32165*	NEW	60100	tub	3	20	18	24	16
NV-CG32165-2-6	*CG32165*	NEW	60100	tub	4	32	16	36	24
50935	*CG32588*	TRiP	25709	act	1	29	42	29	19
50935	*CG32588*	TRiP	25709	act	2	23	30	21	15
50935	*CG32588*	TRiP	25709	tub	2	15	31	13	18
50935	*CG32588*	TRiP	25709	tub	3	20	22	18	22
KK102410-3	*CG32588*	VDRC	60100	act	1	12	20	14	16
KK102410-3	*CG32588*	VDRC	60100	act	1	22	10	7	8
KK102410-3	*CG32588*	VDRC	60100	act	2	18	12	16	16
KK102410-3	*CG32588*	VDRC	60100	act	3	17	16	9	16
KK102410-3	*CG32588*	VDRC	60100	act	4	16	14	16	18
KK102410-3	*CG32588*	VDRC	60100	tub	1	22	12	18	20
KK102410-3	*CG32588*	VDRC	60100	tub	2	10	10	17	11
KK102410-3	*CG32588*	VDRC	60100	tub	3	17	21	16	21
NV-CG33252-4	*CG33252*	NEW	60100	act	1	24	24	21	22
NV-CG33252-4	*CG33252*	NEW	60100	act	1	18	23	23	26
NV-CG33252-4	*CG33252*	NEW	60100	act	2	23	32	25	27
NV-CG33252-4	*CG33252*	NEW	60100	act	2	13	22	22	22
NV-CG33252-4	*CG33252*	NEW	60100	act	3	28	27	31	22
NV-CG33252-4	*CG33252*	NEW	60100	tub	1	15	23	12	21
NV-CG33252-4	*CG33252*	NEW	60100	tub	1	16	24	17	21
NV-CG33252-5	*CG33252*	NEW	60100	act	1	20	21	26	24
NV-CG33252-5	*CG33252*	NEW	60100	act	2	11	15	14	12
NV-CG33252-5	*CG33252*	NEW	60100	act	3	30	23	20	22
NV-CG33252-5	*CG33252*	NEW	60100	act	4	27	19	19	24
NV-CG33252-5	*CG33252*	NEW	60100	tub	1	20	21	18	18
NV-CG33252-5	*CG33252*	NEW	60100	tub	2	19	23	17	18
NV-CG33252-5	*CG33252*	NEW	60100	tub	3	34	40	35	44
NV-CG33252-5	*CG33252*	NEW	60100	tub	4	14	14	14	20
NV-CG33252-5	*CG33252*	NEW	60100	tub	5	18	23	16	16
28540	*CG33470*	TRiP	25710	act	1	30	34	31	20
28540	*CG33470*	TRiP	25710	act	2	28	25	18	17
28540	*CG33470*	TRiP	25710	act	3	22	30	16	19

a: act: *Act5C::GAL4* driver or tub: *αTub84B::GAL4* driver

Table A.3: *Continued from previous page*

line	gene	source	control	driver[a]	rep.	balF	wtF	balM	wtM
28540	*CG33470*	TRiP	25710	tub	1	27	21	18	20
28540	*CG33470*	TRiP	25710	tub	2	19	21	22	26
28540	*CG33470*	TRiP	25710	tub	3	22	14	27	24
28540	*CG33470*	TRiP	25710	tub	4	20	16	13	19
43147	*CG33470*	TRiP	25710	act	1	12	13	7	15
43147	*CG33470*	TRiP	25710	act	5	22	16	13	14
43147	*CG33470*	TRiP	25710	tub	1	16	10	12	10
43147	*CG33470*	TRiP	25710	tub	2	16	16	7	11
43147	*CG33470*	TRiP	25710	tub	3	10	12	10	16
43147	*CG33470*	TRiP	25710	tub	4	9	17	8	5
55264	*CG6300*	TRiP	25709	act	1	34	30	35	9
55264	*CG6300*	TRiP	25709	act	2	15	14	14	14
55264	*CG6300*	TRiP	25709	act	3	26	17	9	2
55264	*CG6300*	TRiP	25709	act	4	28	21	14	23
55264	*CG6300*	TRiP	25709	tub	1	19	24	17	12
55264	*CG6300*	TRiP	25709	tub	2	19	10	14	9
55264	*CG6300*	TRiP	25709	tub	3	16	18	8	12
55264	*CG6300*	TRiP	25709	tub	4	34	20	8	18
55264	*CG6300*	TRiP	25709	tub	5	8	5	13	11
NV-CG6300-1	*CG6300*	NEW	60100	act	1	30	27	21	26
NV-CG6300-1	*CG6300*	NEW	60100	act	2	22	25	22	19
NV-CG6300-1	*CG6300*	NEW	60100	act	3	28	19	22	30
NV-CG6300-1	*CG6300*	NEW	60100	tub	1	25	26	28	22
NV-CG6300-1	*CG6300*	NEW	60100	tub	2	24	24	19	22
NV-CG6300-1	*CG6300*	NEW	60100	tub	3	22	19	28	27
61894	*CG9123*	TRiP	25709	act	1	10	0	14	0
61894	*CG9123*	TRiP	25709	tub	1	16	0	16	0
61894	*CG9123*	TRiP	25709	tub	2	12	0	17	0
61894	*CG9123*	TRiP	25709	tub	3	10	0	16	0
61894	*CG9123*	TRiP	25709	tub	4	18	0	16	0
61894	*CG9123*	TRiP	25709	tub	5	15	0	18	0
62982	*CG9123*	TRiP	25709	tub	1	12	0	14	0
62982	*CG9123*	TRiP	25709	tub	2	10	0	13	0
62982	*CG9123*	TRiP	25709	tub	3	10	0	11	0
62982	*CG9123*	TRiP	25709	tub	4	11	0	10	0
GD24629	*CG9123*	VDRC	60100	act	1	16	0	13	0
GD24629	*CG9123*	VDRC	60100	act	1	16	0	13	0
GD24629	*CG9123*	VDRC	60100	act	2	20	0	23	1
GD24629	*CG9123*	VDRC	60100	tub	1	20	0	28	0
GD24629	*CG9123*	VDRC	60100	tub	2	40	0	39	0
NV-CG9123-8	*CG9123*	NEW	60100	act	1	29	0	24	0
NV-CG9123-8	*CG9123*	NEW	60100	act	2	26	0	26	0
NV-CG9123-8	*CG9123*	NEW	60100	act	3	28	0	32	0
NV-CG9123-8	*CG9123*	NEW	60100	act	4	32	0	31	0
NV-CG9123-8	*CG9123*	NEW	60100	tub	1	30	0	31	0
NV-CG9123-8	*CG9123*	NEW	60100	tub	2	27	0	29	0
NV-CG9123-8	*CG9123*	NEW	60100	tub	3	29	0	24	0
NV-CG9123-8	*CG9123*	NEW	60100	tub	4	28	0	22	0
NV-CG4479-6-2	*ProtA*	NEW	60100	act	1	24	23	24	27
NV-CG4479-6-2	*ProtA*	NEW	60100	act	3	24	24	18	31
NV-CG4479-6-2	*ProtA*	NEW	60100	act	4	13	12	11	13
NV-CG4479-6-2	*ProtA*	NEW	60100	tub	1	19	21	28	25
NV-CG4479-6-2	*ProtA*	NEW	60100	tub	3	29	14	23	16
NV-CG4479-6-3	*ProtA*	NEW	60100	act	1	27	22	23	21
NV-CG4479-6-3	*ProtA*	NEW	60100	tub	1	23	26	33	18
NV-CG4479-6-3	*ProtA*	NEW	60100	tub	2	16	18	17	14
NV-CG4479-6-5	*ProtA*	NEW	60100	act	1	25	22	17	20
NV-CG4479-6-5	*ProtA*	NEW	60100	act	2	19	19	19	21
NV-CG4479-6-5	*ProtA*	NEW	60100	act	3	21	22	17	22

