DROSOPHILA COMPARATIVE GENOMICS: THE EVOLUTION OF PROTEIN-CODING GENES, SEX CHROMOSOMES, AND AN ANCESTRAL Y-AUTOSOME TRANSLOCATION IN *D. PSEUDOOBSCURA*

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Amanda Marie Larracuente

February 2010

DROSOPHILA COMPARATIVE GENOMICS: THE EVOLUTION OF PROTEIN-CODING GENES, SEX CHROMOSOMES, AND AN ANCESTRAL Y-AUTOSOME TRANSLOCATION IN *D. PSEUDOOBSCURA*

Amanda Marie Larracuente, Ph. D.

Cornell University 2010

The recent sequencing of ten new genomes, bringing the total number of sequenced *Drosophila* genomes to 12, allowed *Drosophila* comparative genomics to be done now in the context of a phylogeny. This dissertation describes several studies that take advantage of these advances in *Drosophila* comparative genomics. A central theme in the dissertation is how properties of a gene and its genomic environment affect rates of protein evolution. We find that many genic factors influence protein evolutionary rate, especially the level and breadth of gene expression. The genomic environment, such as local recombination rate and genomic location, can also influence evolutionary rates. Selection at one site can influence selection at linked sites, especially in regions of the genome with low recombination rates. We are able to detect these effects on a genome-wide scale. The sex chromosomes have particularly interesting effects on the evolution of genes. Natural selection is expected to be more efficient on the X chromosome for new, recessive mutations because the X is hemizygous in males. We do not find consistent signals of more efficient positive selection on the X chromosome than the autosomes; however we do find more efficient purifying selection on the X chromosome. A striking example of the impact of genomic location on gene evolution is in *D. pseudoobscura,* where the ancestral *Drosophila* Y chromosome translocated to an autosome. We mapped this translocation to the dot chromosome. We find that the rDNA repeats, which are

responsible for X-Y pairing in male meiosis, were lost from this ancestral Y chromosome, and the current Y chromosome of *D. pseudoobscura* has acquired and amplified the intergenic spacer repeats (IGS) of the rDNA. We hypothesize that the new location of the IGS functions to maintain X-Y pairing in male meiosis in the absence of rDNA. The most interesting feature of the Y-to-dot translocation is that the genes shrank 10-fold after moving. A survey of polymorphism and divergence on the dot revealed significantly reduced levels of variation and frequency spectra skewed towards rare variants. We hypothesize that this is due to selective sweeps from positive selection favoring the shortening of introns and that the most recent selective sweep was approximately 228,000 years ago.

BIOGRAPHICAL SKETCH

Amanda was born and raised in Buffalo, New York. After attending the Buffalo Academy of the Sacred Heart, Amanda went to Canisius College, where she majored in Biology. At Canisius College, she joined the lab of Dr. Sara Morris and participated in research with the Computation Ecology group, which is a collaboration between Dr. Morris and Dr. H. David Sheets in the Physics department. Amanda used open population models to study the ecology of stopover sites, where migrating birds stop along their route to replenish fat stores. Her research focused on neotropical migrants stopping over at the Isles of Shoals in Maine and utilized capture-mark-recapture data. Part of her research involved collecting birds in mist nets on Appledore Island, Maine. During this part of Amanda's undergraduate career, she developed a strong affection for birds and increased love for nature.

The most influential experience in Amanda's academic life was a trip to the Galápagos Islands as part of a natural history course at Canisius College. She spent a week on mainland Ecuador and a week traveling to many of the Galápagos Islands. Witnessing the effects of adaptive radiation in the unique collection of creatures endemic to the Galápagos helped foster an intense interest in evolution for Amanda. She decided to pursue a graduate career in genetics researching evolutionary questions at Cornell University in Ithaca, NY. Impressed by his enthusiasm for evolutionary questions and his unique approach to solving problems, Amanda joined the lab of Andrew G. Clark to study Y chromosome evolution in *Drosophila*. This provided a wealth of opportunities for Amanda: she was fortunate enough to cross paths with many spectacular scientists in the Clark lab and participate in a large collaborative effort by the *Drosophila* community in the *Drosophila* 12 genomes project.

To Mark and Quinn, for making my life wonderful.

ACKNOWLEDGMENTS

Throughout my academic life, I have encountered a countless number of people whom I owe thanks to. The complete devotion of faculty to the students in the Biology Department made Canisius a truly special place to start my scientific career. I am very grateful for the guidance of Dr. Sara Morris, who provided me with so many wonderful opportunities to learn more about the natural world and gave me a strong foundation to grow on as a scientist. I have developed relationships with so many amazing scientists that have affected me in my time at Cornell University that it is difficult to thank them all. In the short time that Daven Presgraves and I overlapped in the Clark Lab, our conversations helped me get started in learning and understanding classical population genetics. I am also grateful for my experience in a course taught by Carlos Bustamante in Advanced Population Genetics, in which I feel I learned more in a single semester than I had in years of college. I'd like to thank my committee members Dan Barbash and Rick Harrison, for helpful discussions and other faculty members, such as Chip Aquadro and Mariana Wolfner, who have influenced my research over the years. For the past 2 years, I have worked very closely with Tim Sackton and Nadia Singh on various *Drosophila* genomics problems. I am grateful for this wonderful and exciting collaboration and will always have very fond memories of the very late nights, the sleep deprivation-induced goofiness and the craziness we went through to pull together the 12 genomes paper.

I am most grateful to my advisor, Andy Clark, for being a terrific mentor. I appreciate Andy's enthusiasm, creative thinking and optimism; these are the things that I have learned the most from and have shaped the way I think about scientific problems. On a personal note, I must thank my family, especially my parents for their unconditional love and support and for providing me with opportunities that allowed me to accomplish everything I have in life. In my six years at Cornell, I have

developed friendships with a group of amazing people that will last a very long time. Thanks to Nirav, Marie, Ryan, Marisa, Leo, Nick F., Chelsea, Nick B., Nadia, Erin and many others for the great times and for helping to balance hard work with great fun. I can't even begin to appropriately express my gratitude to my husband, Mark, for his support through the years, especially since our son Quinn came into the world. Finally, I'd like to thank Quinn for helping me keep everything in perspective.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

Drosophila comparative genomics has moved forward by leaps and bounds since the sequencing of *D. melanogaster* in 2000 (ADAMS *et al*. 2000). The sequencing of the second *Drosophila* species, *D. pseudoobscura,* revealed details of *cis*-regulatory evolution on a genome-wide scale (RICHARDS *et al*. 2005). The addition of ten new *Drosophila* species to the list of whole genome shotgun assemblies (*DROSOPHILA* 12 GENOMES CONSORTIUM 2007) provided an opportunity to explore many important evolutionary questions. These species span a wide range of divergence times, life history traits and ecological histories. This dissertation demonstrates the usefulness of comparative genomics when placed on a phylogeny. A central question in this dissertation is: what affects the evolution of proteins? I explore genome features that influence the rate of protein evolution, both genic traits such as protein length, intron features, the level and breadth of expression, and traits attributable to the genomic environment such as local recombination rate and location in the genome. A large part of this dissertation focuses on the effect of genomic location on the evolution of genes, with special attention to the sex chromosomes. The X chromosome evolution has interesting features due to its hemizygosity in males. Because the X chromosome spends 2/3 of its time in females, it is expected to have an evolutionary history distinct from that of the autosomes. Furthermore, X chromosome hemizygosity in males means that recessive mutations on the X chromosome are immediately exposed to selection when in males. Natural selection should be more efficient on the X chromosome for new mutations that are at least partially recessive. In this dissertation, I discuss tests of this hypothesis and the influence of X-linkage on the rate of protein evolution. The other sex chromosome, the Y chromosome, is male-restricted and non-recombining, which reduces the efficacy of selection. The *Drosophila* Y is relatively gene poor, yet essential for male fertility, and is highly specialized: all identified Y-linked genes have a male-related function. A most

impressive example of the influence of genomic environment on the evolution of genes is the translocation of the ancestral Y chromosome to an autosome in *D. pseudoobscura* (CARVALHO and CLARK 2005). I discuss the mapping of this translocated Y chromosome to a small, mainly heterochromatic autosome called the dot chromosome. Males of most *Drosophila* species do not recombine during meiosis. The autosomes in males pair at multiple homologous sites along their length. However, the X and Y chromosomes do not have many homologous regions outside of the rDNA repeats. Instead, the intergenic spacer regions of the rDNA repeats function to ensure X-Y pairing during meiosis. I discuss the mapping of the rDNA repeats in *D. pseudoobscura* and closely related species and hypothesize about the mechanism of X-Y pairing in *D. pseudoobscura*. Finally, the translocation of genes that have been segregating exclusively in males for millions of years to the dot chromosome where they are passed through males and females, would bring new selection pressures. The most fascinating aspect of this translocation is that the introns and intergenic regions shrank ten-fold after translocating. In this dissertation, I explore patterns of variation on the dot chromosome of *D. pseudoobscura* and test the hypothesis that the drastic size reduction of the region was accomplished through positive selection favoring deletions in introns.

REFERENCES

ADAMS, M. D., S. E. CELNIKER, R. A. HOLT, C. A. EVANS, J. D. GOCAYNE *et al.*, 2000 The genome sequence of Drosophila melanogaster. Science **287:** 2185-2195.

CARVALHO, A. B., and A. G. CLARK, 2005 Y chromosome of *D. pseudoobscura* is not homologous to the ancestral *Drosophila* Y. Science **307:** 108-110.

DROSOPHILA 12 GENOMES CONSORTIUM, 2007 Evolution of Genes and Genomes on the *Drosophila* Phylogeny. Nature **450:** 203-218.

RICHARDS, S., Y. LIU, B. R. BETTENCOURT, P. HRADECKY, S. LETOVSKY *et al.*, 2005 Comparative genome sequencing of *Drosophila pseudoobscura*: Chromosomal, gene, and *cis*-element evolution. Genome Research **15:** 1-18.

CHAPTER 1[1]

EVOLUTION OF PROTEIN-CODING GENES IN DROSOPHILA

Introduction: Determinants of protein evolution across taxa

Understanding what governs variation in evolutionary rate among proteins is a longstanding biological problem (ZUCKERKANDL 1965). The recent availability of several complete genomes from yeast, insects, and mammals, as well as the preponderance of functional genomic datasets measuring factors such as gene expression (CHINTAPALLI *et al.* 2007; ZHANG *et al.* 2007), gene dispensability (effect of the loss of a particular gene has on the fitness of the organism) (GIAEVER *et al.* 2002), and protein–protein interactions (GIOT *et al.* 2003), has rekindled interest in this problem.

The coupling of these genomic datasets with improved statistical methodologies has revealed a core set of factors that correlate with rates of protein evolution. In yeast, gene expression level seems to be the major correlate of protein evolutionary rate, likely reflecting the importance of translational selection (DRUMMOND *et al.* 2005; DRUMMOND *et al.* 2006; PAL *et al.* 2001 but see WALL *et al.* 2005). Indeed, highly expressed genes evolve slowly in green algae (POPESCU *et al.* 2006), bacteria (ROCHA and DANCHIN 2004), *Drosophila* (LEMOS *et al.* 2005; MARAIS *et al.* 2004), vertebrates (SUBRAMANIAN and KUMAR 2004), *Arabidopsis* (WRIGHT *et*

[1] This chapter is a modified version of a publication that appeared in *Trends in Genetics* (Larracuente, A.M., T.B.Sackton, A.J. Greenberg, A. Wong, N.D. Singh, D. Sturgill, Y. Zhang, B. Oliver, and A.G. Clark. 2008. Evolution of protein-coding genes in *Drosophila*. *Trends. Genet.* **24**(3): 114-123) and is reprinted with permission. T.B.S. ran PAML and edited text in the publication, A.J.G. helped set up the partial correlation analyses and dispensability categories, A.W. calculated τ, we used the expression dataset from D.S., Y.Z. and B.O., N.D.S calculated recombination estimates and N.D.S. and A.G.C. contributed to text editing in the publication.

al. 2004), and poplars (INGVARSSON 2007), suggesting the ubiquitous relevance of translational selection to protein evolution.

In yeast, the dispensability of a gene for organismal growth and the number and structure of protein–protein interactions also influence rates of protein evolution. Dispensability and rate of protein evolution are significantly negatively correlated (HIRSH and FRASER 2001), although whether this correlation remains after controlling for gene expression variation remains controversial (DRUMMOND *et al.* 2006; PAL *et al.* 2003; WALL *et al.* 2005; ZHANG and HE 2005). The reported negative correlation between the number of protein–protein interactions and evolutionary rate (FRASER *et al.* 2002; TEICHMANN 2002) has also proven controversial (HAHN *et al.* 2004; JORDAN *et al.* 2003, but see FRASER *et al.* 2003). Thus, although the contributions of protein dispensability, protein interaction networks, and other factors (e.g. protein structural constraints; LIN *et al.* 2007) cannot yet be definitively ruled out in yeast, translational selection (as measured by gene expression levels) seems to be the strongest and most consistent determinant of protein evolutionary rates. Although studies in multicellular organisms lag behind those in yeast, both tissue bias in gene expression (DURET and MOUCHIROUD 2000; INGVARSSON 2007; LIAO *et al.* 2006; WRIGHT *et al.* 2004) and developmental timing (DAVIS *et al.* 2005; GOOD and NACHMAN 2005) appear to correlate with rates of protein evolution, suggesting that developmental processes and cell-type diversity unique to multicellular organisms are also relevant.

Despite our increasing understanding of the correlates of rates of protein evolution, the evolutionary mechanisms by which these factors influence rates of evolution remain unclear. Positive selection can increase rates of protein evolution above neutrality through fixation of advantageous alleles, whereas purifying selection leads to lower rates of protein evolution than expected under neutrality through the removal of deleterious mutations.

Hill-Robertson interference

Tight linkage among sites can also inhibit the fixation of adaptive mutations and the elimination of deleterious mutations due to interference caused by selection acting at linked sites ("Hill-Robertson interference"; FELSENSTEIN 1974; HILL and ROBERTSON 1966; MCVEAN and CHARLESWORTH 2000), reducing the efficacy of selection. There are numerous mechanisms that can generate interference; here, we describe two examples. In clonal interference (Figure 1.1A), lack of recombination between sites results in reduced adaptation (because segregating beneficial mutations on different haplotypes cannot both fix in the population; FISHER 1930). In this case, a deleterious mutation might be dragged to fixation as a consequence (genetic hitchhiking; SMITH and HAIGH 1974), showing how purifying selection can be inefficient at purging comparatively weak deleterious mutations when they are linked to a beneficial mutation. "Ruby in the rubbish" (Figure 1.1B) is a special case where purifying selection against deleterious mutations (background selection; CHARLESWORTH *et al.* 1993) results in reduced adaptation: in this case, the failure to recombine beneficial mutations (with a small fitness effect) off a deleterious background prevents the fixation of the beneficial allele (PECK 1994).

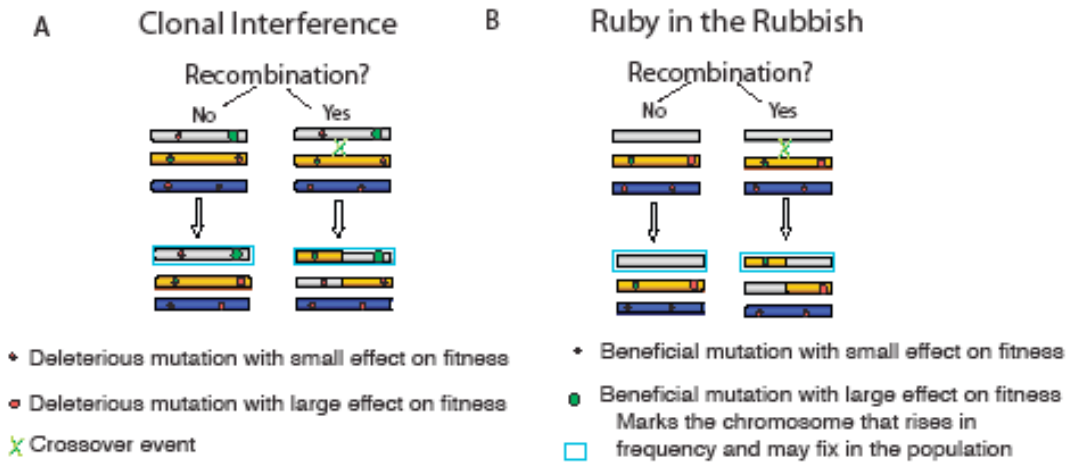


Figure 1.1. Examples of Hill-Robertson Interference. Two models of interference are diagrammed. (A) A demonstration of how positive selection for a beneficial mutation can generate clonal interference, which can lead to reduced adaptation and the hitchhiking of deleterious mutations. (B) A demonstration of the "ruby in the rubbish" model where purifying selection against a deleterious mutation leads to reduced adaptation.

Under a model of interference, we expect that evolutionary dynamics will differ between genes evolving strictly under purifying selection and those which experience positive selection, because interference can lead to an increase in the number of deleterious mutations that fix by drift, while decreasing the fixation of advantageous mutations. More efficient selection will thus lead to increased rates of amino acid fixations in positively selected genes (higher d_N; see BETANCOURT and PRESGRAVES 2002; ZHANG and PARSCH 2005) and a decreased substitution rate due to more efficient removal of deleterious mutations in genes evolving only under purifying selection (lower d_N; see HADDRILL *et al.* 2007). Thus, to detect the signature of interference in genomic data, it is necessary to assess whether or not a substantial number of amino acid fixations in a gene have occurred because of positive selection.