a: act: *Act5C::GAL4* driver or tub: *αTub84B::GAL4* driver

Table A.3: *Continued from previous page*

line	gene	source	control	driver[a]	rep.	balF	wtF	balM	wtM
NV-CG4479-6-5	*ProtA*	NEW	60100	tub	1	28	21	16	35
NV-CG4479-6-5	*ProtA*	NEW	60100	tub	2	13	14	14	19
KK102917	*ProtB*	VDRC	60100	act	1	18	25	20	17
KK102917	*ProtB*	VDRC	60100	act	2	11	14	12	15
KK102917	*ProtB*	VDRC	60100	act	2	11	14	12	15
KK102917	*ProtB*	VDRC	60100	act	3	28	22	28	36
KK102917	*ProtB*	VDRC	60100	act	3	28	22	9	36
KK102917	*ProtB*	VDRC	60100	tub	1	26	17	15	20
KK102917	*ProtB*	VDRC	60100	tub	3	26	15	12	19
KK102918	*ProtB*	VDRC	60101	tub	4	17	21	25	15
NV-CG4478	*ProtB*	NEW	60100	act	1	17	19	18	13
NV-CG4478	*ProtB*	NEW	60100	tub	1	15	22	19	22
NV-CG4478	*ProtB*	NEW	60100	tub	2	20	22	18	13
NV-CG4478-15-1	*ProtB*	NEW	60100	act	1	19	20	10	19
NV-CG4478-15-1	*ProtB*	NEW	60100	act	2	18	15	14	19
NV-CG4478-15-1	*ProtB*	NEW	60100	act	3	20	15	22	27
NV-CG4478-15-1	*ProtB*	NEW	60100	act	4	15	20	23	19
NV-CG4478-15-1	*ProtB*	NEW	60100	tub	1	19	13	12	12
NV-CG4478-4-1	*ProtB*	NEW	60100	act	1	14	16	14	10
NV-CG4478-4-1	*ProtB*	NEW	60100	tub	1	11	21	14	15
NV-CG4478-4-1	*ProtB*	NEW	60100	tub	2	15	23	8	18
NV-CG4478-4-2	*ProtB*	NEW	60100	act	1	18	29	29	25
NV-CG4478-4-2	*ProtB*	NEW	60100	act	2	15	11	18	11
NV-CG4478-4-2	*ProtB*	NEW	60100	act	2	15	11	9	11
NV-CG4478-4-2	*ProtB*	NEW	60100	act	3	25	28	19	26
NV-CG4478-4-2	*ProtB*	NEW	60100	act	4	18	22	18	23
NV-CG4478-4-2	*ProtB*	NEW	60100	tub	1	15	19	21	16
NV-CG4478-4-2	*ProtB*	NEW	60100	tub	2	24	23	17	20
34005	*RpS15Aa*	TRiP	25710	tub	1	20	0	13	0
34005	*RpS15Aa*	TRiP	25710	tub	2	14	1	11	0
34005	*RpS15Aa*	TRiP	25710	tub	3	11	0	15	0
34005	*RpS15Aa*	TRiP	25710	tub	5	10	0	16	0
34005	*RpS15Aa*	TRiP	25710	act	1	22	0	12	2
34005	*RpS15Aa*	TRiP	25710	act	2	17	0	11	0
34005	*RpS15Aa*	TRiP	25710	act	3	15	0	18	0
34008	*RpS15Ab*	TRiP	25710	act	1	15	18	5	14
34008	*RpS15Ab*	TRiP	25710	act	3	26	17	10	8
34008	*RpS15Ab*	TRiP	25710	act	4	17	16	12	13
34008	*RpS15Ab*	TRiP	25710	act	6	15	13	7	16
34008	*RpS15Ab*	TRiP	25710	tub	1	11	11	9	13
34008	*RpS15Ab*	TRiP	25710	tub	4	10	12	14	9
34008	*RpS15Ab*	TRiP	25710	tub	5	12	15	27	23
34008	*RpS15Ab*	TRiP	25710	tub	6	9	12	11	12
GD49103	*RpS15Ab*	VDRC	60100	act	1	32	0	31	0
GD49103	*RpS15Ab*	VDRC	60100	act	2	24	0	19	1
GD49103	*RpS15Ab*	VDRC	60100	tub	1	31	0	23	0
GD49103	*RpS15Ab*	VDRC	60100	tub	2	12	0	12	0
KK109062	*RpS15Ab*	VDRC	60100	act	1	43	0	34	0
KK109062	*RpS15Ab*	VDRC	60100	act	2	31	0	30	0
KK109062	*RpS15Ab*	VDRC	60100	tub	1	21	0	32	0
KK109062	*RpS15Ab*	VDRC	60100	tub	2	26	0	42	0
32969	*sle*	TRiP	25710	tub	1	11	8	23	22
32969	*sle*	TRiP	25710	tub	2	2	8	9	8
28656	*tHMG1*	TRiP	25710	act	1	24	0	16	0
28656	*tHMG1*	TRiP	25710	act	2	21	0	5	0
28656	*tHMG1*	TRiP	25710	act	3	30	0	26	0
28656	*tHMG1*	TRiP	25710	act	4	25	0	34	0
28656	*tHMG1*	TRiP	25710	tub	1	31	0	19	0
28656	*tHMG1*	TRiP	25710	tub	2	6	0	15	0

a: act: *Act5C::GAL4* driver or tub: *αTub84B::GAL4* driver

Table A.3: *Continued from previous page*

line	gene	source	control	driver[a]	rep.	balF	wtF	balM	wtM
28656	*tHMG1*	TRiP	25710	tub	3	31	0	18	0
28656	*tHMG1*	TRiP	25710	tub	4	25	0	20	1
61962	*tHMG1*	TRiP	25709	act	1	38	0	34	0
61962	*tHMG1*	TRiP	25709	act	2	39	0	31	0
61962	*tHMG1*	TRiP	25709	act	3	32	0	30	0
61962	*tHMG1*	TRiP	25709	tub	1	16	0	14	0
61962	*tHMG1*	TRiP	25709	tub	2	19	0	32	0
61962	*tHMG1*	TRiP	25709	tub	3	12	0	14	0
61962	*tHMG1*	TRiP	25709	tub	4	10	0	10	0
61962	*tHMG1*	TRiP	25709	tub	5	16	0	10	0
NV-CG7045-16	*tHMG1*	NEW	60100	tub	5	18	0	13	0
NV-CG7045-16	*tHMG1*	NEW	60100	tub	7	32	0	31	0
NV-CG7045-17	*tHMG1*	NEW	60100	act	1	18	20	13	6
NV-CG7045-17	*tHMG1*	NEW	60100	act	2	19	24	16	11
NV-CG7045-17	*tHMG1*	NEW	60100	act	3	16	31	30	23
NV-CG7045-17	*tHMG1*	NEW	60100	act	4	11	13	15	6
NV-CG7045-17	*tHMG1*	NEW	60100	act	5	31	22	21	8
NV-CG7045-17	*tHMG1*	NEW	60100	tub	1	24	0	25	0
NV-CG7045-17	*tHMG1*	NEW	60100	tub	2	37	0	21	0
NV-CG7045-17	*tHMG1*	NEW	60100	tub	4	17	0	13	0
NV-CG7045-17	*tHMG1*	NEW	60100	tub	5	26	0	15	0
NV-CG7045-5	*tHMG1*	NEW	60100	act	1	12	13	18	7
NV-CG7045-5	*tHMG1*	NEW	60100	act	2	31	20	21	22
NV-CG7045-5	*tHMG1*	NEW	60100	act	3	11	16	14	34
NV-CG7045-5	*tHMG1*	NEW	60100	act	4	38	19	16	7
NV-CG7045-5	*tHMG1*	NEW	60100	act	5	16	15	10	6
NV-CG7045-5	*tHMG1*	NEW	60100	act	6	21	28	24	20
NV-CG7045-5	*tHMG1*	NEW	60100	act	7	18	20	31	8
NV-CG7045-5	*tHMG1*	NEW	60100	tub	1	21	0	12	0
NV-CG7045-5	*tHMG1*	NEW	60100	tub	2	29	0	26	0
NV-CG7045-5	*tHMG1*	NEW	60100	tub	4	15	0	17	2
NV-CG7045-5	*tHMG1*	NEW	60100	tub	5	18	1	12	0
NV-CG7045-5	*tHMG1*	NEW	60100	tub	6	28	0	27	0
NV-CG7045-5	*tHMG1*	NEW	60100	tub	8	19	0	28	0
NV-CG7045-7	*tHMG1*	NEW	60100	act	1	20	6	16	6
NV-CG7045-7	*tHMG1*	NEW	60100	act	2	16	2	22	5
NV-CG7045-7	*tHMG1*	NEW	60100	act	3	26	3	18	7
NV-CG7045-7	*tHMG1*	NEW	60100	act	4	22	2	14	7
NV-CG7045-7	*tHMG1*	NEW	60100	tub	1	14	0	10	0
NV-CG7045-7	*tHMG1*	NEW	60100	tub	2	14	0	20	0
NV-CG7045-7	*tHMG1*	NEW	60100	tub	3	22	1	8	0
NV-CG7045-7	*tHMG1*	NEW	60100	tub	4	18	0	16	0
NV-CG7045-7	*tHMG1*	NEW	60100	tub	5	11	0	11	0
28657	*tHMG2*	TRiP	25710	act	1	32	25	21	19
28657	*tHMG2*	TRiP	25710	act	2	15	21	9	20
28657	*tHMG2*	TRiP	25710	act	3	45	10	24	24
28657	*tHMG2*	TRiP	25710	act	4	10	15	10	12
28657	*tHMG2*	TRiP	25710	tub	1	34	16	16	19
28657	*tHMG2*	TRiP	25710	tub	2	18	14	17	12
28657	*tHMG2*	TRiP	25710	tub	3	25	16	13	14
28657	*tHMG2*	TRiP	25710	tub	4	20	14	13	18
NV-CG7046-7	*tHMG2*	NEW	60100	act	1	25	26	21	23
NV-CG7046-7	*tHMG2*	NEW	60100	act	2	12	19	30	16
NV-CG7046-7	*tHMG2*	NEW	60100	act	3	19	26	28	24
NV-CG7046-7	*tHMG2*	NEW	60100	act	4	21	14	17	22
NV-CG7046-7	*tHMG2*	NEW	60100	act	5	19	13	17	21
NV-CG7046-7	*tHMG2*	NEW	60100	tub	1	16	24	11	16
NV-CG7046-7	*tHMG2*	NEW	60100	tub	2	15	18	13	14
NV-CG7046-7	*tHMG2*	NEW	60100	tub	3	15	17	15	19