In *Drosophila*, interference has been shown to affect weak selection at synonymous sites (BETANCOURT and PRESGRAVES 2002; MARAIS *et al.* 2004; MARAIS

et al. 2005), but empirical evidence for interference at nonsynonymous sites (BETANCOURT and PRESGRAVES 2002; PRESGRAVES 2005; ZHANG and PARSCH 2005) has not been consistently replicated (MARAIS and CHARLESWORTH 2003; MARAIS *et al.* 2004) outside of regions with no recombination (BACHTROG 2003; HADDRILL *et al.* 2007; MCVEAN and CHARLESWORTH 2000).

Molecular evolution in the comparative genomics era

Codon substitution models provide a comprehensive framework for modeling how protein sequences evolve. Originally developed by Goldman and Yang (GOLDMAN and YANG 1994) and Muse and Gaut (MUSE and GAUT 1994), and implemented in the software package PAML (YANG 1997), these models consider the evolution of codons on a phylogeny of species using a maximum likelihood framework, allowing the estimation of parameters such as ω (the d_N/d_S ratio), which is often used as a measure of the amount of evolutionary constraint on a protein. Furthermore, by comparing the likelihood of the data under different models that make different assumptions about how ω varies among codons in a gene or among lineages in a phylogeny, these maximum likelihood models make it possible to test a number of evolutionary hypotheses (see YANG 2002 for a good review). In particular, by comparing the likelihood of the data under a model that requires a certain proportion of codons in a gene to have $\omega > 1$ (a commonly used signature for positive selection), to the likelihood of the data under a model that assumes all codons in a gene have $\omega \leq 1$, it is possible to test for a signature of selection on a subset of codons in a gene (YANG *et al.* 2000). Using a Bayesian approach based on this framework, it is also possible to estimate both ω and probability of positive selection at individual codons in a gene (YANG *et al.* 2005).

These models make two simplifying assumptions that are probably rarely true in real data: that silent substitutions are neutral, and that the mutational process is at

equilibrium. Although the genomic data from *Drosophila* violate these assumptions, there are several reasons to believe that our conclusions are not substantially affected. First, if variation in selection at silent sites does not substantially affect patterns of protein evolution that we observe, we would expect d_N and ω to show similar patterns, which is the case (Appendix Tables 1.1-1.5). Furthermore, there are no significant differences in either d_S or divergence at four-fold degenerate synonymous sites between positively selected and not positively selected genes, suggesting that variation in synonymous site evolution does not tend to bias our detection of positive selection.

The availability of complete genome sequences from a large number of related species (DROSOPHILA 12 GENOMES CONSORTIUM 2007) provides the opportunity to use these models across entire genomes to estimate ω and the probability of positive selection for every orthologous gene in the genome (limited only by the ability to accurately identify orthologs and produce alignments, as these methods can be sensitive to misidentification of orthologs and inaccurate alignments). The *Drosophila* genomes (Figure 1.2) are ideally suited to this sort of analysis, as six of the 12 species with sequenced genomes are members of a closely related clade with an ideal level of evolutionary divergence for codon-based methods.

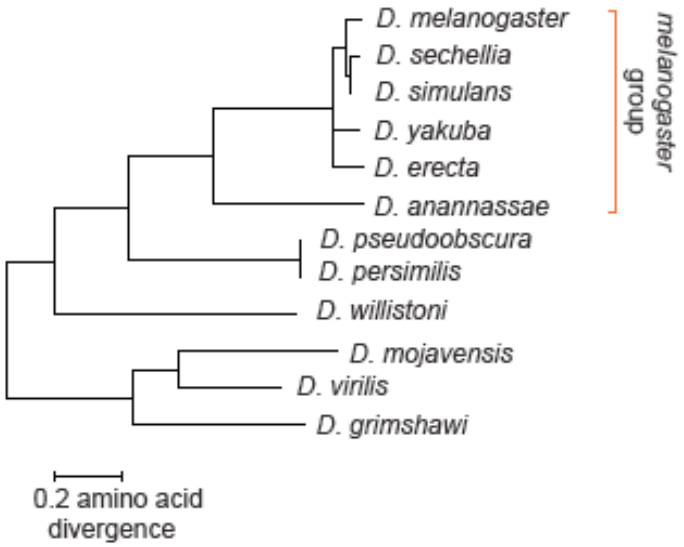


Figure 1.2. Phylogeny of the 12 sequenced *Drosophila* species. Phylogeny of the 12 sequenced *Drosophila* species where the branch lengths are scaled by the amino acid divergence.

Here, we use the newly available *Drosophila* genomes to infer the history of purifying or positive selection, and estimate rates of protein evolution, in single-copy orthologs. Using these data, we discuss why proteins evolve at different rates and shed new light on the factors impacting protein evolution.

Materials and Methods

Fitting codon based maximum likelihood models to *Drosophila* genomic data

The 12 sequenced *Drosophila* genomes present a unique opportunity to fit codon-based maximum likelihood models of molecular evolution on a genomic scale. We started with the masked alignments of all single-copy orthologs in the *melanogaster* group (the following six species: *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, *D. ananassae*), available at ftp://ftp.flybase.net/genomes/12_species_analysis/clark_eisen/alignments/melangaster_group.guide_tree.longest.cds.masked.tar.gz. We exclude paralogs from consideration because of difficulties in computationally verifying the accuracy of phylogenies and of

alignments, as the methods we describe are very sensitive to alignment quality. Although these species represent only half of the 12 for which we now have genome sequence, divergence at silent sites is too great (saturated) beyond the *melanogaster* group. Saturation at silent sites prevents accurate estimation of d_S, and thus would erode the power to accurately estimate both rates of evolution (d_S and ω) and patterns of positive selection. Restricting our analysis to just the set of single-copy orthologs introduces an ascertainment bias such that some rapidly evolving genes or genes restricted to a particular lineage are not examined. A discussion of gene families and lineage-restricted genes appears in *Drosophila* 12 Genomes Consortium (2007).

We used PAML version 3.15 to fit codon-substitution models of molecular evolution. For each alignment of a set of genes with only a single ortholog in the *melanogaster* group, we ran PAML models M0, M7, and M8 (GOLDMAN and YANG 1994; YANG 1997; YANG and NIELSEN 2002; YANG *et al.* 2005). Model M0 assumes a single ω for each gene, and is thus the simplest model. We use the M0 estimates of ω for all cases where we need a single point estimate of the degree of constraint on a given gene across the entire *melanogaster* group phylogeny. Models M7 and M8 allow ω to vary among sites in a given gene. Model M7 assumes that the distribution of ω among sites follows a beta (0,1) distribution. This allows a wide number of possible shapes of the beta distribution, but constrains the distribution to exclude positively selected sites (those with $\omega > 1$). To test for positive selection, we compare model M8 with model M7. Model M8 is equivalent to model M7, except that it adds an additional class of codons with $\omega > 1$. Thus, if M8 fits the data significantly better than M7, it indicates statistical support for a class of positively selected codons.

Because the topology of the *melanogaster* species group (*D. melanogaster, D. simulans* and *D. sechellia*) relative to *D. yakuba* and *D. erecta* is uncertain (POLLARD *et al.* 2006; WONG *et al.* 2007) we ran every model on all three possible tree

topologies for every gene. For the analyses presented here, we used the results of the tree with the best likelihood, averaged across all models ran, although we note that the results across trees are extremely consistent. The parameter estimates generated using the data for just the best-supported tree (the tree with *D. yakuba* and *D. erecta* as sister species) are essentially identical to the parameter estimates generated using the maximum likelihood tree for each gene (data not shown). In order to eliminate the possibility that our maximum likelihood estimates represent local, rather than global, optima, we ran all PAML models multiple times, and used the results from the run with the best likelihood (although likelihoods differ between runs for only a very small fraction of alignments). The PAML output for the best tree for each alignment is available at ftp://ftp.flybase.net/genomes/12_species_analysis/clark_eisen/paml.

To obtain P values for the M8 vs. M7 comparison, we simulated alignments using a modified version of evolver (evolverNSsites) provided in the PAML package to generate data under the null model M7. For each of the null hypotheses, 1000 alignments were chosen at random with the sole criterion that the alignments have more than 100 codons. The empirical codon frequencies, estimated κ, and estimated ω were obtained from codeml output for each randomly chosen alignment and were used to generate 12,000 simulated alignments (12 replicates of 1000 alignments) for each null model. Model M7 and M8 were then run on these simulated alignments to generate an empirical null distribution of likelihood ratio test statistics. This empirical null distribution was converted to an empirical cumulative probability distribution using the ecdf function in R, and then used to calculate P values for M7 vs. M8 test of positive selection.

In addition to these site models, we also ran a series of branch models, also implemented in PAML version 3.15. These models allow ω to vary among branches, but not among sites (YANG 1998). In each case, we compare a model where there are

two estimates of ω allowed on the phylogeny to model M0, which allows only one estimate of ω. We ran five of these branch models, one for each of the terminal lineages (excluding *D. ananassae*) in the *melanogaster* subgroup, so that we have an estimate of ω for the *D. melanogaster* terminal lineage, the *D. simulans* terminal lineage, the *D. sechellia* terminal lineage, the *D. yakuba* terminal lineage, and the *D. erecta* terminal lineage. Previous simulations have shown that likelihood ratio tests from these branch models are well-behaved, so we used the standard χ^2 approximation to estimate whether each terminal branch has a significantly different ω from the rest of the phylogeny. Based on these branch models, we can then classify, for each terminal lineage, genes as significantly accelerated, significantly decelerated, or unchanged with respect to the ω. We use the *D. melanogaster* lineage-specific estimates of ω to compare to parameters that are estimated in *D. melanogaster* and might change between species (*e.g.*, local recombination rate).

To correct for multiple tests, we controlled the false discovery rate (FDR) by calculating q-values (STOREY and TIBSHIRANI 2003). We used the qvalue package in R to perform the calculations. We used the default parameters for P values for branch tests. The distribution of likelihood ratio test statistics for the M8 vs. M7 site test has a point mass at 0, which results in a maximal P value of 0.7354. Therefore, we used the lambda range (0, 0.48] to estimate the underlying uniform distribution of the true null P values (STOREY and TIBSHIRANI 2003). Unless otherwise noted, we consider genes with evidence for positive selection (based on the comparison of the M7 and M8 models) at an FDR of 10% to be "positively selected" (M8_qval column ≤ 0.10 in the PAML_data_summary.tsv, available at ftp://ftp.flybase.net/genomes/12_species_analysis/clark_eisen/paml).

The PAML models implemented implicitly assume that synonymous sites evolve neutrally and that the mutational process is at equilibrium; however most genes

in *Drosophila* violate these assumptions. Selection at synonymous sites might be expected to reduce the number of synonymous fixations. The extent to which selection on synonymous sites influences inferences of protein evolution in PAML as a consequence of selection on synonymous sites has not been assessed as of yet. However, several lines of evidence suggest that these non-equilibrium processes do not substantially impact our conclusions. There are no significant differences in d_S between genes in which we infer positive selection (median d_S = 1.772) and genes where there is no evidence for positive selection (median d_S = 1.776; Mann-Whitney U test P=0.31). Furthermore, positively selected genes are actually slightly less likely to have a d_S value in either the upper or lower quartile of d_S (where quartiles are calculated based on the entire dataset). This argues strongly against any d_S bias in our test for positive selection, as selection on d_S would be predicted to lead to more genes with low d_S in the positive selection class, not a slight deficit. Because these tests were done with d_S estimated using codon substitution models implemented in PAML using all synonymous sites, we repeated them with divergence at only four-fold degenerate synonymous sites (d_{S4} ; estimated using baseml; SINGH *et al.* 2008) as these sites should not lead to biased estimates of divergence (BIELAWSKI *et al.* 2000; BIERNE and EYRE-WALKER 2003). When we compared d_{S4} between positively selected and not positively selected genes, we also did not see a significant difference (median d_{S4} for positively selected and not positively selected genes are 1.3721 and 1.36774, respectively; Mann-Whitney U test P=0.7964). There is also no bias toward genes in the tails of the d_{S4} distribution being positively selected. Therefore, variation in synonymous site evolution does not appear to bias our detection of positive selection.

We used the test of a lineage-specific ω in *D. melanogaster* in order to create a dataset enriched for genes that have experienced recent positive selection, because evolutionary dynamics differ between genes under purifying selection and those under

positive selection. We attempt to elucidate these differences in selective regimes and detect the consequences of Hill-Robertson interference on synonymous and nonsynonymous sites. We achieve this by separately analyzing two datasets: one that is enriched for genes evolving under positive selection (called the "accelerated" set) and one that is not ("not accelerated"). Genes were placed in "accelerated" set if the P value for the *D. melanogaster* branch test for acceleration was significant ($P \leq 0.0045$ corresponding to a false discovery rate of 10%; 219 genes; data from PAML_data_summary.tsv, available at ftp://ftp.flybase.net/genomes/12_species_analysis/clark_eisen/paml). We fully acknowledge that a test for acceleration is not a stringent test for positive selection, and because our "accelerated" dataset is based on an accelerated rate of evolution in *D. melanogaster*, rather than a strict test for positive selection, we cannot rule out the possibility that some of the genes in our "accelerated" set represent genes evolving under relaxed constraint in the *D. melanogaster* lineage. However, several lines of evidence suggest that our conclusions are nonetheless robust. First, our results are qualitatively similar if the "accelerated" and "not accelerated" datasets are defined on the phylogeny-wide test for positive selection (see 'Robustness' section, below). Second, genes with evidence for positive selection based on the more robust phylogeny-wide test (M8 vs. M7) are significantly overrepresented in our *D. melanogaster*-specific "accelerated" set. Finally, contamination of our "accelerated" set with genes evolving by relaxed constraint should make our results conservative, as neutral fixations will not be affected by changes in the efficacy of selection, thus tending to bias the estimates of correlations with ω towards zero.

The alignments used, as well as the output of all PAML models run, are available at FlyBase (ftp://ftp.flybase.net/genomes/12_species_analysis/clark_eisen/).

Estimation of covariates of rates of protein evolution

We compiled data on numerous different properties of *Drosophila* proteins from several different data sources. Here, we briefly describe the source of the data for each parameter.

Expression parameters. Measuring the relevant expression level for studies of this sort is not straightforward, especially considering that expression level measurements tend to be noisy, and that whole organism expression is not necessarily the most relevant value for multicellular organisms with differentiated tissue types. We used data from two different expression studies in this paper: tissue-biased expression from 11 adult tissues (brain, midgut, hindgut, head, crop, Malphigian tubule, testis, ovary, accessory gland, thoracic and abdominal carcass and thoracico-abdominal ganglia) from FlyAtlas (www.flyatlas.org) (CHINTAPALLI *et al.* 2007), and sex-specific whole adult fly expression from species-specific NimbleGen arrays (ZHANG *et al.* 2007). As codon bias is known to correlate strongly with expression level, we use the frequency of optimal codons (FOP) as a third source of information on evolutionarily relevant expression levels. To reduce noise and improve the accuracy of our expression measure, we used the first principle component of either maximal tissue expression from FlyAtlas and codon bias, or the median maximal sex-specific expression across species from Zhang *et al.* (2007) and codon bias. As discussed below, both of these expression estimators give similar results in our analysis.

Degree of tissue bias of expression was measured based on the FlyAtlas data, using the statistic $\tau = \sum_{j=1}^{n} 1 - \left(\frac{\log S(j)}{\log S_{\max}} \right) / (n-1)$, where S is the signal intensity and n is the number of tissues (YANAI *et al.* 2005). The log of signal intensity for a tissue was set to 0 for any gene that was detected on fewer than 2 arrays in that tissue. τ

ranges from 0 to 1, with values close to 0 indicating broadly expressed genes and values close to 1 indicated highly biased genes. At $\tau = 1$, expression is only detectable in a single tissue.

In this study we also estimate divergence in gene expression among the species of the *melanogaster* group, using data from species-specific arrays designed to *D. melanogaster, D. simulans, D. yakuba,* and *D. ananassae* by (ZHANG *et al.* 2007). To calculate pairwise divergence in expression, we rank ordered genes within each sex and each species by expression signal intensity. Subsequently, for pairs of species *i* and *j,* ranks for species *j* were regressed on ranks for species *i*. Our distance measure, D_{ij} was calculated as the studentized residuals (corrected for the standard deviation) between the observed rank in species *j* and its predicted rank according to the regression model. For each sex and each gene, we generated a pairwise distance matrix from all D*ij*. We used the phylip package (FELSENSTEIN 1989) to convert this distance matrix into branch lengths, using the unrooted species tree (((dmel, dsim), dyak), dana). This then allowed us to calculate total tree length for the *melanogaster* group, as well as individual branch lengths for each species. Unless otherwise noted, we present data on the total expression tree length only.

Essential genes. We downloaded information on mutant phenotypes of each gene from FlyBase (http://flybase.net/). We then noted if a gene had mutations described as "lethal," "sterile," "viable" or "visible." Regardless of phenotypic information, all Minute genes and those named *l(x)y* (e.g., *l(2)k14710*) were recorded as having lethal alleles. Likewise, all genes named (*f/m*)*s(x)y* (e.g., *fs(1)K10*) were recorded as having sterile alleles. Some genes have more than one type of allele. We grouped genes into four sets – "essential" (those with lethal or sterile alleles); "viable" (those that have visible and viable mutations, but no lethal or sterile ones); "no

information" (those that have alleles, but no information on their phenotype); and "no alleles" (no alleles listed).

Recombination. Local recombination rate in *D. melanogaster* was calculated by taking all genes that have been mapped on the physical and genetic maps from Release 4.3 of the *D. melanogaster* genome, and fitting a third order polynomial to the genetic position as a function of physical position for each chromosome arm. Recombination rate was estimated as the derivative of this polynomial at the midpoint of each gene. This regression polynomial approach is referred to as the "RP" method throughout.

Protein-protein interaction data. The protein-protein interaction data used was the number of high confidence (as defined by GIOT *et al.* 2003) protein-protein interactions (GIOT *et al.* 2003) downloaded from http://www.thebiogrid.org/.

Statistical analysis

We are interested in estimating the effect of each network parameter on rates of protein evolution. Most of the variables we examined are correlated with one another. Thus it is inadvisable to apply multiple regression due to the problem of collinearity (DRUMMOND *et al.* 2006; WEISBERG 1985). Moreover, pairwise correlations can be misleading (WHITTAKER 1990). Therefore, we used partial correlations, which are defined as correlations between pairs of variables calculated conditional on all other parameters (WHITTAKER 1990). This approach has been widely used in the literature (*i.e.*, WALL *et al.* 2005), and has a number of attractive features: for example, it allows the estimation of associations among all the variables under consideration and, unlike regression, does not imply directionality of effect.

To estimate partial correlations, we calculated the pseudoinverse of correlation matrices, as implemented in the R package corpcor (written by J. Schafer, R. Opgen-

Rhein and K. Strimmer 2006 corpcor: Efficient Estimation of Covariance and Partial Correlation; http://www.strimmerlab.org/software/corpcor/). We estimated partial correlations by inverting the Spearman correlation matrices because of the highly irregular distributions of most of our parameters. To assess significance, we randomly assigned the values of a given parameter to genes, keeping the others constant, and re-estimated the partial correlation matrix. These permutations were performed using the boot package from R (written by A. Canty and B. Ripley 2006 boot: Bootstrap R (S-Plus) Functions; http://CRAN.R-project.org). We then repeated the process for each parameter. Each two-tailed P value was thus calculated twice, and we used the bigger of the two (the slight differences arise due to randomness of the permutations).

It has been argued that noise in the data leads to under-estimation of correlations and thus potentially to spurious partial correlations, and that Principal Component Regression (PCR) might better control collinearity (DRUMMOND *et al.* 2006). This approach involves constructing principal components (PCs) from predictor variables and using the PCs as predictors in regression (DRUMMOND *et al.* 2006; JOLLIFFE 1986). Principal components constructed from our variables were complicated in structure and not easily interpretable. Therefore, we transformed the regression coefficients for the principal components to coefficients for the variables, as described in Jolliffe (chapter 8, JOLLIFFE 1986). We then calculated right-tailed P values by permuting the response variables. We used linear regression when ω was the response variable, and logistic regression when positive selection (a binary character) was the response variable. Although we present only the results of partial correlation analyses, principal component regression leads to generally qualitatively similar conclusions, suggesting that noise in our data does not substantially alter our conclusions.

In order to assess the effect of essentiality on variation in rates of protein evolution, we used a principal component ANCOVA. We constructed PCs from all eight predictor variables (intron number, exon length, intron length, expression level, tissue-bias of expression, number of protein-protein interactions, recombination rate and codon bias) and used them as covariates in analysis of covariance, with gene essentiality as a categorical variable. We calculated P values by randomly re-assigning the values of the response variable to genes and comparing the coefficients estimated from the data to the resulting null distribution.

To assess the effects of factors discussed in this paper (ω, intron number, protein length, intron length, expression, τ, protein-protein interactions, recombination rate, expression divergence and gene essentiality) on whether a gene is likely to experience positive selection or not, we used a logistic regression in R (using GLM and family = binomial; results presented in Appendix table 1.6). We used an FDR cutoff of 10% to identify genes likely evolving under positive selection.

Robustness

In the preparation of a large dataset of genomic covariates of protein evolution, many choices have to be made about how to estimate parameters. In many cases, these choices are arbitrary, as several equally acceptable options exist. In order to determine the extent to which choices about estimation methods impact our conclusions, we tested the robustness of the partial correlation results to alternative methods of estimating expression level and rate of protein evolution. For all the results we present, we observe consistent effects whether we use the FlyAtlas data or the Zhang et al. (ZHANG *et al.* 2007) data (and for a subset of tests, we achieved similar results with Gibson *et al.* 2004 data) to estimate expression level, and whether we use ω estimated by PAML model M0, or *D. melanogaster* lineage-specific ω

estimated by branch models in PAML as our estimate of the rate of protein evolution (data not shown).

Codon bias and d_S

Selection for translational efficiency/accuracy and robustness is expected to lower substitution rates at synonymous sites, leading to a negative correlation between codon bias and d_S. However, some studies that use the maximum likelihood estimates of d_S obtained from codon substitution models implemented in PAML find an unexpected positive correlation (BETANCOURT and PRESGRAVES 2002; MORIYAMA and POWELL 1996), including this paper. This problem has to do with the method of counting synonymous sites and is discussed in detail in Bierne and Eyre-Walker (BIERNE and EYRE-WALKER 2003) and Bielawski et al. (BIELAWSKI *et al.* 2000).

Recombination rates are not conserved across the phylogeny

The rate of recombination for a gene in *Drosophila* can change between species due to changes in recombination environment associated with a change in genomic location or inversion. Recombination rates, in general, appear to be relatively labile across *Drosophila* species (HAMBLIN and AQUADRO 1999; ORTIZ-BARRIENTOS *et al.* 2006) even in the closely related species of the *melanogaster* species complex (i.e. TRUE *et al.* 1996 and HAMBLIN and AQUADRO 1996). We chose to report the partial correlation analysis done on the 'accelerated' and 'not accelerated' datasets using the melanogaster-specific estimates of ω, d_N and d_S because we only have estimates of local recombination rates for *D. melanogaster*. Therefore the estimates of evolutionary rate for the recombination analyses are specifically from the *D. melanogaster* branch rather than from the *melanogaster* group phylogeny.

Results: Correlates of variation in rates of protein evolution in Drosophila

The recent publication of the genome sequences from 12 *Drosophila* species (DROSOPHILA 12 GENOMES CONSORTIUM 2007) presents a unique opportunity to dissect evolutionary mechanisms underlying variation in rates of protein evolution in *Drosophila*. Using sophisticated models of molecular evolution and the largest subset of the phylogeny (the *melanogaster* group) in which synonymous sites are not saturated, we can disentangle purifying and positive selection on a genomic scale across all single-copy orthologous protein coding genes. We identify positively selected genes as those with statistical support (based on codon substitution models implemented in PAML) for a subset of codons where replacement mutations have fixed more rapidly than silent mutations.

Table 1.1 Factors affecting rates of protein evolution. Additional details about how each factor was estimated are available in Materials and Methods

Factor	How it was measured	Correlation with ω	Refs
Gene Expression	Quantified by the first principle component of the maximum expression across tissues from FlyAtlas and codon bias measured by FOP	Strong negative correlation with ω; likely driven by purifying selection against mutations that reduce transcriptional efficiency, translational efficiency, or translational robustness.	(DRUMMOND *et al.* 2005; DRUMMOND *et al.* 2006; INGVARSSON 2007; LEMOS *et al.* 2005; MARAIS *et al.* 2004; PAL *et al.* 2001; POPESCU *et al.* 2006; ROCHA and DANCHIN 2004; SUBRAMANIAN and KUMAR 2004; WRIGHT *et al.* 2004)
Tissue bias in expression	Calculated as τ (YANAI *et al.* 2005), based on expression data for 11 adult tissues from FlyAtlas. τ ranges from 0 to 1, with high values indicating more biased expression	Strong positive correlation with ω; likely driven by more positive selection on tissue biased genes, as well as increased purifying selection on broadly expressed genes	(DURET and MOUCHIROUD 2000; INGVARSSON 2007; LIAO *et al.* 2006; WRIGHT *et al.* 2004)

Table 1.1 (Continued)

Essentiality	Genes assigned to either "essential," "viable", or "no phenotype" based on mutation information available from FlyBase. The "no phenotype" class includes genes with no alleles and genes with uncharacterized alleles	Essential genes are more conserved than non-essential genes; no difference in proportion of positively selected genes between essential and viable classes	(HIRSH and FRASER 2001; HURST and SMITH 1999; LIAO *et al.* 2006; WALL *et al.* 2005; ZHANG and HE 2005)
Intron Number	The number of introns in each gene	Negative correlation with ω; potentially driven by conservation of exonic splice site enhancers	(MARAIS *et al.* 2005; PARMLEY *et al.* 2007)
Intron Length	Average length of all introns in each gene	Non-significant weak negative correlation.	(CARVALHO and CLARK 1999; COMERON and KREITMAN 2000; COMERON and KREITMAN 2002; MARAIS *et al.* 2005)
Protein Length	Length of the coding sequence of each gene	Weak, but significant negative correlation.	(COMERON 2004; COMERON and KREITMAN 2002; COMERON *et al.* 1999; DURET and MOUCHIROUD 1999)
Protein-Protein Interactors	Number of high confidence protein-protein interactors from Giot et al. (GIOT *et al.* 2003)	No significant effect in our study; controversy in the literature over the role of network structure in protein evolution	(FRASER *et al.* 2002; FRASER *et al.* 2003; GIOT *et al.* 2003; HAHN *et al.* 2004; JORDAN *et al.* 2003; TEICHMANN 2002)
Recombination	Estimated based on the physical and genetic maps from *D. melanogaster* genome release 4.3 using a regression polynomial approach	Positive correlation with the efficacy of selection	(BETANCOURT and PRESGRAVES 2002; HADDRILL *et al.* 2007; MARAIS *et al.* 2004; ZHANG and PARSCH 2005)

Similar to other multicellular organisms, gene expression levels in *Drosophila* are negatively correlated with evolutionary rate (measured as ω, the d_N/d_S ratio), and tissue bias in expression is independently positively correlated with evolutionary rate (Figure 1.3A; Table 1.1; Appendix Table 1.1). Additionally, both intron number and

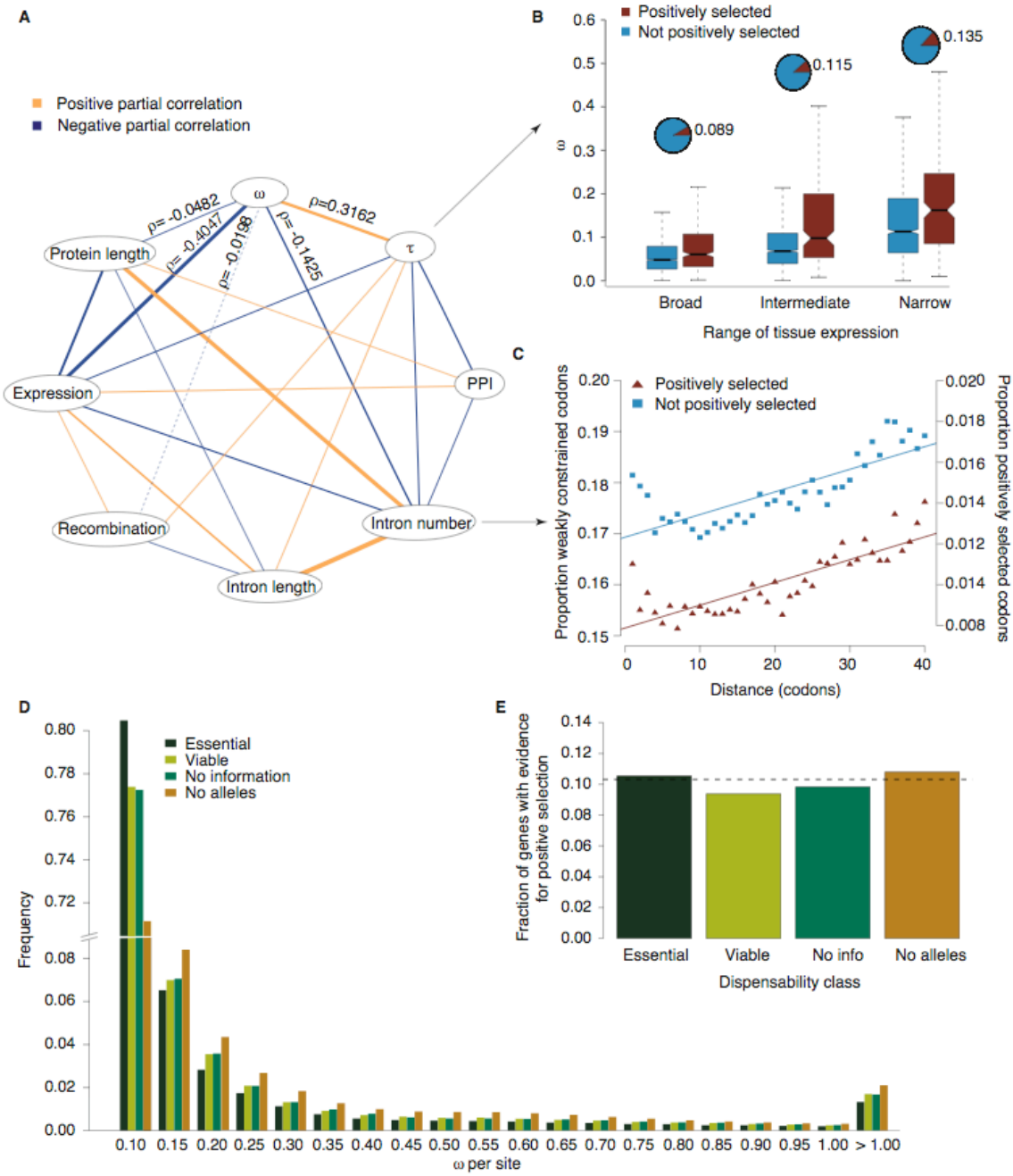
protein length are significantly negatively correlated with ω (Figure 1.3A; Appendix Table 1.1). The number of protein-protein interactions does not appear to significantly correlate with evolutionary rate (Figure 1A; Appendix Table 1.1), although our knowledge of protein-protein interactions in *Drosophila* is still rudimentary (GIOT *et al.* 2003). These results are robust to alternative methods of estimating covariates. The correlation between ω and protein length (Figure 1.3A), however, appears to be at least partially driven by a positive correlation between d_S and protein length (Spearman's partial $\rho = 0.1833$, $P = 4x10^{-4}$).

We also find a significant effect of gene essentiality on rates of protein evolution: when controlling for all continuous variables using principle component regression, essentiality significantly associates with rates of protein evolution (ANCOVA, $P = .0028$). These *Drosophila* data thus tentatively confirm the general model of protein evolution that has recently emerged. In the remaining sections, we present results from the *Drosophila* genomes that illuminate how tissue bias, intron number, essentiality, and interference relate to variation in rates of protein evolution.

Is the relationship between evolutionary rate and degree of tissue bias driven by positive selection?

Beyond overall levels of gene expression, the breadth of tissues across which genes are expressed correlates with rates of protein evolution in multicellular organisms (DURET and MOUCHIROUD 2000; LIAO *et al.* 2006; ZHANG and LI 2004) (Figure 1.3A, 1.3B): ubiquitously expressed genes evolve more slowly than genes with more restricted expression.

Figure 1.3. Factors affecting rates of protein evolution in *Drosophila*. (A) Diagram of all factors included in the partial correlation matrix. Orange lines connecting two factors represent significant positive partial correlations, and blue lines represent significant negative partial correlations. The thickness of the lines correspond to the magnitude of the partial ρ. For correlations with ω, non-significant partial correlations are indicated as dashed lines, and the actual values of ρ are shown. (B) Box plot of ω for genes with broad ($\tau \leq 0.50$), intermediate ($0.50 < \tau < 0.90$), or narrow ($\tau \geq 0.90$) range of tissue expression. Blue boxes show genes with no evidence for positive selection, red boxes show genes with evidence for positive selection. The pie charts show the fraction of genes in each expression class with evidence for positive selection. (C) Plot showing fraction of non-positively-selected codons with $\omega > 0.1$ (blue) and fraction of codons with probability of positive selection > 0.50 (red), as a function of the distance from an exon/intron boundary, measured in codons.
(D) Distributions of ω per-site for each dispensability class ("essential", "viable", "no information", "no alleles"). Essential genes have a significant excess of codons with $\omega < 0.1$. (E) Fraction of genes with evidence for positive selection (FDR 10%) in each dispensability class. There is no significant difference among classes. The dashed line shows the overall proportion of genes with evidence for positive selection.


A
Positive partial correlation
Negative partial correlation
ω
τ
ρ= -0.0482
ρ= -0.4047
ρ= -0.0198
ρ=0.3162
ρ= -0.1425
Protein length
Expression
PPI
Recombination
Intron number
Intron length
B
Positively selected
Not positively selected
0.089
0.115
0.135
ω
Broad
Intermediate
Narrow
Range of tissue expression
C
Positively selected
Not positively selected
Proportion weakly constrained codons
Proportion positively selected codons
Distance (codons)
D
Essential
Viable
No information
No alleles
Frequency
ω per site
E
Fraction of genes with evidence for positive selection
Essential
Viable
No info
No alleles
Dispensability class

This correlation could be driven in part by genes with detectable expression only in male reproductive tissues, which evolve rapidly and are frequent targets of positive selection in many taxa, including *Drosophila* (ELLEGREN and PARSCH 2007; HAERTY *et al.* 2007; PROSCHEL *et al.* 2006) and mammals (CLARK and SWANSON 2005; ELLEGREN and PARSCH 2007). The new *Drosophila* genome data, and the recent publication of FlyAtlas, a *Drosophila* expression atlas covering 11 adult tissues (CHINTAPALLI *et al.* 2007) facilitates explicitly testing this hypothesis.