a: act: *Act5C::GAL4* driver or tub: *αTub84B::GAL4* driver

Table A.3: *Continued from previous page*

line	gene	source	control	driver[a]	rep.	balF	wtF	balM	wtM
NV-CG7046-7	*tHMG2*	NEW	60100	tub	4	15	18	13	19
NV-CG7046-7	*tHMG2*	NEW	60100	tub	5	8	18	16	15
38967	*Ubi-p5E*	TRiP	25709	act	3	16	15	10	16
28967	*Ubi-p5E*	TRiP	25709	act	1	24	20	9	14
38967	*Ubi-p5E*	TRiP	25709	tub	1	18	14	14	11
38967	*Ubi-p5E*	TRiP	25709	tub	3	14	12	13	18
NV-CG32744-6	*Ubi-p5E*	NEW	60100	act	1	12	23	24	18
NV-CG32744-6	*Ubi-p5E*	NEW	60100	act	2	24	29	16	23
NV-CG32744-6	*Ubi-p5E*	NEW	60100	act	3	19	23	13	21
NV-CG32744-6	*Ubi-p5E*	NEW	60100	tub	1	10	22	25	10
NV-CG32744-6	*Ubi-p5E*	NEW	60100	tub	2	22	21	22	27
NV-CG32744-6	*Ubi-p5E*	NEW	60100	tub	3	18	11	24	27

a: act: *Act5C::GAL4* driver or tub: *αTub84B::GAL4* driver

Table A.4: Welch's t statistics comparing the proportion of balancer F1 flies from RNAi crosses and control crosses for males and females separately or pooled. Benjamini-Hochberg corrected $*p < 0.05$, $**p < 0.01$, $***p < 0.001$