Tissue bias in gene expression is measured using τ (YANAI *et al.* 2005): low values indicate ubiquitous expression and high values indicate highly biased expression in one or a few tissues. After removing genes with testes- or accessory gland-biased expression, τ remains significantly positively correlated with ω (Spearman's partial $\rho_\tau = 0.2641$, $P=2\text{x}10^{-4}$) and d_N (Spearman's partial $\rho_\tau = 0.2586$, $P=2\text{x}10^{-4}$). Pooled across all tissues, genes with detectable expression in only one tissue have a higher rate of evolution (median $\omega = 0.125$) than ubiquitously expressed genes (those with detectable expression in all 11 tissues; median $\omega = 0.047$; Mann-Whitney U $P < 1\text{x}10^{-16}$); this pattern also holds when each of the 11 tissues in the FlyAtlas dataset are considered individually (data not shown). Thus, the increased evolutionary rate associated with high tissue bias does not appear to be primarily driven by evolutionary patterns among genes expressed in any particular tissue.

However, the elevated rate of protein evolution among narrowly expressed genes appears to be at least partially driven by positive selection (Figure 1.3B). Genes with higher tissue bias are more likely to show evidence for positive selection (logistic regression $\beta = 0.291$, $P = 0.040$; Appendix Table 1.6) and a significantly higher proportion of narrowly expressed genes reject the null hypothesis of no positive selection at a 10% false discovery rate (FDR; 13.7% versus 8.97%, Fisher's Exact Test, $P = 1.84 \text{ x } 10^{-6}$), even after removing genes with testes or accessory gland biased

expression patterns (data not shown). However, differences in patterns of positive selection cannot completely explain the observed pattern: ω and τ remain significantly correlated among genes with no evidence for positive selection (Spearman's partial $\rho = 0.3093$, $P < 2 \times 10^{-4}$; Figure 1.3B).

Ubiquitously expressed genes appear to experience both stronger purifying selection and less frequent positive selection than narrowly expressed genes. It is possible that ubiquitously expressed genes are involved in more cellular and physiological processes than narrowly expressed genes, leading to more extensive pleiotropy, as has been suggested previously (DURET and MOUCHIROUD 2000). Pleiotropy is expected to constrain the fixation of beneficial mutations, as well as increase the strength of purifying selection on a gene (FISHER 1930) consistent with the observation that broadly expressed genes evolve more slowly.

Intron number constrains the rate of protein evolution

Exonic splice site enhancers (ESEs) are short sequences in exons near intron-exon boundaries that aid in ensuring proper intron excision (BLENCOWE 2000). Mammalian ESEs appear highly constrained, with lower rates of both nonsynonymous (d_N) and synonymous site evolution (d_S) near intron-exon boundaries (PARMLEY *et al.* 2007). In *Drosophila*, genes containing introns have a significantly lower d_N than genes lacking introns (median $d_{N\,(\text{introns})} = 0.103$, median $d_{N\,(\text{no introns})} = 0.199$; Mann-Whitney U, $P < 2\text{x}10^{-16}$, see also MARAIS *et al.* 2005), as well as a significantly lower d_S (median $d_{S\,(\text{introns})} = 1.748$, median $d_{S\,(\text{no introns})} = 1.964$; Mann-Whitney U, $P < 2\text{x}10^{-16}$). Furthermore, intron number is significantly negatively correlated with ω (Spearman's partial $\rho = -0.1425$ $P=2\text{x}10^{-4}$; Figure 1.3A; Appendix Table 1.1), d_N (Spearman's partial $\rho = -0.1669$ $P = 2\text{x}10^{-4}$), and d_S (Spearman's partial $\rho = -0.094$ $P=2\text{x}10^{-4}$).

Could constraint in ESEs mediate the effects of intron number on the evolution of proteins in *Drosophila*? If so, we expect a decreased rate of evolution for codons potentially overlapping ESEs. Indeed, codons near intron–exon boundaries have a significantly lower proportion of weakly constrained codons (proportion of non-positively-selected codons with $\omega > 0.1$; Figure 1.3C; $R^2 = 0.61$, $P = 1.33 \times 10^{-9}$). Interestingly, the proportion of codons with evidence for positive selection also decreases near intron–exon boundaries (Figure 1.3C; $R^2 = 0.68$, $P = 3.18 \times 10^{-11}$), suggesting that ESEs might also limit the adaptive evolution of codons, although there does not appear to be a significant decrease in the probability of positive selection in genes with many introns (logistic regression $\beta = -0.001$, $P = 0.948$).

Essential genes evolve more slowly but are no less likely to evolve adaptively

Gene essentiality has been previously associated with significant, if small, decreases in the rate of protein evolution across numerous taxa (HIRSH and FRASER 2001; HURST and SMITH 1999; LIAO *et al.* 2006; WALL *et al.* 2005; ZHANG and HE 2005 but see DRUMMOND *et al.* 2006; PAL *et al.* 2003). Early reports suggested that the elevated ω of non-essential genes results from an excess of positive selection among non-essential genes, as excluding putatively rapidly evolving immune genes eliminated the difference in evolutionary rates between non-essential genes and essential genes in mammals (HURST and SMITH 1999). Alternatively, essential genes might be more conserved because of more intense purifying selection (due to, for example, more severe fitness consequences of mutations that reduce protein function).

To distinguish between these models, we used estimates of ω per-codon to infer the proportion of codons that face strong selective constraint. In *Drosophila,* there is a higher proportion of highly-constrained codons ($\omega < 0.1$) in essential genes than in any other dispensability class (Figure 1.3D), even among genes that share a

similar level of expression (not shown), suggesting that essential genes have a higher fraction of sites that are unavailable for evolutionary modification. By contrast, the fraction of codons with $\omega > 1$ is similar among dispensability classes (Figure 1.3D) indicating that essential genes are equally likely to experience positive selection as non-essential genes. Further, the number of genes rejecting the null hypothesis of no positive selection (at a 10% FDR) is similar among all dispensability classes (Figure 1.3E; df = 3, $\chi^2 = 2.96$, $P = 0.399$); this result is robust to the FDR cutoff used (data not shown) and the confounding effects of other variables (logistic regression $\beta = -0.169$, $P = 0.2611$). Therefore, the essentiality of a gene does not appear to affect the number of mutations that fix by positive selection, but stronger purifying selection on essential genes appears to decrease the number of mutations that fix by drift alone.

Factors contributing to the efficacy of selection

Hill-Robertson interference (HILL and ROBERTSON 1966) reflects the interaction between selective forces at linked sites and is predicted to lead to a reduced efficacy of natural selection as genetic linkage increases (FELSENSTEIN 1974). These effects can result from interference generated by selected alleles at neighboring loci as well as interference between selected sites within a gene. The evolutionary dynamics should differ between genes evolving solely under purifying selection and those under positive selection, as factors increasing the efficacy of selection will correlate with higher ω among positively selected genes but lower ω among negatively selected genes. Previous studies attempting to detect interference focused on either small datasets enriched for rapidly evolving genes (BETANCOURT and PRESGRAVES 2002; ZHANG and PARSCH 2005), or on large datasets without factoring in the mode of selection acting on genes (HADDRILL *et al.* 2007; MARAIS *et al.* 2004), yielding inconsistent results. To resolve these conflicts, we partitioned a dataset enriched for

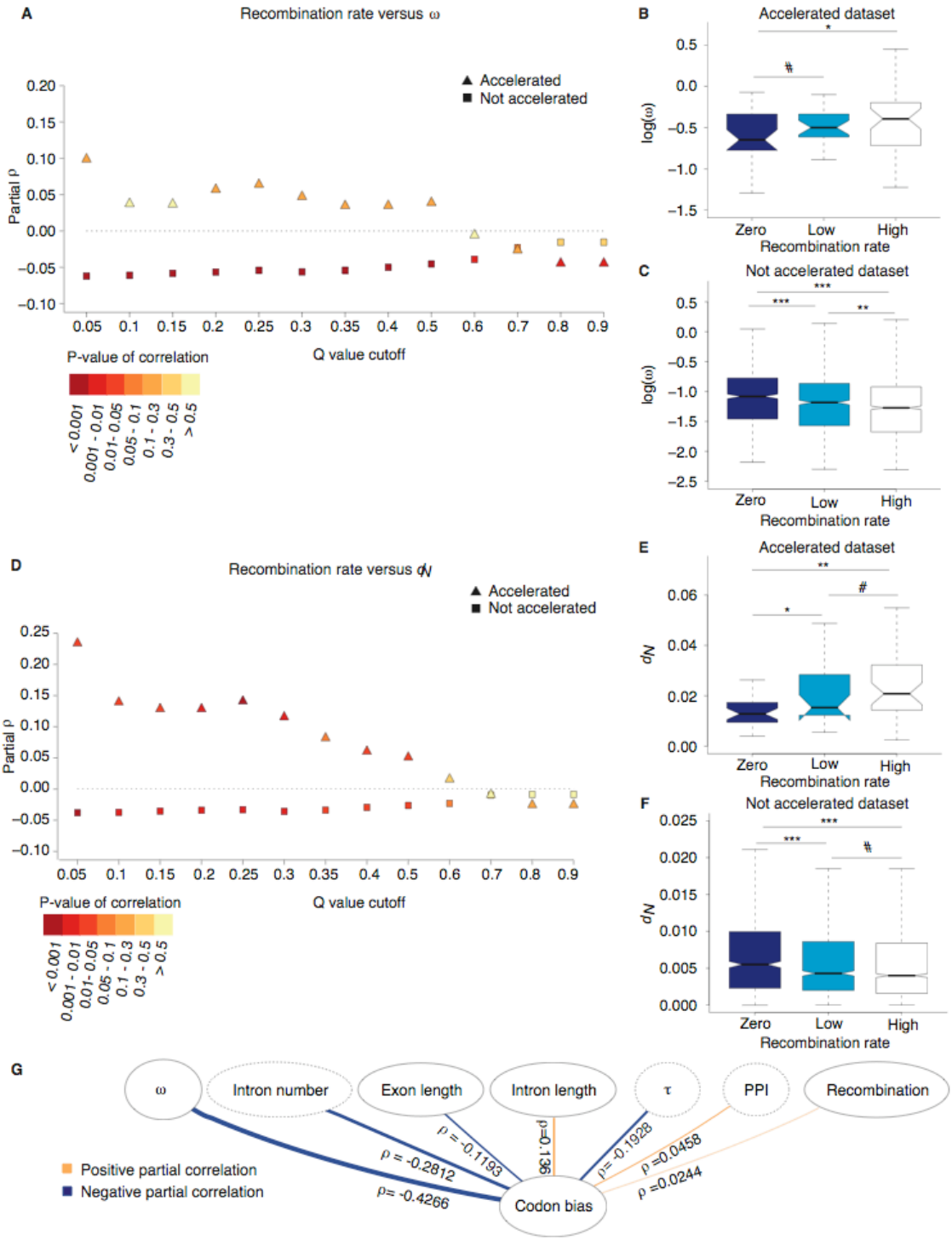
positively selected genes from the rest of the data, which is likely to be enriched for genes evolving mainly under purifying selection.

We assigned to the "accelerated" dataset all genes with evidence for a significantly accelerated rate of evolution (FDR 10%) along the *D. melanogaster* branch, compared to the rest of the phylogeny. Because the "accelerated" dataset is not based on a strict test for positive selection, it is possible that some genes in our "accelerated" set are evolving under relaxed constraint in the *D. melanogaster* lineage. However, several lines of evidence suggest these genes are the minority: our results are qualitatively similar when we define our "accelerated" dataset based on genes with evidence for positive selection from codon-based models, and those positively-selected genes are significantly overrepresented among "accelerated" genes. Additionally, inclusion of comparatively unconstrained genes makes our analysis conservative, as neutral fixations are unaffected by changes in the efficacy of selection.

Recombination enhances the efficacy of purifying and positive selection

Considerable attention has focused on recombination rate variation as a driver of differences in the efficacy of selection (MARAIS and CHARLESWORTH 2003; PRESGRAVES 2005), because recombination increases the independence of sites between loci and, to a lesser degree, within a gene. A comparison of 255 *D. melanogaster* and *D. simulans* orthologs (~25% rapidly evolving *Acps*), revealed a positive correlation between d_N and recombination rate, suggesting that regions of low recombination experience limited adaptation in *Drosophila* (BETANCOURT and PRESGRAVES 2002). Similarly, in *Drosophila* genes with male-biased expression, which are known to evolve rapidly (ELLEGREN and PARSCH 2007; PROSCHEL *et al.* 2006), ω and d_N are significantly positively correlated with recombination rate (ZHANG and PARSCH 2005) .

Figure 1.4. Hill-Robertson interference in *Drosophila*. (A) Scatter plot showing partial correlations between recombination rate and ω for different false discovery rate cutoffs (Q-values) for including genes in the "accelerated" dataset. The x-axis shows the Q-value used to determine whether a gene is assigned to the "accelerated" dataset, so increasing values of this axis indicate increasing numbers of false positives in the "accelerated" set, but also larger sample sizes. Points are shaded relative to the *P*-value of the correlation, with darker points indicating more significant correlations. Note that the magnitude of the correlation between recombination rate and ω decreases with increasing false positives, as expected. (B) Box plots of ω for "accelerated" and (C) "not accelerated" datasets, divided by genes with high (top quintile), low (bottom quintile), or zero recombination. Genes with low recombination have lower ω for the "accelerated" dataset, and higher ω for the "not accelerated" dataset, as predicted by the interference model (# $0.05 < P < 0.1$, * $0.01 < P < 0.05$, ** $0.001 < P < 0.01$, *** $P < 0.001$). (D), (E), and (F) are the same as (A) , (B), and (C), except with d_N instead of ω. (G). Partial correlations between codon bias and other factors in the model. Only significant correlations are shown; variables with solid outlines appear to influence codon bias at least partially via changes in the efficacy of selection at synonymous sites. Orange lines connecting two circles represent significant positive partial correlations, and blue lines represent significant negative partial correlations. The thickness of the lines correspond to the relative magnitude of ρ for each partial correlation, and the actual value of ρ is shown next to the line.


A
Recombination rate versus ω
Accelerated
Not accelerated
Partial ρ
Q value cutoff
P-value of correlation
< 0.001
0.001 - 0.01
0.01- 0.05
0.05 - 0.1
0.1 - 0.3
0.3 - 0.5
> 0.5
B
Accelerated dataset
log(ω)
Zero
Low
High
Recombination rate
C
Not accelerated dataset
D
Recombination rate versus dN
E
Accelerated dataset
dN
F
Not accelerated dataset
G
ω
Intron number
Exon length
Intron length
τ
PPI
Recombination
Codon bias
ρ = -0.2812
ρ = -0.1193
ρ=-0.136
ρ= -0.1928
ρ =0.0458
ρ =0.0244
ρ= -0.4266
Positive partial correlation
Negative partial correlation

However, larger studies that estimated pairwise d_N, d_S, and ω between *D. melanogaster-D. yakuba* orthologs instead found a negative correlation (HADDRILL *et al.* 2007; MARAIS *et al.* 2004).

Because the predicted effect of recombination rate on the rate of protein evolution depends on the mode of selection acting on a gene, we examined the effect of recombination rate on the efficacy of both purifying and positive selection. For the "accelerated" dataset, median ω (Figure 1.4B) and d_N (Figure 1.4E) increase with increasing recombination rate, as predicted by the interference model for genes evolving under positive selection. For the "not accelerated" dataset, median ω (Figure 1.4C) and d_N (Figure 1.4F) are lower in regions of high recombination, also consistent with the interference model. Even low levels of recombination can markedly improve the efficacy of selection (Figure 1.4B, 1.4C, 1.4E, and 1.4F), as predicted by population genetic theory (FISHER 1930).

We also used partial correlations to examine the relationship between evolutionary and recombination rates at a genomic scale in *D. melanogaster.* In the "not accelerated" dataset recombination rate is negatively correlated with both ω (Spearman's partial $\rho = -0.061$, $P < 2 \times 10^{-4}$; Appendix Table 1.2; Figure 1.4A) and d_N (Spearman's partial $\rho = -0.0374$, $P < 8 \times 10^{-4}$; Figure 1.4D; Appendix Table 1.2) consistent with an increased efficacy of selection against deleterious mutations with increased recombination. By contrast, for the "accelerated" dataset, recombination rate and d_N are significantly positively correlated, suggesting that positive selection is more efficacious with increasing recombination (Spearman's partial $\rho = 0.1395$, $P = 0.0382$; Figure 1.4D; Appendix Table 1.3). The correlation between ω and recombination rate trends in the predicted direction (Spearman's partial $\rho = 0.0375$, $P = 0.5932$; Figure 1.4A; Appendix Table 1.3). We do not believe this relationship is mediated by other factors, such as d_S, which is not significantly different between the "accelerated" and

"not accelerated" datasets (permutation test $P = 0.316$). In short, there is no conflict between previous studies: the datasets used in Betancourt and Presgraves (BETANCOURT and PRESGRAVES 2002), and Zhang and Parsch (ZHANG and PARSCH 2005) were similar to our "accelerated" dataset (high fractions of male reproductive genes) and the datasets used in Marais *et al.* (MARAIS *et al.* 2004) and Haddrill *et al.* (HADDRILL *et al.* 2007) were analogous to our "not accelerated" dataset where most genes evolve under purifying selection. On a genome-wide scale, at least in *Drosophila*, there is strong support for the theory that recombination enhances the efficacy of natural selection.

Intragenic interference is supported by patterns of selection at synonymous sites

Weak but pervasive selection on synonymous sites (AKASHI 1995), such as selection for translational efficiency and/or accuracy leading to codon bias, is expected to be especially vulnerable to interference (AKASHI 1995; BETANCOURT and PRESGRAVES 2002; COMERON and GUTHRIE 2005; HILL and ROBERTSON 1966; MCVEAN and CHARLESWORTH 2000). Thus, genes in which positive selection has fixed many amino acids are expected to be unable to simultaneously maintain high levels of codon bias (measured as FOP, the frequency of optimal codons ; BETANCOURT and PRESGRAVES 2002): genes in the "accelerated" dataset have significantly less codon bias than genes in the "not accelerated" dataset (median *D. melanogaster* $FOP_{(accelerated)}=0.480$; $FOP_{(not\ accelerated)}=0.524$; Mann-Whitney U $P=9.78\times10^{-14}$), even when only genes with similar expression levels are analyzed (data not shown). Furthermore, the central-most regions of long exons have less codon bias (COMERON and GUTHRIE 2005), and exon length correlates negatively with codon bias (Spearman's partial $\rho_{exon\ length} = -0.1193$ $P = 2\times10^{-4}$; Figure 1.4G; Appendix Table 1.4) suggesting that interference from nearby selected sites reduces the efficacy of selection at synonymous sites. Whether these patterns of interference are generated by

nearby nonsynonymous sites under stronger selection (LOEWE and CHARLESWORTH 2007) or interference between weakly selected synonymous sites (COMERON and GUTHRIE 2005) remains unresolved. Consistent with previous studies, recombination rate (COMERON *et al.* 1999; HEY and KLIMAN 2002; MARAIS and CHARLESWORTH 2003; MARAIS *et al.* 2001) and intron length (COMERON and KREITMAN 2002) both positively correlate with codon bias (Spearman's partial $\rho_{recombination} = 0.0244$, $P = 0.0252$; Spearman's partial $\rho_{intron\ length} = 0.136$ $P = 2x10^{-4}$; Figure 1.4G; Appendix Table 1.4), suggesting that decreasing linkage increases the efficacy of purifying selection at synonymous sites. Although alternative explanations have been proposed to explain the correlation between codon bias and recombination rate (MARAIS *et al.* 2001; SINGH *et al.* 2005), given the signature of interference in evolutionary rates at nonsynonymous sites, combined with the theoretical expectation that weak selection at synonymous sites should be particularly vulnerable to interference, it is likely that interference at least contributes to the pattern of evolution at synonymous sites.

Beyond protein divergence

In addition to protein divergence, changes in expression pattern have long been thought to have a central role in interspecific differentiation (KING and WILSON 1975; PRUD'HOMME *et al.* 2006, but see HOEKSTRA and COYNE 2007). The 12 *Drosophila* genome sequences facilitated the generation of expression data from species-specific arrays to *D. melanogaster, D. simulans, D. yakuba, D. ananassae, D. pseudoobscura, D. virilis* and *D. mojavensis* (ZHANG *et al.* 2007), which allowed us to estimate rates of expression divergence among the *melanogaster* group species in a phylogenetic framework (see online Materials and Methods).

Expression level (Spearman's partial $\rho_{expression\ level} = 0.041$, $P = 0.0012$; Appendix Table 1.5) and tissue bias (Spearman's partial $\rho_{\tau} = 0.043$, $P = 2x10^{-4}$; Appendix Table 1.5) both significantly positively correlate with expression

divergence, as do intron number and length (Spearman's partial $\rho_{\text{intron length}} = 0.0284$, $P = 0.0176$; Spearman's partial $\rho_{\text{intron number}} = 0.0265$, $P = 0.0234$; Appendix Table 1.5); protein length negatively correlates with expression divergence (Spearman's partial $\rho_{\text{protein length}} = -0.0914$, $P = 2 \times 10^{-4}$; Appendix Table 1.5). However, expression divergence and the probability of positive selection appear unrelated (Logistic regression $\beta = 0.05$, P=0.770), and rate of protein divergence across the phylogeny is only weakly positive correlated with rate of expression divergence (Spearman's partial $\rho_{dN} = 0.0254$, $P = 0.0304$; Spearman's partial $\rho_{\omega} = 0.0353$ $P = 0.0016$; Appendix Table 1.5).

These results contrast with previous reports in *Drosophila,* which suggested moderate to strong correlations between expression divergence and rates of protein divergence (LEMOS *et al.* 2005; NUZHDIN *et al.* 2004). However, these studies have limited sample sizes and rely on single-species arrays for multiple species hybridizations, subjecting them to hybridization mismatch errors. Our results suggest that overall, genes that are rapidly diverging at the protein sequence level are not rapidly diverging in expression level and *vice versa*, although genes that diverge in pattern of expression (*i.e.*, degree of sex-bias) might have a different pattern (ZHANG *et al.* 2007). Although some genes must experience positive selection for both changes in gene expression and changes in protein sequence, there is as yet no clear evidence that this is a general pattern.

Concluding Remarks

Although several factors that correlate with the observed diversity of evolutionary rates among proteins have been identified, the evolutionary mechanisms through which they influence rates of protein divergence are poorly understood. Translational selection (as measured by gene expression) appears to be one of the most important determinates of evolutionary rate, although tissue bias in expression can

have equally strong independent effect on evolutionary rates, alluding to important differences between yeast and *Drosophila*. The extent to which stronger purifying selection on ubiquitously expressed genes, excess positive selection among biased genes, or increased translational selection affecting genes with broad expression affects the relationship between tissue bias and protein evolution remains unresolved. Factors such as gene essentiality and intron number impose additional constraints and stronger purifying selection on genes, but neither factor affects the rate of adaptive evolution.

An advantage of the *Drosophila* dataset is that we can separate genes based on the type of selection they experience. This is of particular importance because recombination rates affect the efficacy of both purifying and positive selection, influencing rates of protein evolution in opposite ways depending on the mode of selection acting on a gene. These weak but significant effects due to Hill-Robertson interference should not be overlooked.

Almost all studies of this kind have, by necessity, assumed that most parameters remain constant throughout the evolutionary time captured by measures of protein divergence. The use of species-specific expression microarrays for eight of the 12 sequenced species has already been shown that expression profiles can diverge over relatively short time scales (ZHANG *et al.* 2007). The next level of analysis will be to consider not only how proteins evolve, but how changes in genomic context, transcriptional properties, and physiological roles of proteins over evolutionary time affect protein evolution.

REFERENCES

AKASHI, H., 1995 Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. Genetics **139:** 1067-1076.

BACHTROG, D., 2003 Adaptation shapes patterns of genome evolution on sexual and asexual chromosomes in *Drosophila*. Nat Genet **34:** 215-219.

BETANCOURT, A. J., and D. C. PRESGRAVES, 2002 Linkage limits the power of natural selection in *Drosophila*. Proc Natl Acad Sci U S A **99:** 13616-13620.

BIELAWSKI, J. P., K. A. DUNN and Z. YANG, 2000 Rates of nucleotide substitution and mammalian nuclear gene evolution. Approximate and maximum-likelihood methods lead to different conclusions. Genetics **156:** 1299-1308.

BIERNE, N., and A. EYRE-WALKER, 2003 The problem of counting sites in the estimation of the synonymous and nonsynonymous substitution rates: implications for the correlation between the synonymous substitution rate and codon usage bias. Genetics **165:** 1587-1597.

BLENCOWE, B. J., 2000 Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. Trends Biochem Sci **25:** 106-110.

CARVALHO, A. B., and A. G. CLARK, 1999 Intron size and natural selection. Nature **401:** 344.

CHARLESWORTH, B., M. T. MORGAN and D. CHARLESWORTH, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289-1303.

CHINTAPALLI, V. R., J. WANG and J. A. DOW, 2007 Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. Nat Genet **39:** 715-720.

CLARK, N. L., and W. J. SWANSON, 2005 Pervasive adaptive evolution in primate seminal proteins. PLoS Genet **1:** e35.

COMERON, J. M., 2004 Selective and mutational patterns associated with gene expression in humans: influences on synonymous composition and intron presence. Genetics **167:** 1293-1304.

COMERON, J. M., and T. B. GUTHRIE, 2005 Intragenic Hill-Robertson interference influences selection intensity on synonymous mutations in *Drosophila*. Mol Biol Evol **22:** 2519-2530.

COMERON, J. M., and M. KREITMAN, 2000 The correlation between intron length and recombination in *Drosophila*. Dynamic equilibrium between mutational and selective forces. Genetics **156:** 1175-1190.

COMERON, J. M., and M. KREITMAN, 2002 Population, evolutionary and genomic consequences of interference selection. Genetics **161:** 389-410.

COMERON, J. M., M. KREITMAN and M. AGUADE, 1999 Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. Genetics **151:** 239-249.

DAVIS, J. C., O. BRANDMAN and D. A. PETROV, 2005 Protein evolution in the context of *Drosophila* development. J Mol Evol **60:** 774-785.

DROSOPHILA 12 GENOMES CONSORTIUM, 2007 Evolution of genes and genomes on the *Drosophila* phylogeny. Nature **450:** 203-218.

DRUMMOND, D. A., J. D. BLOOM, C. ADAMI, C. O. WILKE and F. H. ARNOLD, 2005 Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A **102:** 14338-14343.

DRUMMOND, D. A., A. RAVAL and C. O. WILKE, 2006 A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol **23:** 327-337.

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

DURET, L., and D. MOUCHIROUD, 2000 Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate. Mol Biol Evol **17:** 68-74.

ELLEGREN, H., and J. PARSCH, 2007 The evolution of sex-biased genes and sex-biased gene expression. Nat Rev Genet **8:** 689-698.

FELSENSTEIN, J., 1974 The evolutionary advantage of recombination. Genetics **78:** 737-756.

FELSENSTEIN, J., 1989 PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics **5:** 164-166.

FISHER, R. A., 1930 *The Genetical Theory of Natural Selection*. Oxford University Press, Oxford.

FRASER, H. B., A. E. HIRSH, L. M. STEINMETZ, C. SCHARFE and M. W. FELDMAN, 2002 Evolutionary rate in the protein interaction network. Science **296:** 750-752.

FRASER, H. B., D. P. WALL and A. E. HIRSH, 2003 A simple dependence between protein evolution rate and the number of protein-protein interactions. BMC Evol Biol **3:** 11.

GIAEVER, G., A. M. CHU, L. NI, C. CONNELLY, L. RILES *et al.*, 2002 Functional profiling of the *Saccharomyces cerevisiae* genome. Nature **418:** 387-391.

GIBSON, G., R. RILEY-BERGER, L. HARSHMAN, A. KOPP, S. VACHA *et al.*, 2004 Extensive sex-specific nonadditivity of gene expression in *Drosophila melanogaster*. Genetics **167:** 1791-1799.

GIOT, L., J. S. BADER, C. BROUWER, A. CHAUDHURI, B. KUANG *et al.*, 2003 A protein interaction map of *Drosophila melanogaster*. Science **302:** 1727-1736.

GOLDMAN, N., and Z. YANG, 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol **11:** 725-736.

GOOD, J. M., and M. W. NACHMAN, 2005 Rates of protein evolution are positively correlated with developmental timing of expression during mouse spermatogenesis. Mol Biol Evol **22:** 1044-1052.

HADDRILL, P. R., D. L. HALLIGAN, D. TOMARAS and B. CHARLESWORTH, 2007 Reduced efficacy of selection in regions of the *Drosophila* genome that lack crossing over. Genome Biol **8:** R18.

HAERTY, W., S. JAGADEESHAN, R. J. KULATHINAL, A. WONG, K. RAVI RAM *et al.*, 2007 Evolution in the fast lane: rapidly evolving sex-related genes in *Drosophila*. Genetics **177:** 1321-1335.

HAHN, M. W., G. C. CONANT and A. WAGNER, 2004 Molecular evolution in large genetic networks: does connectivity equal constraint? J Mol Evol **58:** 203-211.

HAMBLIN, M. T., and C. F. AQUADRO, 1996 High nucleotide sequence variation in a region of low recombination in *Drosophila simulans* is consistent with the background selection model. Molecular Biology and Evolution **13:** 1133-1140.

HAMBLIN, M. T., and C. F. AQUADRO, 1999 DNA sequence variation and the recombinational landscape in *Drosophila pseudoobscura*: a study of the second chromosome. Genetics **153:** 859-869.

HEY, J., and R. M. KLIMAN, 2002 Interactions between natural selection, recombination and gene density in the genes of *Drosophila*. Genetics **160:** 595-608.

HILL, W. G., and A. ROBERTSON, 1966 The effect of linkage on limits to artificial selection. Genet Res **8:** 269-294.

HIRSH, A. E., and H. B. FRASER, 2001 Protein dispensability and rate of evolution. Nature **411:** 1046-1049.

HOEKSTRA, H. E., and J. A. COYNE, 2007 The locus of evolution: evo devo and the genetics of adaptation. Evolution Int J Org Evolution **61:** 995-1016.

HURST, L. D., and N. G. SMITH, 1999 Do essential genes evolve slowly? Curr Biol **9:** 747-750.

INGVARSSON, P. K., 2007 Gene expression and protein length influence codon usage and rates of sequence evolution in Populus tremula. Mol Biol Evol **24:** 836-844.

JOLLIFFE, I. T., 1986 *Principal component analysis*. Springer, New York.

JORDAN, I. K., Y. I. WOLF and E. V. KOONIN, 2003 No simple dependence between protein evolution rate and the number of protein-protein interactions: only the most prolific interactors tend to evolve slowly. BMC Evol Biol **3:** 1.

KING, M. C., and A. C. WILSON, 1975 Evolution at two levels in humans and chimpanzees. Science **188:** 107-116.

LEMOS, B., B. R. BETTENCOURT, C. D. MEIKLEJOHN and D. L. HARTL, 2005 Evolution of proteins and gene expression levels are coupled in *Drosophila* and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. Mol Biol Evol **22:** 1345-1354.

LIAO, B. Y., N. M. SCOTT and J. ZHANG, 2006 Impacts of gene essentiality, expression pattern, and gene compactness on the evolutionary rate of mammalian proteins. Mol Biol Evol **23:** 2072-2080.

LIN, Y. S., W. L. HSU, J. K. HWANG and W. H. LI, 2007 Proportion of solvent-exposed amino acids in a protein and rate of protein evolution. Mol Biol Evol **24:** 1005-1011.

LOEWE, L., and B. CHARLESWORTH, 2007 Background selection in single genes may explain patterns of codon bias. Genetics **175:** 1381-1393.

MARAIS, G., and B. CHARLESWORTH, 2003 Genome evolution: recombination speeds up adaptive evolution. Curr Biol **13:** R68-70.

MARAIS, G., T. DOMAZET-LOSO, D. TAUTZ and B. CHARLESWORTH, 2004 Correlated evolution of synonymous and nonsynonymous sites in *Drosophila*. J Mol Evol **59:** 771-779.

MARAIS, G., D. MOUCHIROUD and L. DURET, 2001 Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc Natl Acad Sci U S A **98:** 5688-5692.

MARAIS, G., P. NOUVELLET, P. D. KEIGHTLEY and B. CHARLESWORTH, 2005 Intron size and exon evolution in *Drosophila*. Genetics **170:** 481-485.

MCVEAN, G. A., and B. CHARLESWORTH, 2000 The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics **155:** 929-944.

MORIYAMA, E. N., and J. R. POWELL, 1996 Intraspecific nuclear DNA variation in *Drosophila*. Mol Biol Evol **13:** 261-277.

MUSE, S. V., and B. S. GAUT, 1994 A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol **11:** 715-724.

NUZHDIN, S. V., M. L. WAYNE, K. L. HARMON and L. M. MCINTYRE, 2004 Common pattern of evolution of gene expression level and protein sequence in *Drosophila*. Mol Biol Evol **21:** 1308-1317.

ORTIZ-BARRIENTOS, D., A. S. CHANG and M. A. NOOR, 2006 A recombinational portrait of the *Drosophila pseudoobscura* genome. Genet Res **87:** 23-31.

PAL, C., B. PAPP and L. D. HURST, 2001 Highly expressed genes in yeast evolve slowly. Genetics **158:** 927-931.

PAL, C., B. PAPP and L. D. HURST, 2003 Genomic function: Rate of evolution and gene dispensability. Nature **421:** 496-497; discussion 497-498.

PARMLEY, J. L., A. O. URRUTIA, L. POTRZEBOWSKI, H. KAESSMANN and L. D. HURST, 2007 Splicing and the evolution of proteins in mammals. PLoS Biol **5:** e14.

PECK, J. R., 1994 A ruby in the rubbish: beneficial mutations, deleterious mutations and the evolution of sex. Genetics **137:** 597-606.

POLLARD, D. A., V. N. IYER, A. M. MOSES and M. B. EISEN, 2006 Widespread discordance of gene trees with species tree in Drosophila: evidence for incomplete lineage sorting. PLoS Genet **2:** e173.