			Act5C::GAL4			***Tub84B::GAL4***		
line	**target**	$\mathbf{s}_{19}^{a}$	**males**	**females**	**pooled**	**males**	**females**	**pooled**
31735	*Cdc23*	1	1.74	-0.98	1.08	0.19	0.4	2.13
55268^b	*CG11659*	1	1.95	0.73	1.77	-1.83	1.95	1.41
NV-CG11659-2^b	*CG11659*	1	1.65	*-3.95	-0.46	-1.52	-0.53	-1.56
32933^b	*CG11700*	1	1.1	0.3	1.04	1.16	-0.71	-0.02
NV-CG11700-1^b	*CG11700*	1	1.21	-1.56	-0.07	-0.52	0.18	-0.7
44037	*CG12264*	1	NA	NA	NA	NA	NA	NA
36079^b	*CG12592*	1	-0.25	-0.58	-0.6	-0.62	-2.67	-1.78
NV-CG12608-11-4^b	*CG12608*	1	1.26	-3.26	-0.81	-0.55	0.79	0.02
NV-CG12608-11-7^b	*CG12608*	1	3.07	-2.24	0.44	-0.67	1.14	0.24
51878^b	*CG18278*	1	-1.51	-1.37	-1.75	0.26	0.45	0.7
55375^b	*CG18787*	1	-1.54	-1.3	-2.09	-0.83	2.16	0.43
55663^b	*CG18789*	0	**-14.48	**-17.75	**-16.51	***-22.08	**-9.34	***-26.48
28520^b	*CG30059*	0	0.63	0.39	1.28	-0.89	*-4.94	-2.28
28607^b	*CG30059*	0	0.73	1.11	1.33	0.35	-3.04	-1.17
31569^b	*CG31958*	0.34	1.35	0.36	1.23	0.56	-1.75	-1.01
NV-CG31958-1	*CG31958*	1	2.51	**-5.17	-1.62	-1.54	-0.78	-1.58
NV-CG31958-2^b	*CG31958*	1	-0.27	-0.38	-0.3	-0.31	0.64	0.12
NV-CG31960-1^b	*CG31960*	1	1.5	-4.25	-2.34	-0.9	1.08	-0.29
28692^b	*CG32164*	0.63	1.93	1.88	3.28	0.14	-0.66	0.11
NV-CG32164-7^b	*CG32164*	1	1.26	-2.23	-0.28	-0.27	2.7	0.49
NV-CG32164_1C	*CG32164*	1	-0.03	-2.43	-2.3	-2.14	-0.58	-1.85
NV-CG32164_1R3	*CG32164*	1	-0.6	-1.94	-1.04	-0.6	-1.94	-1.04
GD49306^b	*CG32165*	0.10	0.04	-1.82	-0.72	-1.21	0.76	-0.28
GD49307	*CG32165*	0.10	4.2	-3.6	-0.91	-0.21	-0.71	-0.35
NV-CG32165-2	*CG32165*	1	-1.41	-4.55	***-10.7	-2.89	-1.8	*-3.46
NV-CG32165-2-4^b	*CG32165*	1	-2.08	*-4.84	-3.49	-3.15	-3.77	*-3.95
NV-CG32165-2-6	*CG32165*	1	*-3.68	**-6.51	**-6.36	-2.83	*-3.8	*-3.81
50935^b	*CG32588*	1	-3.09	3.41	0.28	1.85	1.42	1.67
KK102410-3^b	*CG32588*	1	2.19	-2.3	-0.81	-1.21	-0.66	-1.28
NV-CG33252-4^b	*CG33252*	1	0.84	-0.97	-0.08	0.18	*4.37	1.3
NV-CG33252-5^b	*CG33252*	1	1.12	-2.4	-1.84	-0.98	0.92	-0.37
28540^b	*CG33470*	0	0.69	1.38	1.36	0.15	-3.75	-1.99
43147^b	*CG33470*	0	1.77	0.03	1.11	-0.15	-0.72	-0.46
55264^b	*CG6300*	1	-1.27	-1	-1.38	-0.04	-0.49	-0.86
NV-CG6300-1^b	*CG6300*	1	1.46	-2.79	-1.55	-1.66	-0.76	-1.57
61894	*CG9123*	1	***-22.08	**-9.34	***-26.48	***-22.08	**-9.34	***-26.48
62982^b	*CG9123*	0	***-22.08	**-9.34	***-26.48	***-22.08	**-9.34	***-26.48
GD24629	*CG9123*	0.66	***-14.32	***-36.83	***-32.36	**-11.95	***-24.34	***-16.39
NV-CG9123-8^b	*CG9123*	1	***-16.44	***-36.83	***-37.29	**-11.95	***-24.34	***-16.39
NV-CG4479-6-2^b	*ProtA*	0.98	2.37	**-5.48	0.21	-2.46	-0.83	-1.61
NV-CG4479-6-3	*ProtA*	0.98	-2.6	0.94	-1.9	-2.6	0.94	-1.9
NV-CG4479-6-5^b	*ProtA*	0.98	2.54	*-4.33	-0.44	0.66	-0.74	0.24
KK102917^b	*ProtB*	0.94	1.91	-1.87	1.23	-0.71	-1.17	-2.31
NV-CG4478	*ProtB*	0.99	-1.37	1.19	-0.47	-1.37	1.19	-0.47
NV-CG4478-15-1	*ProtB*	0.99	1.84	-2.53	-0.02	NA	NA	NA
NV-CG4478-4-1	*ProtB*	0.99	0.22	3.65	1.64	0.22	3.65	1.64
NV-CG4478-4-2^b	*ProtB*	0.99	0.86	-1.8	-0.63	-1.32	0.34	-1.28
34005^b	*RpS15Aa*	1	*-5.04	**-16.08	***-13.18	**-14.62	***-27.61	***-27.69
34008^b	*RpS15Ab*	1	1.94	0.09	1.43	-0.64	-1.02	-1.24
GD49103	*RpS15Ab*	0.6	**-11.66	***-36.83	***-28.37	**-11.95	***-24.34	***-16.39
KK109062^b	*RpS15Ab*	0.7	***-16.44	***-36.83	***-37.29	**-11.95	***-24.34	***-16.39
32969	*sle*	1	-1.26	0.26	-0.22	-1.26	0.26	-0.22
28656^b	*tHMG1*	0.84	*-7.39	**-16.08	**-16.7	**-13.57	***-52.98	***-30.6
61962	*tHMG1*	0	**-14.48	**-17.75	**-16.51	***-22.08	**-9.34	***-26.48
NV-CG7045-16	*tHMG1*	1	**-11.95	***-24.34	***-16.39	**-11.95	***-24.34	***-16.39
NV-CG7045-17	*tHMG1*	1	-2.84	-0.99	-2.02	**-11.95	***-24.34	***-16.39
NV-CG7045-5^b	*tHMG1*	1	-0.86	-2.39	-1.6	***-10.91	***-22.07	***-15.24
NV-CG7045-7	*tHMG1*	1	*-4.74	**-12.18	***-13.48	**-11.95	***-22.1	***-15.9
28657^b	*tHMG2*	0.87	1.72	-0.17	0.67	-0.25	*-6.29	*-4.2
NV-CG7046-7^b	*tHMG2*	1	0.49	-1.74	-1.81	-0.62	1.99	0.51
38967^b	*Ubi-p5E*	1	-0.05	-0.4	-0.32	-0.05	-0.4	-0.32
NV-CG32744-6^b	*Ubi-p5E*	1	1.22	0.05	1.76	-1.3	0.09	-1.51

a: fraction of all possible 19-mers (siRNAs) specifically matching target gene; off-targets are always only the other duplicate
b: plotted in Figure 2.2

Table A.5: Compiled phenotype data for *D. melanogaster*-specific genes and their parents.