POPESCU, C. E., T. BORZA, J. P. BIELAWSKI and R. W. LEE, 2006 Evolutionary rates and expression level in *Chlamydomonas*. Genetics **172:** 1567-1576.

PRESGRAVES, D. C., 2005 Recombination enhances protein adaptation in Drosophila melanogaster. Curr Biol **15:** 1651-1656.

PROSCHEL, M., Z. ZHANG and J. PARSCH, 2006 Widespread adaptive evolution of *Drosophila* genes with sex-biased expression. Genetics **174:** 893-900.

PRUD'HOMME, B., N. GOMPEL, A. ROKAS, V. A. KASSNER, T. M. WILLIAMS *et al.*, 2006 Repeated morphological evolution through *cis*-regulatory changes in a pleiotropic gene. Nature **440:** 1050-1053.

ROCHA, E. P., and A. DANCHIN, 2004 An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol Biol Evol **21:** 108-116.

SINGH, N. D., P. F. ARNDT and D. A. PETROV, 2005 Genomic heterogeneity of background substitutional patterns in *Drosophila melanogaster*. Genetics **169:** 709-722.

SINGH, N. D., A. M. LARRACUENTE and A. G. CLARK, 2008 Contrasting the efficacy of selection on the X and autosomes in *Drosophila*. Molecular Biology and Evolution **25:** 454-467.

SMITH, J. M., and J. HAIGH, 1974 The hitch-hiking effect of a favourable gene. Genet Res **23:** 23-35.

STOREY, J. D., and R. TIBSHIRANI, 2003 Statistical significance for genomewide studies. Proc Natl Acad Sci U S A **100:** 9440-9445.

SUBRAMANIAN, S., and S. KUMAR, 2004 Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. Genetics **168:** 373-381.

TEICHMANN, S. A., 2002 The constraints protein-protein interactions place on sequence divergence. J Mol Biol **324:** 399-407.

TRUE, J. R., J. M. MERCER and C. C. LAURIE, 1996 Differences in crossover frequency and distribution among three sibling species of *Drosophila*. Genetics **142:** 507-523.

WALL, D. P., A. E. HIRSH, H. B. FRASER, J. KUMM, G. GIAEVER *et al.*, 2005 Functional genomic analysis of the rates of protein evolution. Proc Natl Acad Sci U S A **102:** 5483-5488.

WEISBERG, S., 1985 *Applied linear regression*. John Wiley, New York.

WHITTAKER, J., 1990 *Graphical Models in Applied Multivariate Statistics*. John Wiley, Chichester.

WONG, A., J. D. JENSEN, J. E. POOL and C. F. AQUADRO, 2007 Phylogenetic incongruence in the *Drosophila melanogaster* species group. Mol Phylogenet Evol **43:** 1138-1150.

WRIGHT, S. I., C. B. YAU, M. LOOSELEY and B. C. MEYERS, 2004 Effects of gene expression on molecular evolution in Arabidopsis thaliana and Arabidopsis lyrata. Mol Biol Evol **21:** 1719-1726.

YANAI, I., H. BENJAMIN, M. SHMOISH, V. CHALIFA-CASPI, M. SHKLAR *et al.*, 2005 Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics **21:** 650-659.

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

YANG, Z., 1998 Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol **15:** 568-573.

YANG, Z., 2002 Inference of selection from multiple species alignments. Curr Opin Genet Dev **12:** 688-694.

YANG, Z., and R. NIELSEN, 2002 Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol **19:** 908-917.

YANG, Z., R. NIELSEN, N. GOLDMAN and A. M. PEDERSEN, 2000 Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155:** 431-449.

YANG, Z., W. S. WONG and R. NIELSEN, 2005 Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol **22:** 1107-1118.

ZHANG, J., and X. HE, 2005 Significant impact of protein dispensability on the instantaneous rate of protein evolution. Mol Biol Evol **22:** 1147-1155.

ZHANG, L., and W. H. LI, 2004 Mammalian housekeeping genes evolve more slowly than tissue-specific genes. Mol Biol Evol **21:** 236-239.

ZHANG, Y., D. STURGILL, M. PARISI, S. KUMAR and B. OLIVER, 2007 Constraint and turnover in sex-biased gene expression in the genus *Drosophila*. Nature **450:** 233-237.

ZHANG, Z., and J. PARSCH, 2005 Positive correlation between evolutionary rate and recombination rate in *Drosophila* genes with male-biased expression. Mol Biol Evol **22:** 1945-1947.

ZUCKERKANDL, E., PAULING, L., 1965 *Evolutionary divergence and convergence in proteins*. Academic Press, New York.

CHAPTER 2[2]

CONTRASTING THE EFFICACY OF SELECTION ON THE X AND AUTOSOMES IN *DROSOPHILA*

Introduction

The efficacy of natural selection depends on the strength of selection, allele dominance and the effective population size. Selection at one site can affect selection at linked sites, especially in regions of low recombination resulting in a decreased efficacy of selection (HILL and ROBERTSON 1966). In addition to varying across different species, the efficacy of selection can vary within a species genome due to effects of chromosomal location and local recombination rate. *Drosophila* males are hemizygous for the X chromosome, making new mutations immediately visible to selection in male. It is hypothesized that this increased visibility should lead to an increase in the efficacy of natural selection on the X chromosome relative to the autosomes. However, since there are only three X chromosomes for every four autosomes, the reduced effective size of the X chromosome, assuming equal effective numbers of breeding males and females, implies that weakly selected variants on the X may have their dynamics mediated to a greater degree by genetic drift. The reduced efficacy of selection expected from the smaller effective size of the X could counter the increase in efficacy of selection cause by hemizygosity of the X chromosome in males.

If positive selection is more efficacious on the X chromosome, it is expected that there will be increased rates of substitutions resulting from the fixation of

---

[2] This chapter is a modified version of a paper published in *Molecular Biology and Evolution* (Singh, N.D., A.M. Larracuente, A.G. Clark. 2008. Contrasting the efficacy of selection on the X and autosomes in *Drosophila*. *Mol. Biol. Evol.* **2**:454-467.) and is reprinted with permission. N.D.S. contributed significantly to the text that appeared in the publication, estimated d_{S4} and performed the paired analyses and A.G.C. contributed to editing the text in the publication. A.M.L. and N.D.S were first co-authors of this publication.

beneficial mutations. This expectation relies on several assumptions and under certain conditions. Table 2.1 presents theoretical ratios of substitution rates of the X chromosome to the autosomes in a single-locus model with selection coefficients s_f and s_m in females and males respectively, mutation rates μ_f and μ_m in males and females, respectively, and dominance parameter h (CHARLESWORTH *et al.* 1987; VICOSO and CHARLESWORTH 2006). Natural selection is expected to be more efficient if new mutations are on average at least partially recessive (defined here as $0 < h < 0.5$). If this condition is met, then rates of adaptive evolution on the X chromosome should exceed those on the autosomes (traditionally referred to as the "faster-X" hypothesis; AVERY 1984; CHARLESWORTH *et al.* 1987); this inequality holds for both small and large coefficients of selection (BETANCOURT *et al.* 2004).

Table 2.1: Expected ratio of substitution rates on the X and the autosomes under different parameter combinations assuming equal numbers of effective males and females. †k is the ratio of fitness effects on females and males.
*When the assumption of equal numbers of breeding males and females is relaxed, the ratio of breeding males to breeding females also plays a role in the expected relative substitution rates on the X and the autosomes (See Appendix 2 and Appendix Table 2.1). These results are presented in Appendix Figure 2.4.

Male and Female Mutation Rate (μ_m and μ_f)	Sex-specific Selective Effects (s_m and s_f)	Dominance coefficient (h)	Expected Ratio of X/A Substitution rates
$\mu_m = \mu_f$	$s_m = s_f; s_m, s_f > 0$	$0 < h < 0.5$	> 1*
$\mu_m = \mu_f$	$s_m = s_f; s_m, s_f > 0$	$h = 0.5$	1*
$\mu_m = \mu_f$	$s_m = s_f; s_m, s_f > 0$	$0.5 < h < 1$	< 1*
$\mu_m = \mu_f$	$s_f > 0; s_m = -ks_f$ †	$h < k/2$	< 1
$\mu_m = \mu_f$	$s_f > 0; s_m = -ks_f$ †	$h > k/2$	≥ 1
$\mu_m = \mu_f$	$s_m > 0, s_f = -ks_m$ †	$0.5 < h < 1$	< 1
$\mu_m = \mu_f$	$s_m > 0, s_f = -ks_m$ †	$0 < h < 0.5$	> 1
$\mu_m = \mu_f$	$s_m = s_f; s_m, s_f < 0$	$0.5 < h < 1$	> 1
$\mu_m = \mu_f$	$s_m = s_f; s_m, s_f < 0$	$0 < h < 0.5$	< 1
$\mu_m > \mu_f$	N/A	N/A	< 1
$\mu_m < \mu_f$	N/A	N/A	> 1

It is important to note that the above predictions are based on the assumptions of equal numbers of breeding males and females, equal mean and variance in reproductive success for the sexes (*i.e.* no segregating variation in fitness apart from newly arisen mutations), and identical distributions of selection and dominance coefficients acting on new mutations. If selection acts on standing variation rather than new mutations, then it is expected that rates of adaptive evolution on the autosomes will exceed those on the X chromosome (CHARLESWORTH *et al.* 1987). If in the case where segregating mutations that were previously deleterious become beneficial, this inequality holds for all coefficients of dominance (Orr and Betancourt 2001). If the selective effects of mutations differ between males and females or if there is evidence of sexual antagonism, where there are opposing selective pressures in the two sexes, then alleles that favor males at the expense of females enjoy an evolutionary advantage when they are autosomal rather than X-linked (Rice 1984), and partially recessive alleles that favor females at the expense of males show the opposite tendency (CHARLESWORTH *et al.* 1987; Table 2.1).

The X is also expected to have an increased efficacy of purifying selection under the same conditions that are expected to lead to an increase in the efficiency of positive selection (Table 2.1). This will have the opposite effect on rates of substitution. More efficient purifying selection is expected to reduce the substitution rate as new, at least partially recessive deleterious mutations are purged from the population (CHARLESWORTH *et al.* 1987), which would lead to decreased rates of substitution on the X chromosome. Weak selection is expected to be especially vulnerable to this effect. Codon bias is a form of weak but pervasive selection on synonymous sites and is a consequence of selection-drift-mutation balance (AKASHI 1997; BULMER 1991; MCVEAN and CHARLESWORTH 1999; SHARP and LI 1986). The

X chromosome should have higher levels of codon bias than the autosomes because of an increase in the efficacy of purifying selection.

The recent sequencing of the eukaryotic genomes of ten additional species of *Drosophila* brings the total number of sequenced species to twelve. This dataset provides an opportunity to revisit the question of whether the X chromosome has an increased efficacy of selection in the context of the *Drosophila* phylogeny on a genome-wide scale. We examined potential differences in the efficacy of positive selection between the X and the autosomes by comparing substitution rates of X-linked and autosomal genes. We took advantage of a set of genes that were predicted to evolve under positive selection across the phylogeny (*DROSOPHILA* 12 GENOMES CONSORTIUM 2007; LARRACUENTE *et al.* 2008) and on individual branches of the phylogeny. This affords us the opportunity to test for increased efficacy of positive selection on the X chromosome in subsets of rapidly evolving genes.

Our results provide strong support for an increased efficacy of purifying selection on the X chromosome across the *Drosophila* phylogeny. However, there does not appear to be a strong signal of an increased efficacy of positive selection on the X chromosome in these species, as rates of substitution are not systematically increased on this chromosome. The results are sensitive to the metric of substitution employed in the comparison, and vary considerably among species. We suggest that while positive selection may be more efficacious on the X chromosome, adaptive evolution from new mutations is not sufficiently pervasive relative to the amount of purifying selection to systematically inflate substitution rates on the X chromosome in *Drosophila*.

Materials and Methods

Coding sequence alignments

Two sets of alignments were used for this analysis, both of which are based on the masked alignments as described elsewhere.(*DROSOPHILA* 12 GENOMES CONSORTIUM 2007). The first set includes 8510 genes with a single ortholog in the six species in the *melanogaster* group, the largest clade without saturation at synonymous sites. This set was used for inferences of ω, d_N, and d_S in the *melanogaster* group (described below). The second set includes 6698 genes with a single ortholog in all 12 fully sequenced *Drosophila* genomes. This set was used for inferences of amino acid divergence and codon bias across the *Drosophila* phylogeny (described below).

Evolutionary analysis

Estimates of ω, d_N, and d_S were obtained for each of the 8510 alignments in the *melanogaster* group from branch models run in PAML (version 3.1; Yang 1998). The codon substitution model used here makes a number of assumptions whose validity we discuss throughout the paper. An assumption in this model is that there is no heterogeneity among sites in selection pressure (*i.e.* ω is constant across sites). These models allowed us to obtain branch-specific estimates of evolutionary rate parameters. Five branch tests were used, where for each test, one terminal lineage of the *melanogaster* subgroup was allowed to have a different ω than the rest of the *melanogaster* group phylogeny. Note that *D. ananassae* was included in the phylogeny but given the near-saturation at synonymous sites, we did not use branch models for this lineage.

We used two methods to estimate rates of adaptation on a particular branch of the phylogeny. Our first method was to assess the rate of evolution of each gene on a terminal branch relative to the rest of the phylogeny and identify genes that show a

significant relative acceleration in evolutionary rate on that branch using the branch-specific codon substitution models described above. To test for significant differences in ω between each terminal lineage and the rest of the tree, we performed likelihood ratio tests (LRT) assuming that the LRT statistic follows a χ^2 distribution. Inspection of the distribution of the LRT statistic revealed that this assumption is appropriate, as it conforms well to the χ^2 distribution. A significant P-value for this test coupled with the observation that the terminal branch ω exceeds the estimate of ω for the rest of the phylogeny indicates a branch-specific acceleration for a gene. Our second method was to test for positive selection across the entire phylogeny. The test for positive selection across the phylogeny compares models that allow ω to vary among sites (M7 and M8) and identifies genes that show support for a class of codons within the gene with $\omega > 1$. The test for positive selection was done by comparing models M7 and M8 and using simulations to generate a null distribution of likelihood ratio test statistics to generate P-values for this test (for a full description of methods see *DROSOPHILA* 12 GENOMES CONSORTIUM 2007 or LARRACUENTE *et al.* 2008). Any genes that had a significant deceleration on a particular lineage from the branch models were removed from that species' analysis. All PAML results (including P-values) are available for download at FlyBase (ftp://ftp.flybase.net/12_species_analysis).

We also estimated amino acid divergence for orthologous sequences in the set of 6698 alignments for all twelve sequenced *Drosophila* genomes. We used a model implemented in the CODEML package in PAML that translates codons to amino acids to estimate amino acid divergence for each branch of the phylogeny. We report the terminal branch lengths for individual species with two exceptions. Because of the short terminal branch lengths for *D. persimilis* and *D. pseudoobscura*, the amino acid divergence on the shared branch immediately preceding the split of these two lineages

was added to the terminal amino acid divergence of each of these two species. Thus, the intragenomic comparisons of rate of evolution are expected to be similar for *D. persimilis* and *D. pseudoobscura* given our methodology.

Estimates of amino acid divergence for clades were obtained by summing relevant internal and external branches for only those genes whose Muller element locations had been conserved. For instance, for the *melanogaster* complex clade, we included each of the three terminal lineages in addition to the shared *sechellia*/*simulans* lineage. Only the 1878 genes for which the tree topology with *D. yakuba* and *D. erecta* as sister species had highest support were included in these clade-specific analyses. This was out of necessity, as these species do not form a clade in the other tree topologies. This approach does limit the impact of phylogenetic incongruence across the genome, as it focuses the analysis on those genes that do not appear to bear strong signatures of lineage-sorting. However, one issue we cannot account for is phylogenetic incongruence within the context of a certain gene. While this can pose challenges for phylogenetic inference (WONG *et al.* 2007), we believe that this issue will introduce noise rather than a systematic biases in our results. For the analyses of pairs of orthologs for which in one species the pair is X-linked and the other species the pair is autosomal, we compared the relative amino acid divergence. The relative amino acid divergence was calculated for each gene as the amino acid divergence for the clade-specific branch for that gene normalized by the mean amino acid divergence across genes for that clade.

To estimate divergence at four-fold degenerate sites, we extracted the four-fold degenerate sites from the alignments of coding sequences in the *melanogaster* group and used BASEML with an unrooted tree to estimate terminal branch lengths. We restricted ourselves to those genes with at least 100 four-fold degenerate sites; in total we have divergence estimates for 4712 genes. For clade-specific analyses on

divergence at four-fold degenerate sites, only the 3068 genes for which the tree topology with *D. yakuba* and *D. erecta* as sister species had highest support were included, and genes whose Muller element locations were not conserved across species were not considered.

Lineage-specific evolutionary parameter estimates were based on the tree topology with the highest likelihood. For the clade-specific analysis, we restricted the analysis to the 4925 genes for which the tree topology with *D. yakuba* and *D. erecta* as sister species had highest support, and we further restricted ourselves to the subset of genes whose Muller element locations have been conserved across species. For each gene, the clade-specific branch lengths were obtained by summing relevant internal and external branches of the phylogeny (see above).

Estimates of ω for the four-species comparisons were taken from PAML model M0 based on two-species alignments of orthologous sequences in *D. melanogaster*/*D. simulans* and *D. pseudoobscura*/*D. persimilis* species pairs. We generated those alignments by extracting the appropriate sequences from the multi-species alignments in the dataset of 6698 genes with single orthologs in all twelve genomes. We limited this analysis to the subset of genes for which Muller element locations had been conserved in all four species.

Statistics and multiple test correction

All reported P-values are based on two-tailed tests unless specifically stated otherwise. Given the number of statistical comparisons performed, we employed several different corrections for multiple testing. For the M7 versus M8 comparison, we controlled the false discovery rate (FDR) by estimating *q*-values (STOREY AND TIBSHIRANI 2003) using the qvalue package in R (described in *DROSOPHILA* 12 GENOMES CONSORTIUM 2007 and LARRACUENTE *et al.* 2008). Unless otherwise stated,

we used a FDR threshold of 0.1, implying that the set of genes satisfying this criterion is expected to include 10% false positives. For the other statistical comparisons requiring correction for multiple tests, we used Holm's method for sequential Bonferroni correction (HOLM 1979); these P-values are referred to as "adjusted" P-values throughout the text.

Genic features

We estimated the degree of codon bias for all genes in the set of alignments based on orthologous sequences in all twelve genomes. We used a standalone implementation of codonW (downloaded from http://codonw.sourceforge.net), and used the codons defined as preferred in *D. melanogaster* to estimate the frequency of optimal codons (FOP) for each gene in each species. The application of preferred codon definitions from *D. melanogaster* to the remaining species in the genus is not likely to adversely affect our results, as codon preferences appear highly conserved across the phylogeny (*DROSOPHILA* 12 GENOMES CONSORTIUM 2007; VICARIO *et al.* 2007) .

Details of the data used to measure the breadth of expression are found in Larracuente et al. (2008). Briefly, we obtained tissue-specific expression data for seven adult tissues (brain, midgut, hindgut, Malphigian tubule, testis, ovary, accessory gland) from FlyAtlas (www.flyatlas.org) (see CHINTAPALLI *et al.* 2007; WANG *et al.* 2004). Expression had been assayed on Affymetrix Dros2 microarrays with four independent replicates for each tissue. With the exception of the testis, ovary and accessory gland, these expression estimates are from tissues dissected from equal numbers of males and females. Specificity of expression was measured by

$\tau = \sum_{j=1}^{n} 1 - \left(\frac{\log(S_j)}{\log(S_{\max})} \right) / (n-1)$ with S representing the signal intensity and n representing the number of tissues (YANAI *et al.* 2005). The $\log(S_j)$ was set to 0 for any gene detected on 0 or 1 out of 4 arrays for a given tissue. To limit our analysis to genes with no evidence of male-specific expression patterns, genes with $\tau \geq 0.9$ and expressed in the testes or accessory glands were removed.

Chromosomal locations of every gene in each species were based on scaffold-to-Muller-element maps kindly provided by AJ Bhutkar based on methodology described elsewhere (BHUTKAR *et al.* 2006; BHUTKAR *et al.* 2008). Only genes whose locations could be unambiguously mapped to a particular Muller element in the species under study were included in this analysis. We will refer to Muller element A as the 'X' chromosome in all twelve species, and in *D. willistoni*, *D. persimilis* and *D. pseudoobscura,* Muller element D is referred to as the 'neo-X.'

Results and Discussion

Because of the hemizygosity of the X chromosome in *Drosophila* males, the efficacy of both positive and purifying selection is expected to be greater than that of autosomes provided that new mutations are at least partially recessive (Table 2.1). Under this model, substitution rates between the X and the autosomes should differ. Theory predicts differences in the efficacy of selection on the X chromosome and the autosomes, which has led to several testable hypotheses (for review see (VICOSO AND CHARLESWORTH 2006). Here we explore differences in the efficacy of natural selection between the X and the autosomes in the context of the 12 species *Drosophila* phylogeny by comparing rates and patterns of molecular evolution on the X chromosome and the autosomes.

Neutral evolution

The prediction that the X chromosome should evolve faster than the autosomes relies on a number of assumptions, namely that there are equal effective numbers of breeding males and females, that adaptive mutations are new, that the adaptive mutation rate on the X chromosome is at least that of the autosomes, and that adaptive mutations are at least partially recessive (AVERY 1984; BETANCOURT *et al.* 2004; CHARLESWORTH *et al.* 1987). To assess the validity of these assumptions, we compared rates of substitution at potentially neutral sites, as selection is not the only cause for differences in the rates of substitution on the X and autosomes. Differences in the male or female mutation rate can lead to higher substitution rates on the autosomes of X chromosomes, respectively, for neutral sites (Table 2.1). In mammals, males have a higher mutation rate because the male germline has more mitotic divisions than in females (DROST AND LEE 1995), leading to lower divergence on the X chromosome relative to autosomes at neutral sites. However evidence in *Drosophila* suggests that male and females germlines tend to have similar number of mitotic divisions (DROST AND LEE 1995), but it is still possible that males and females have different mutation rates.

Differences in the numbers of breeding males and females or a difference in the variance in male and female reproductive success will impact the neutral substitution rates on the X chromosome and autosomes. If the effective number of breeding females increases relative to the effective number breeding males, then the ratio of effective sizes of the X and autosomes increases from the expected ¾ ratio under an assumption of equal numbers of effective males and females, and can approach or even exceed unity (HARTL AND CLARK 2007). Polymorphism data suggests that ratio of the effective sizes of the X and autosomes deviates from the

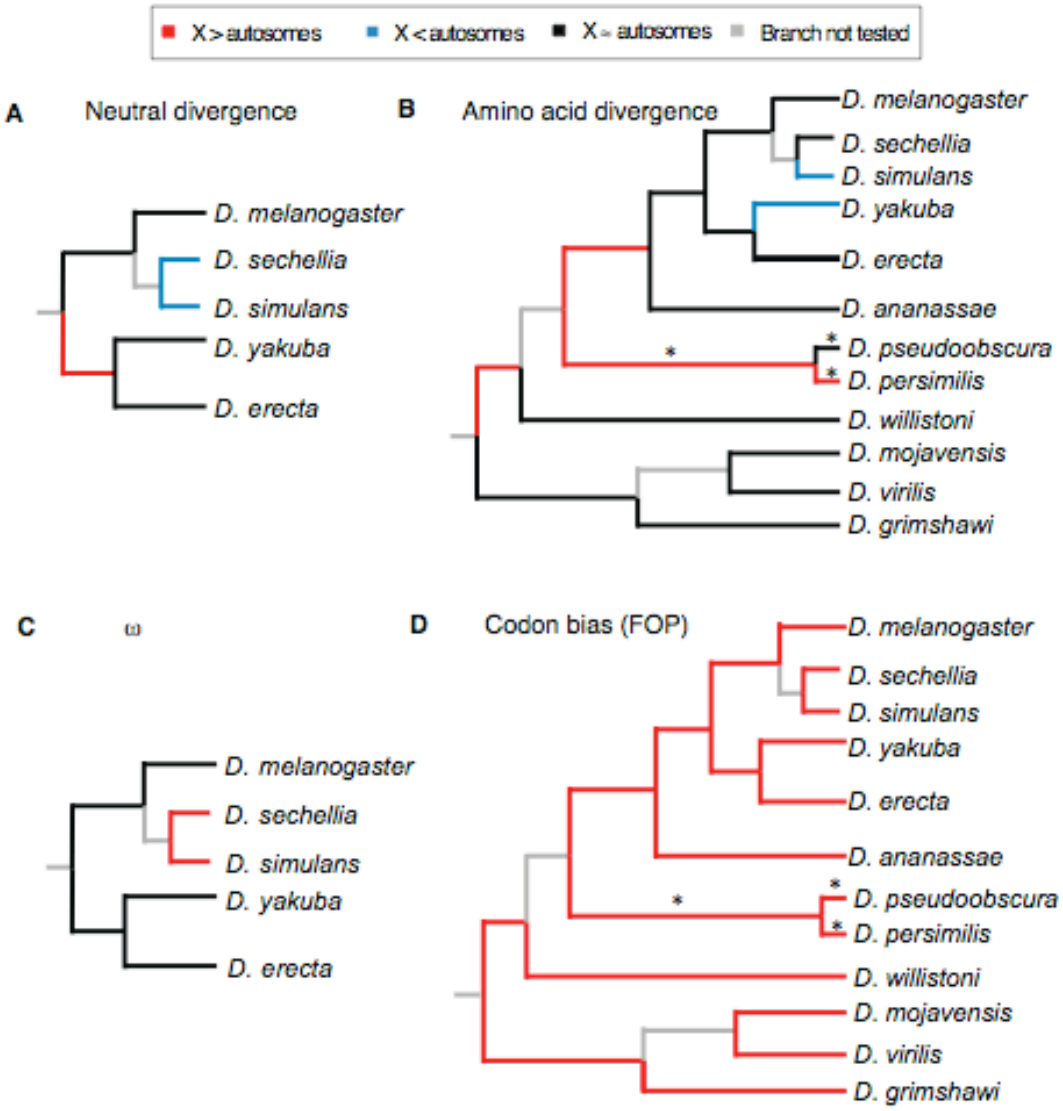
expected ¾ in several populations of *D. melanogaster* (HUTTER *et al.* 2007; SINGH *et al.* 2007b), though the direction and magnitude of the deviation vary across populations. An increase in effective size of the X will decrease the fixation probability of nearly neutral, slightly deleterious mutations and should reduce the overall substitution rate (KIMURA 1983). Furthermore, demographic events such as population bottlenecks affect the X chromosome and autosomes differently: under a model with equal numbers of effective males and females, a population bottleneck has a more severe effect on the X chromosome than on the autosomes (POOL and NIELSEN 2007; WALL *et al.* 2002). Therefore, the substitution rates on the X and autosomes are affected by differences in mutation rates between males and females, the effective number of breeding males and females, and the demographic history of the population.

We compared rates of evolution at four-fold degenerate synonymous sites in the *melanogaster* subgroup to assess whether differences in male and female mutation rates or demographic history on neutral substitutional patterns. These results are presented in Figure 2.1A and Table 2.2 (see also Appendix Figure 2.1).

Table 2.2: Median (mean) divergence at four-fold synonymous sites (d_{S4}), d_N and ω for the five *melanogaster* subgroup species. [a] The means of the distributions of the ωs per branch are not interpretable and therefore only the medians are listed.
[b] There are few enough F-linked genes in each species (out of all genes with a single ortholog in the *melanogaster* group: dmel=39, dsec=39, dsim=35, dyak=39, dere=39) that removing these genes does not affect the summary statistics for the autosomes pooled together.

	X			Autosomes[b]			F		
	d_{S4}	d_N	ω^a	d_{S4}	d_N	ω^a	d_{S4}	d_N	ω^a
D. melanogaster	0.054 (0.059)	0.0048 (0.0078)	0.061	0.053 (0.056)	0.0045 (0.0068)	0.065	0.039 (0.041)	0.0086 (0.0088)	0.19
D. sechellia	0.015 (0.017)	0.0023 (0.0034)	0.10	0.021 (0.023)	0.0024 (0.0035)	0.088	0.0059 (0.0061)	0.0018 (0.0018)	0.23
D. simulans	0.012 (0.014)	0.0013 (0.0029)	0.076	0.017 (0.019)	0.0014 (0.0026)	0.059	0.0052 (0.0065)	0.0014 (0.0033)	0.21
D. yakuba	0.069 (0.072)	0.0072 (0.012)	0.070	0.071 (0.073)	0.0070 (0.011)	0.072	0.055 (0.057)	0.016 (0.016)	0.18
D. erecta	0.066 (0.068)	0.0086 (0.014)	0.091	0.063 (0.064)	0.0078 (0.012)	0.092	0.058 (0.060)	0.013 (0.014)	0.19

Figure 2.1: Comparison of median X-linked and autosomal (A) divergence at fourfold degenerate sites for the *melanogaster* subgroup phylogeny, (B) amino acid divergence for the complete phylogeny, (C) ω for the *melanogaster* group phylogeny and (D) the frequency of optimal codons (FOP) for the complete phylogeny. Branches with statistical evidence in support of increases or decreases for X-linked genes are shaded in red and blue, respectively. Branches where there appeared to be no statistically significant differences between the X and the autosomes are in black. Results for both lineage-specific (terminal branches) and clade-specific analysis (internal branches, where possible) are included. For the clade-specific analysis, the internal branch leading to the ancestor of the clade is shaded (to distinguish these results from those lineage-specific analysis), but it is important to note that this branch is not included in the analysis (with the exception of the shared *pseudoobscura/persimilis* lineage; see Materials and Methods). Only branches descendant to the ancestor of a given clade are considered in the clade-specific analyses. In addition to the ancestral X, *D. willistoni*, *D. pseudoobscura* and *D. persimilis* also have a neo-X chromosome; asterisks denote species for which the metric of interest is significantly higher on the neo-X chromosome than on the autosomes.


X > autosomes
X < autosomes
X ≈ autosomes
Branch not tested
A
Neutral divergence
D. melanogaster
D. sechellia
D. simulans
D. yakuba
D. erecta
B
Amino acid divergence
D. melanogaster
D. sechellia
D. simulans
D. yakuba
D. erecta
D. ananassae
D. pseudoobscura
D. persimilis
D. willistoni
D. mojavensis
D. virilis
D. grimshawi
C
ω
D. melanogaster
D. sechellia
D. simulans
D. yakuba
D. erecta
D
Codon bias (FOP)
D. melanogaster
D. sechellia
D. simulans
D. yakuba
D. erecta
D. ananassae
D. pseudoobscura
D. persimilis
D. willistoni
D. mojavensis
D. virilis
D. grimshawi

It is important to note that these synonymous sites are likely subject to weak selection on codon bias, and are thus not truly "neutral" (*e.g.* AKASHI 1994; POWELL and MORIYAMA 1997; SINGH *et al.* 2007a). Despite this caveat, we expect that divergence at fourfold degenerate sites can shed light onto putative mutation differences between the sexes because within the context of protein-coding sequences, synonymous sites are less constrained than nonsynonymous sites and because selection on codon bias is weak. While we acknowledge that selection may affect divergence at fourfold degenerate sites to some degree, we nonetheless expect genetic drift to play a large role the evolution of fourfold degenerate sites. Therefore, we will refer to these fourfold-degenerate synonymous sites as "neutral" throughout the remainder of this discussion, consistent with the terminology of the nearly-neutral model.

It has been suggested that the neutral substitution rates on the X chromosome and autosomes in *D. melanogaster* and *D. simulans* are very similar (BAUER DUMONT AND AQUADRO 1997). Consistent with these observations, we do not find significant differences between the distributions of divergence at X-linked and the pooled autosomal loci in *D. melanogaster* (median divergence = 0.054 and 0.053 for the X and autosomes, respectively; $P = 0.12$, Mann-Whitney U-test, MWU), *D. yakuba* (median divergence = 0.069 and 0.071 for the X and autosomes, respectively; $P = 0.12$, MWU), or in *D. erecta* (median divergence = 0.066 and 0.063 for the X and autosomes, respectively; $P = 0.13$, MWU). However, in *D. sechellia* and *D. simulans*, divergence is significantly higher at the pooled autosomal loci (0.021 and 0.017, respectively) than at X-linked loci (0.015 and 0.012, respectively; $P << 0.0001$, MWU, both comparisons). We see similar results when we examine only the genes with the least amount of codon bias (lowest quartile; data not shown). This supports the idea that fourfold-degenerate synonymous sites can be used as a proxy for neutrality. When we compare rates of divergence on the X chromosome to individual

chromosome arms in *D. sechellia* and *D. simulans*, we find that the X-linked divergence is significantly lower than on every individual Muller element in both species after correction for multiple tests (adjusted $P < 0.022$, MWU, all comparisons). However, while the individual *D. yakuba* and *D. erecta* lineages show no differences in neutral rates of evolution between X-linked and autosomal genes, clade-specific analysis reveals a mild yet significant increase in neutral substitution rates on X chromosome (median divergence = 0.170 and 0.162 for the X and autosomes, respectively; $P = 0.03$, MWU). However, this result should be interpreted with caution because the clade-specific analysis is based only on the subset of genes whose most well-supported tree topology is that in which *D. yakuba* and *D. erecta* are sister species (approximately 50% of the dataset).

Thus the pattern of reduced X divergence at neutral sites in *D. sechellia* and *D. simulans* is consistent with elevated male mutation rates. It is possible that the reduced divergence at fourfold degenerate synonymous sites on the X chromosome is partially due to greater selective constraint at synonymous sites given the increased codon bias of X-linked genes in *Drosophila* (COMERON *et al.* 1999; HAMBUCH and PARSCH 2005; SINGH *et al.* 2005). It is unclear why these results would be seen in just these two lineages. It is important to note that for closely related species pairs such as *D. simulans* and *D. sechellia,* and *D. pseudoobscra* and *D. persimilis* we are conflating polymorphism and divergence. Therefore, sites that are polymorphic in one or both species can be erroneously counted as fixed between the species, thus inflating divergence, especially at autosomal loci because X-linked polymorphism is lower *D. simulans* (BEGUN AND WHITLEY 2000). It cannot be determined in this study whether these differences in rates of divergence at four-fold synonymous sites between species are due to differences in mutation patterns or X-linked and autosomal polymorphism,

demographic history, patterns of weak selection at synonymous sites or breeding structure.