Gene	**Phenotype**	**Source**	**Df, TEI confirmed?**	**Reference**
CG31958	Viable, fertile	NV-CG31958, TRiP 31569	-	-
CG31960	Viable, fertile	NV-CG31960	-	-
Ada1-1	Unknown	-	-	-
Ada1-2	Unknown	-	-	-
CG18789	Unclear	PEPCG18789^{G2531} P-element insertion in exon 1 of either copy, non-specific RNAi with TRiP 55663	no, seq. too similar	[235]
CG18787	Unclear	P{EP}CG18789^{G2531} P-element insertion in exon 1 of either copy, non-specific RNAi with TRiP 55663	-	-
Qtzl	Probably sterile	P{GSV}Qtzl^{GS8052} P-element insertion in 5'UTR	see [81]	[81, 236]
escl	Viable, fertile	P{XP}escl^{d01514} in exon 1	TE present, no expression	-
CG12264	Lethal	TRiP 44037		
ProtB	Viable, fertile	Df(2L)Δ*Mst35B* Deletion of *ProtA* and *ProtB*, NV-CG4478, VDRC 102917	see [237]	[237]
ProtA	Viable, fertile	Df(2L)Δ*Mst35B* Deletion of *ProtA* and *ProtB*, NV-CG4479	see [237]	[237]
CG31683	Unknown	-	-	-
CG18858	Unknown	-	-	-
CG31687	Unknown	-	-	[238]
Cdc23	Lethal	VDRC stock 52279 and 52280	-	[239, 240]
CG31688	Viable, fertile	VDRC stocks 103257 and 102282, P{SUPor-P}CG31688^{KG07854}	not determined	
RpS15Ab	Viable, fertile	TRiP 34008	-	-
RpS15Aa	Lethal	TRiP 34005, non-specific RNAi with VDRC 49103, 109062	-	-
CG33470	Viable, fertile	Non-specific RNAi with TRIP 28540, 43147	-	-
IMPPP	Viable, fertile	Non-specific RNAi with TRIP 28540, 43147	-	-
CG30059	Viable, fertile	Non-specific RNAi with TRIP 28520, 28627	-	-
CG18278	Viable, fertile	Non-specific RNAi with TRIP 28520, 28627	-	-
CG32165	Lethal, fertile	P{EPgy2}EY14634, NV-CG32165, non-specific RNAi with TRiP 60487	TE confirmed, but expressed	-
CG32164	Viable, fertile	NV-CG32164, non-specific RNAi with TRiP 60487	-	-
CG12592	Viable, fertile	TRiP 36079	-	-
CG18545	Unknown	-	-	
sle	Viable,sterile	Several alleles, including P{GSV1}sle^{GS3144} P-element insertion, TRiP 32969	no, many alleles	[241]
CG11659	Viable, fertile	NV-CG11659, TRiP 55268	-	-
CG6300	Viable, fertile	NV-CG6300, TRiP 55264	-	-
tHMG1	Lethal (Viable, fertile)	NV-CG7045, non-specific RNAi with TRiP 28656, TRiP 61962 (ΔtHMG1/tHMG2 mutant)	see [122] and discussion	[122]
tHMG2	Viable, fertile	NV-CG7046, non-specific RNAi with TRiP 28657, ΔtHMG1/tHMG2 mutant	see [122] and discussion	[122]
CG11700	Viable, lower fertility	CG11700^{null}	see [82]	[82]
Ubi-p5E	Viable, fertile	TRiP 38967, NV-CG32744, Ubi-p5E^{null},	see [82]	[82]
CG32588	Viable, fertile	VDRC 102410, NV-CG32588	-	-
CG33252	Viable, fertile	NV-CG33252	-	-
CG9123	Lethal	NV-CG9123, TRiP 61894, non-specific RNAi with TRiP 24629 and 62982	-	-
CG12608	Viable, fertile	NV-CG12608	-	-

Table A.6: Primers used for reverse-transcriptase quantitative PCR.

Target	**F Primer**	**R Primer**	**Length**	**eff.**[b]
CG31958	AGACGACGAGCTGGAAGAGA	TGTAATTTAGAATGGGCGTGGT	107	104%
CG31960	CTGGAGGAGATGATACGCGAG	GAAAGCGTGTGCGTGTAGATT	106	99%
CG32165	AGCCGACGATCTGCAAGTTT	TTTCGAATCCGGCTCACCAA	102	100%
CG32164	TGCGTCCTGTTGCCTTTCATA	ATCGTCTTCCTTGAGGTCGTC	372	99%
CG9123	GCTGGATTTCCAAACGGGCT	CGCATAACATGGGCGTTCTC	111	101%
CG12608	CATAGTGGGCACCTACGAG	GGAGCTGTCTGCAAAAGTCTG	115	96%
ProtA	GCCAGGCTCTCGGAGAATAG	GTATTGCTGGCAAATCCGTCG	99	98%
ProtB	AGAGCCTGTGGAATGGCATAAT	GGGCGGTGCTCTCCTCTTT	98	101%
tHMG2	AGACGGAGTCGTTTCCCCATA	GGGTTGGCCTTTGGTTAGTTT	75	101%
tHMG1	GACAAGATCGTGTGGCAGGA	ACACAAAAAGGGTGGGGCAT	151	97%
CG33252[a]	CGCACCAACCTAAATATACCACT	TGAGAACGCTCGAAGGATACC	123	94%
CG32588[a]	TTACTGGTGAGAGCGTACATGC	GCAAAAAGCGGAACGAAGATATT	86	97%
RpS15Ab	CCACGAGGAGGCTAGGAGAA	ACATATCAAACTCCATCCCTCTAC	100	102%
RpS15Aa	GCGGTACAGTGATAAATCAATAGCG	TAAGTAACTCCGGTCGAGGT	92	94%
CG12592	GACGGGGAAGTCTCGAATGG	GGTGGCGCTGAATTACCTTC	107	103%
sle[a]	GTCGCTTGTCCCTTCTGGAAA	TCCTCACATCTAAAGTGGACGAG	123	98%
CG11659	TGATGACGGCGAATCGCTTG	CCCGACCAGTGCTGGTTATT	72	96%
CG6300	AATGATGACGGCGAATCCCTG	TTCCGTAGTATCCCGACCAGT	85	102%
RpL32	AGCATACAGGCCCAAGATCG	TGTTGTCGATACCCTTGGGC	112	96%

a: From FlyPrimerBank [110].

b: Efficiencies were calculated using a 8-$\log_2$ dilution series of a common control cDNA template. Efficiency = $10^{-1/s} - 1) \times 100$ and slope is the slope of the line of best fit for a plot of $\log_2$ (template concentration) versus C_T. All correlations were >0.99.

Table A.7: McDonald-Kreitman test results for *D. melanogaster*-specific duplicates.

Gene	RP[a]	RF	SP	SF	Codons[b]	Frac. Codons[c]	Sam. Size[d]	FET *p*
CG31958	4	3	0	9	48	0.28	16.71	0.020
CG31960	0	3	0	11	59	0.4	16.81	1.000
Ada1-2	0	3	0	16	239	0.78	16.97	1.000
Ada1-1	0	4	0	16	239	0.78	17	1.000
CG18789	0	10	0	10	240	0.59	16.99	1.000
CG18787	0	9	0	10	222	0.55	16.96	1.000
ProtB	2	3	1	7	58	0.27	17	0.510
ProtA	2	2	2	4	56	0.26	16.98	1.000
CG31683	0	5	0	11	321	0.76	16.96	1.000
CG18858	0	3	1	11	321	0.76	17	1.000
RpS15Ab	0	3	3	7	118	0.91	16.99	0.530
RpS15Aa	0	0	0	9	118	0.91	16.99	1.000
CG33470	0	4	0	8	177	0.62	17	1.000
IMPPP	0	4	0	8	177	0.62	17	1.000
CG30059	1	7	16	12	387	0.79	16.65	0.040
CG18278	1	7	17	10	388	0.79	16.89	0.018
CG32165	4	34	1	41	761	0.7	16.99	0.190
CG32164	1	25	0	40	761	0.7	16.99	0.394
CG11659	10	8	19	9	408	0.74	16.74	0.530
CG6300	5	8	13	12	404	0.74	16.92	0.506
tHMG1	1	16	1	6	65	0.47	17	0.510
tHMG2	5	8	1	1	68	0.49	16.99	1.000
CG11700	0	29	3	32	242	0.45	16.97	0.250
Ubi-p5E	0	0	2	26	386	0.72	16.98	1.000
CG32588	0	18	0	2	47	0.25	16.96	1.000
CG33252	0	11	1	2	49	0.27	17	0.214
CG9123	2	27	0	10	268	0.56	16.98	1.000
CG12608	0	13	5	11	275	0.58	16.95	0.048

a: Polymorphism and divergence data come from the 17 DPGP2 'core' RG genomes with low admixture. Changes were polarized using D. simulans and D. yakuba. Singletons were excluded from the analysis. R replacement, S synonymous, P polymorphism, F fixed.

b: Number of codons included in calculations

c: Number of codons included in calculations / total number of codons in the gene

d: Average number of genomes (out of 17) a codon was called in

Table A.8: PAML analysis of d_N, d_S and d_N/d_S (ω) values under a free-ratio model following the tree ((((Dmel new gene, Dmel parent gene),(Dsim,Dsec)),(Dyak,Dere)),Dana).

		ω anc-new			**ω anc-parent**			**ω sim-anc**		
New	**Parent**	**dN**	***dS***	ω	***dN***	***dS***	ω	***dN***	***dS***	ω
CG31958	*CG31960*	0.009642	0.019632	0.49	0.000005	0.052297	0	0.030247	0.058723	0.52
Ada1-2	*Ada1-1*	0.000002	0.000001	2.76	0.001393	0.005031	0.28	0.006976	0.112934	0.06
CG18789	*CG18787*	0.004669	0.005496	0.85	0.005581	0.011568	0.48	0.01538	0.057273	0.27
ProtB	*ProtA*	0.009972	0.031103	0.32	0.036486	0.013323	2.74	0.016156	0.066823	0.24
CG31683	*CG18858*	0.002055	0.000002	999	0.000002	0.000001	2.15	0.006284	0.048499	0.13
RpS15Ab	*RpS15Aa*	0.009468	0.000031	302.96	0.000006	0.064202	0	0.000013	0.125585	0
CG33470	*IMPPP*	0.000002	0	47.48	0.000002	0	11.45	0.006913	0.053216	0.12
CG30059	*CG18278*	0.002712	0.01032	0.26	0.001762	0.016556	0.11	0.005532	0.038009	0.15
CG32165	*CG32164*	0.007162	0.007769	0.92	0.002846	0.006544	0.43	0.015713	0.069361	0.23
CG11659	*CG6300*	0.00435	0.002763	1.57	0.001615	0.019219	0.08	0.011346	0.063072	0.18
tHMG1	*tHMG2*	0.099642	0.153305	0.65	0.040811	0.039624	1.03	0.00788	0.063649	0.12
CG32588	*CG33252*	0.263115	0.163404	1.61	0.108715	0.108714	1	0.007168	0.000007	999
CG9123	*CG12608*	0.041308	0.026096	1.58	0.000736	0.031434	0.02	0.021871	0.037971	0.58

a: anc - *D. melanogaster*-specific ancestral, pre-duplication gene copy (i.e. the gene copy that duplicated to form the new gene and parent gene. Branch sim-anc, for example, is the branch from the *D. simulans / D. melanogaster* split to the pre-duplication ancestral gene.

Table A.9: Tajima's *D*, Fay and Wu's *H*, and Hudson-Kreitman-Aguadé-like statistic *p*-values for windows containing *D. melanogaster*-specific duplications.

	100 IS Window, 50 IS step			250 IS window, 50 IS step			
genes[a]	*D*	*H*	HKAl *p*[b]	*D*	*H*	HKAl *p*[b]	Sweep Finder[c]
CG31958; CG31960	-0.97	-6.15	+0.099	-0.84	-3.55	+0.001	0
CG18789; Ada1-2; Qtzl; CG18787; Ada1-1	**-1.70	-4.94	-1.21e-11	***-1.76	**-51.51	-1.22e-19	1
ProtA; ProtB	-1.23	-4.03	-0.449	*-1.20	-5.26	-0.1855	1
CG31683; CG31687; CG18858	-1.18	***-43.37	-0.174	***-1.75	***-103.21	-6.87e-6	0
RpS15Ab	-0.54	-8.27	-0.400	-0.80	-22.06	+0.010	0
CG30059; CG33470; CG18278; IMPPP	-0.71	3.59	+0.155	-0.78	-11.56	+0.021	0
CG32165; CG32164	**-1.65	-4.45	-1.41e-19	-0.78	-10.54	-4.60e-46	0
CG12592	0.3	2.42	-5.11e-13	-0.26	0.41	-9.07e-27	0
CG11659; CG6300	-0.54	-10.67	-0.189	-0.57	-11.54	-0.785	0
tHMG1; tHMG2	-0.66	-2.29	-0.928	*-1.06	6.34	+0.030	1
CG11700; Ubi-p5E	**-1.82	-5.18	+0.361	**-1.61	-2.06	+2.74e-4	0
RpS15Aa	-0.6	-1.87	+0.122	-0.69	-19.67	+0.100	0
CG32588	-0.95	-1.49	+2.64e-6	-1.16	0.9	+3.82e-8	0
CG9123; CG12608	*-1.63	4.66	-1.43e-5	*-1.37	-3.35	-8.55e-17	1
CG33252	-0.86	4.42	+0.017	-0.92	-3.93	+7.44e-4	0

a: Genes contained within a single duplication are analysed together because duplicate regions were masked for this analysis.

b: Signed HKAl χ^2 test *p* values. Negative values indicate a deficiency of segregating sites relative to divergent sites.

c: From Pool et al. (2012)

note: *p* values are determined relative to the empirical distribution of values for windows on the same chromosome arm.

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Table A.10: Summary of *CG32164* and *CG32165* fertility assays.

	Males			**Females**		
	reps	**sterile**	**mean offspring (95% CI)**	**reps**	**sterile**	**mean offspring (95% CI)**
CG32165^{p4d1}	50	0	55.4 (35.0-80.3)	50	9	34.6 (20.0-51.0)
CG32165^{p1c5}	50	50	-	44	13	43.0 (12.8-55.5)
CG32165^{p4g2}	50	50	-	48	8	40.8 (22.0-57.2)
CG32165^{RC}	30	0	52.1 (39.5-65.3)	30	0	31.5(21.9-42.0)
CG32164^{p3d8}	50	0	55.3 (34.2-84.7)	50	5	33.7 (18.1-49.9)
CG32164^{p7f7}	48	7	64.5 (35.0-94.0)	50	50	-
CG32164^{p7c8}	50	5	63.2 (36.8-99.3)	42	42	-
CG32164^{RC}	30	0	51.6 (41.2-63.3)	30	0	38.6 (21.6-36.3)

APPENDIX B

ADDITIONAL PUBLICATIONS

This appendix contains publications that I co-authored with citations and details about my contributions to each project.

B.0.1 Vibranovski et al. (2012) BMC Evolutionary Biology

An excess of new genes formed by retrotransposition (retrogenes) were duplicated from X-linked genes and inserted onto autosomes. Furthermore, new retrogenes tend to be expressed at higher levels in males than in females (i.e. have male-baised expression). Several hypotheses have been put forward to explain these observations [28, 242–244]. Metta and Schlötterer (2010) claimed that non-random X→A movement is an intrinsic property of *Drosophila* retrogenes [245]. Vibranovski et al. (2012) re-analyzed the dataset used by Metta and Schlötterer (2010) to evaluate their claims [246]. I contributed to the discussion shaping the criticisms of Metta and Schlötterer that we raised in the manuscript. I also analyzed male and female fly whole body whole transcriptome sequences to identify genes with sex-biased expression. That analysis was not included in the final paper, but shaped the major critique that whole-body expression analyses cannot accurately and powerfully detect sex-biased gene expression. I also helped write the manuscript. Vibranovski M.D.,

Zhang Y.E., Kemkemer C., VanKuren N.W., Lopes H.F, Karr T.L, and Long M. Segmental dataset and whole body expression data do not support the hypothesis that non-random movement is an intrinsic property of Drosophila retrogenes. *BMC Evolutionary Biology* **12**, 169, (2012). doi:10.1186/1471-2148-12-169

B.0.2 VanKuren and Vibranovski (2013) Journal of Genomics

Males and females of a single species frequently have strikingly different morphologies and behaviors despite the fact that the two sexes share most of the same DNA. These phenotypic differences between sexes are primarily produced by the differential expression of genes between the sexes (i.e. sex-biased gene expression). Thus, accurate identification and characterization of genes that exhibit sex-biased or even sex-specific expression is crucial for understanding the causes of sexual dimorphism, the evolution of sex-biased genes, and selection in relation to sex (see ref. [247] for review). The majority of sex-biased gene expression in Drosophila occurs in the gonads [248] and while many resources exist for studying *D. melanogaster* gene expression, few exist for other Drosophila species. The goal of VanKuren and Vibranovski (2013) was to generate accurate whole-transcriptome expression data for sex-specific tissues (testis, accessory gland, and ovary) for a broad range of Drosophila species to enable accurate identification genes with sex-biased expression.

This paper describes the first phase of the sequencing project conceived of and started by Maria D. Vibranovski, currently an Associate Professor at the Universidade de São Paulo, Brazil. Maria and I collected testis and ovary samples. I carried out all analyses. I wrote the paper under Maria's supervision. VanKuren N.W. and Vibranovski M.D. A novel dataset for identifying sex-biased genes in Drosophila. *Journal of Genomics* **2**, 64-67 (2014). doi: 10.7150/jgen.7955

B.0.3 Long et al. (2013) Annual Review of Genetics

This review focus on the rates, patterns, and models of new gene origination in a wide range of taxa. I collected references, contributed to discussions about content, and co-wrote this review with M.L. based on outlines and information provided by S. C. and M. D. V. Long M., VanKuren N.W., Chen S., and Vibranovski M.D. New Gene Evolution: Little Did We Know. *Annual Review of Genetics* **47**, 307-333 (2013). doi:10.1146/annurev-genet-111212-133301

B.0.4 Gao et al. (2014) Genome Research

Genes with male-biased expression patterns are underrepresented on the X chromosomes of mammals, flies, and worms. Several hypotheses have been proposed to explain this non-random distribution and each has experimental support from analyses of male-biased protein coding genes. This paper investigated the existence and chromosomal distribution of male-biased intergenic non-coding RNAs in *D. melanogaster* and tested models of male-biased gene location evolution using whole-genome tiling arrays. I collected thorax and gut tissue samples used to generate Figures 1 and 2 and helped write the paper. Gao G., Vibranovski

M.D., Zhang L., Li Z., Liu M., Zhang Y.E., Li X., Zhang W., Fan Q., VanKuren N.W., Long M., and Wei L. A long-term demasculinization of X-linked intergenic noncoding RNAs in *Drosophila melanogaster*. *Genome Research* **24**, 629-638 (2014). doi:10.1101/gr.165837.113