Adaptive evolution

The X chromosome is expected to have an increase in the efficacy of positive selection relative to the autosomes for new mutations that are at least partially recessive, on average. This increased efficacy of positive selection should lead to increased rates of adaptive substitution on the X chromosome, assuming equal numbers of effective males and females (AVERY 1984; BETANCOURT *et al.* 2004; CHARLESWORTH *et al.* 1987). It appears as though deleterious mutations are partially recessive in *Drosophila* (for review see GARCIA-DORADO *et al.* 2004). However, there is little empirical evidence on the distribution of dominance effects of adaptive mutations. Most inferences of the recessivity of beneficial mutations are based on comparing patterns of polymorphism and divergence between X-linked and autosomal genes (BEGUN and WHITLEY 2000; LU and WU 2005; SCHOEFL and SCHLOETTERER 2004). For the purposes of this paper, we assume that new beneficial mutations are mostly recessive.

Several recent studies suggest that a substantial fraction of the *Drosophila* genome is subject to positive selection (BIERNE and EYRE WALKER 2004; *DROSOPHILA* 12 GENOMES CONSORTIUM 2007; SAWYER *et al.* 2003; SAWYER *et al.* 2007; WELCH 2006), therefore we expect to observe any increase in the efficacy of positive selection on the X chromosome to be reflected in the substitution rate at nonsynonymous sites. Empirical studies have yielded inconsistent results with regard to the test of this hypothesis: Musters et al. (2006) found evidence of higher substitution rates on the X chromosome while Betancourt et al. (2002) did not. Furthermore, comparisons of pairs orthologs between species or duplicate genes where one gene is X-linked and the other is autosomal also yield contradictory results: some studies revealing faster

evolutionary rates for X-linked paralogs/orthologs (COUNTERMAN *et al.* 2004; THORNTON and LONG 2002), and another did not (THORNTON *et al.* 2006). It is possible that the differences in results arise from sampling differences.

We took advantage of a large set of single-copy orthologs in the *melanogaster* group and single-copy orthologs in the whole *Drosophila* phylogeny to examine differences in the efficacy of selection between the X chromosome and the autosomes. To do this, we used several different measure of the rate of protein evolution: amino acid divergence for all 12 species and ω (the ratio of nonsynonymous to synonymous) for the *melanogaster* group. We use both paired and unpaired comparisons to compare the efficacy of positive selection on the X chromosome with the autosomes. We have limited our study to just those genes with identifiable orthologs in either the entire phylogeny or within the *melanogaster* group. Only approximately 1/2 of the genes annotated in *D. melanogaster* are contained within dataset of orthologs across the whole phylogeny. These genes are biased towards those genes that are evolving sufficiently slowly such that orthologs can be readily identified. Therefore, this dataset is missing some of the most rapidly evolving genes. It is difficult to assess the magnitude of this bias, but assuming that the genes not captured by our analysis evolve similarly to the genes included in our analysis, this ascertainment bias makes our analysis regarding positive selection conservative.

Amino acid divergence

We compared rates of amino acid divergence of X-linked and autosomal genes within each of the twelve *Drosophila* genomes. For most species, the X chromosome and the autosomes evolve at similar rates (Figure 2.1B, Table 2.3, Appendix Figure 2.2).

Table 2.3. Median (mean) amino acid divergence (a.a. div) and FOP for the X chromosome, autosomes and F element (dot chromosome) for all 12 sequenced *Drosophila* species. [a] For species with both an X and a neo-X chromosome, the statistics for the ancestral X (Muller A) appear above the divider and the neo-X (Muller D) below the divider. [b] There are few enough F-linked genes in each species (out of all genes with a single ortholog in all 12 species: dmel=32, dsec=32, dsim=28, dyak=32, dere=32, dana=28, dpse=30, dper=30, dwil=30, dmoj=36, dvir=30, dgri=30) that removing these genes does not affect the summary statistics for the autosomes pooled together.

	X		Autosomes[b]		F	
	a.a. div	**FOP**	**a.a. div**	**FOP**	**a.a. div**	**FOP**
D. melanogaster	0.0070 (0.011)	0.554 (0.557)	0.0075 (0.011)	0.526 (0.530)	0.013 (0.014)	0.253 (0.267)
D. sechellia	0.0038 (0.0057)	0.568 (0.574)	0.0041 (0.0058)	0.537 (0.540)	0.0034 (0.0034)	0.255 (0.266)
D. simulans	0.0014 (0.0042)	0.578 (0.577)	0.0022 (0.0043)	0.537 (0.542)	0.0027 (0.0056)	0.247 (0.268)
D. yakuba	0.012 (0.018)	0.580 (0.581)	0.013 (0.019)	0.538 (0.542)	0.025 (0.029)	0.249 (0.268)
D. erecta	0.014 (0.023)	0.580 (0.581)	0.014 (0.020)	0.541 (0.544)	0.021 (0.024)	0.254 (0.269)
D. ananassae	0.068 (0.088)	0.595 (0.587)	0.065 (0.082)	0.519 (0.523)	0.16 (0.14)	0.218 (0.226)
D. pseudoobscura[a]	0.082 (0.011)	0.573 (0.570)	0.077 (0.10)	0.538 (0.537)	0.064 (0.067)	0.295 (0.292)
	0.076 (0.095)	0.551 (0.554)				
D. persimilis[a]	0.089 (0.12)	0.566 (0.565)	0.080 (0.10)	0.539 (0.536)	0.064 (0.065)	0.297 (0.291)
	0.083 (0.10)	0.548 (0.550)				
D. willistoni[a]	0.11 (0.13)	0.375 (0.378)	0.11 (0.13)	0.362 (0.369)	0.080 (0.10)	0.327 (0.333)
	0.10 (0.13)	0.363 (0.368)				
D. mojavensis	0.068 (0.089)	0.531 (0.524)	0.067 (0.085)	0.473 (0.473)	0.093 (0.099)	0.304 (0.303)
D. virilis	0.043 (0.055)	0.515 (0.508)	0.043 (0.054)	0.482 (0.479)	0.050 (0.053)	0.337 (0.336)
D. grimshawi	0.075 (0.096)	0.476 (0.473)	0.072 (0.091)	0.457 (0.455)	0.083 (0.093)	0.339 (0.342)

However, in *D. persimilis* when all the autosomal genes are pooled, the X chromosome shows significantly increased rates of amino acid divergence (median amino acid divergence is 0.0797 and 0.0886 for the autosomes and X chromosome,

respectively; adjusted $P = 0.0008$, MWU). In addition, in *D. pseudoobscura* and *D. persimilis*, rates of amino acid divergence on the X chromosome exceed those rates on the neo-X chromosome (*D. pseudoobscura*: 0.0758 and 0.0819 for the neo-X and X, respectively; adjusted $P = 0.044$, MWU; *D. persimilis*: 0.0833 and 0.0886 for the neo-X and X, respectively; adjusted $P = 0.03$, MWU). In *D. yakuba* and *D. simulans*, however, autosomal genes show significantly increased rates of amino acid divergence (*D. yakuba*: 0.0127 and 0.0116 for the autosomes and X, respectively; $P = 0.039$, MWU; *D. simulans*: 0.00219 and 0.00144 for the autosomes and X, respectively; $P = 0.029$, MWU). This may in part be due to non-selective effects, as both *D. simulans* and *D. yakuba* show increased autosomal divergence at four-fold degenerate synonymous sites. Although we have less confidence in the genome sequences of *D. persimilis D. sechellia* and *D. simulans* because their low sequence depth and mosaic assembly, respectively, we see similar patterns in these species as we do in the species sequenced to high coverage, suggesting that these patterns are not artifacts.

Because of the putative lack of recombination on dot chromosome (Muller F), we expect to find an increased substitution rate because a lowered efficacy of purifying selection should result in the increased fixation of deleterious mutations. As expected, amino acid divergence on the dot chromosome is significantly higher than the other autosomes in *D. melanogaster*, *D. yakuba*, *D. erecta*, *D. ananassae*, and *D. mojavensis* ($P < 0.048$, all comparisons, MWU). We repeated all analyses after removing the dot-linked genes and find similar results, which is in part due to the small number of F-linked genes (Table 2.3).

Importantly, differences in gene composition on the X chromosome and the autosomes can also lead to differences in rates of evolution between the two chromosome sets. In *D. melanogaster*, for instance, genes with sex-specific biases in expression pattern appear to be distributed differently throughout the genome.

Accessory gland proteins, which play key roles in male reproduction, have a distribution significantly biased toward autosomes (SWANSON *et al.* 2001), and other proteins with male-biased expression patterns are also depleted on the X chromosome (PARISI *et al.* 2003; RANZ *et al.* 2003). These genes with sex-biased expression patterns may evolve more rapidly than other types of genes, particularly if they are involved in reproduction, as reproductive genes in *Drosophila* do appear to evolve rapidly (for review see HAERTY *et al.* 2007; PANHUIS *et al.* 2006; SWANSON and VACQUIER 2002). Because these differences in the gene composition of the X and the autosomes may confound our comparisons of evolutionary rates, we repeated our analyses removing genes with male-specific expression patterns (see Materials and Methods). The removal of these genes does not dramatically alter the amino acid divergence results: *D. persimilis* still shows evidence in support of increased substitution rates on the X chromosome while *D. simulans* shows the opposite.

We looked at the substitution rates on individual Muller elements for species that showed differences between the X chromosome and pooled autosomes. In *D. persimilis*, while median amino acid divergence is higher on the X than on every other individual chromosome arm, the increase is only significant in comparison with Muller elements E and F (adjusted $P < 0.036$, both comparisons, MWU). Similarly, in *D. simulans*, while median amino acid divergence is reduced on the X in relation to all other chromosomes, this decrease is not statistically significant in any single arm comparison (adjusted $P > 0.27$, all comparisons MWU). Likewise, in *D. yakuba*, rates of evolution are lower on the X than every autosomal chromosome arm, but only significantly so in the comparison with Muller elements B and F (adjusted $P < 0.015$, both comparisons, MWU). Because we only see a significant excess of X-linked divergence when the autosomes are pooled in these species, we suggest that the

evidence of the apparent increased efficacy of selection on the X chromosome is weak.

We can also use evolutionary parameter estimates from internal branches of the phylogeny to investigate clade-specific trends in comparative rates of evolution between the X and the autosomes. In the shared *obscura* group lineage, which includes both the branch leading to *D. pseudoobscura* and *D. persimilis* as well as the terminal branches (see Materials and Methods), rates of amino acid divergence are significantly increased on the X chromosome relative to both the autosomes and the neo-X chromosome (median divergence is 0.0827, 0.0854 and 0.0924 for the autosomes, neo-X and X, respectively; adjusted $P < 0.014$, both comparisons, MWU). Interestingly, amino acid divergence on the X chromosome also appears to be elevated in the *melanogaster* group, as well as in the subgenus *Sophophora*. It seems that this result may reflect an increased power to detect subtle differences in X-linked versus autosomal evolutionary rate by pooling information across internal branches, however it remains possible that this is not characteristic of the genome because this clade-specific analysis was limited to 1878 genes.

The pattern of substitution rates of the X chromosome and autosomes is not consistent across the phylogeny. While the *obscura* group, *melanogaster* group, and *Sophophora* subgenus in general and *D. persimilis* appear to support an increased efficacy of positive selection on the ancestral X chromosome (and neo-X) relative to the autosomes, *D. yakuba* and *D. simulans* show the opposite. Either way, the magnitude of these effects appears to be small. However, there are several other confounding factors that may also contribute to the observed patterns. Recent demographic events such as population bottlenecks are likely to play a role, though little is known about the demographic histories of most of the species studied here. Furthermore, with respect to *D. persimilis*, there are inversions on both the X and the

neo-X chromosomes that may have been fixed by positive selection (MACHADO *et al.* 2007); the inversions themselves as well as recent selective pressures on these inversions may also contribute to the pattern of increased rates of evolution on the *D. persimilis* X and neo-X chromosomes. The extent to which the patterns of amino acid divergence in these species are driven by differences in underlying neutral substitution rate between the X and the autosomes, demographic history, inversions or selective effects unfortunately remains unclear.

Divergence estimated by ω

To control for differences in the substitution rates at synonymous sites, we examined ω, or the ratio of nonsynonymous to synonymous substitution rates per site for all five species in this subgroup (Figure 2.1C, Table 2.2, Appendix Figure 2.3). This is particularly important given that rates of substitution at four-fold degenerate synonymous sites are consistently lower on the X than on the autosomes in two of the five species. We used the set of single-copy orthologs in the *melanogaster* subgroup because saturation at synonymous sites has not as yet been reached. While there are no clade-specific increases in estimates of ω within the *melanogaster* subgroup, *D. sechellia* and *D. simulans* show significant increases in ω for X-linked genes as compared with the pooled autosomal genes (*D. sechellia*: median ω is 0.0876 and 0.102 for the autosomes and the X, respectively; $P = 0.0002$, MWU; *D. simulans*: 0.0589 and 0.0757 for the autosomes and X, respectively; $P = 0.0003$, MWU). This is likely due at least in part to the decrease in neutral substitution rate on the X chromosome in these species, as there are no significant differences in substitution rates at nonsynonymous sites between the X and autosomes in these species (data not shown).

Because the dot chromosome (Muller F) is likely to upwardly bias ω for the autosomes due to the relaxed constraints from a putative lack of recombination on this chromosome, we repeated these analyses on the dataset with dot-linked genes removed. We did not see any significant differences in results, again this is probably because of the small number of dot-linked genes (Table 2.2). Repeating the analysis with male-specific genes removed also yielded similar results.

The patterns that we observe are not driven by individual autosomes: in *D. sechellia*, estimates of ω are higher on the X than on each of the four major autosomal arms, and significantly higher than on elements B, C, and E (adjusted $P > 0.032$, all comparisons, MWU). Estimates of ω are also significantly higher on the X than on each of the four major autosomes in *D. simulans* (adjusted $P > 0.035$, all comparisons, MWU).

Similar to our results using just divergence at nonsynonymous sites, we see support for an increased efficacy of positive selection for X-linked genes in some lineages when comparing ω between the X and autosomes. In particular, *D. simulans* and *D. sechellia* show higher estimates of ω for genes on the X relative to genes on the autosomes. It is possible that the higher ω in these species is driven at least in part by a reduced synonymous substitution rate arising from a lineage-specific increased constraint on synonymous sites in these species, or because we are conflating polymorphism and divergence.

Paired comparisons

Comparing pairs of orthologs between species or paralogs within species where one copy is X-linked and the other is autosomal provides an independent test of whether substitution rates differ between the X chromsome and the autosomes (THORNTON *et al.* 2006). We took advantage of an X-Autosome fusion in *D. persimilis*

and *D. pseudoobscura*, where the Muller element D was fused to ancestral X (Muller A), resulting in a neo-X chromosome. Muller D-linked genes are X-linked in *D. persimilis* and *D. pseudoobscura* and are autosomal in the rest of the phylogeny. Therefore, if there is an increased efficacy of positive selection on the X chromosome, Muller D-linked genes should evolve higher in *D. pseudoobscura*/*D. persimilis* than in *D. melanogaster*/*D. simulans*. These paired comparisons of orthologs in which one pair of orthologs is X-linked and the other pair is autosomal are particularly appropriate for testing for a greater efficacy of positive selection on the X chromosome, as the assumption is that the only difference between the pairs of genes is their chromosomal location. Thus, many potentially confounding factors such as gene function and degree of constraint are controlled for. We use ω estimated from paired alignements of *D. pseudoobscura*/*D. persimilis* and *D. melanogaster*/*D. simulans* as our metric for the substitution rate because it takes synonymous substitution rates into account, which is especially important given the underlying differences in neutral substitution rate between the X and the autosomes observed in several species. We only used genes that map unambiguously to Muller element D in all four species. We then compared the number of genes for which estimates of ω are higher in the *D. pseudoobscura*/*D. persimilis* comparison versus the *D. melanogaster*/*D. simulans* comparison to a baseline standard. Rather than using a single chromosome (the ancestral X chromosome, or Muller element A) as the baseline as has been done in other studies (THORNTON *et al.* 2006), we chose to do all possible comparisons of each major Muller element, individually (Table 2.4). We find a significant excess of these are D-linked genes with higher estimates of ω in *D. pseudoobscura*/*D. persimilis* than *D. melanogaster*/*D. simulans,* when Muller elements B, C, or E are used as the null ($P < 0.03$, Fisher's exact test, all

comparisons). These results suggest that there is an increased efficacy of positive selection on the X chromosome.

Table 2.4: Counts by Muller element of genes with estimates of ω higher in *D. melanogaster*/*D. simulans* comparison or *D. persimilis*/*D. pseudoobscura* comparison. *P*-values are based on Fisher's exact test, using a 2x2 contingency table comparing the counts from element D with counts from each other Muller element.

Muller Element	Number of genes with $\omega_{mel/sim} > \omega_{per/pse}$	Number of genes with $\omega_{mel/sim} < \omega_{per/pse}$	*P*-value
A	350	528	0.71
B	414	506	**0.0069**
C	453	584	**0.030**
D	434	677	N/A
E	741	957	**0.016**

However, *D. pseudoobscura* and *D. persimilis* are so closely related that divergence at nonsynonymous sites is very low (mean $d_N = 0.008$; median $d_N = 0.003$), which may limit our power to detect differences in evolutionary rate among chromosomes. To increase our power, we used a similar approach instead with clade-specific relative rates of amino acid divergence (see Materials and Methods) to test for an increased efficacy of positive selection on the X chromosome. For each gene, we estimated relative amino acid divergence in the *melanogaster* subgroup, the *obscura* group, as well as in *D. willistoni*, where Muller element D has also independently become X-linked. For each Muller element, we counted the number of genes for which (relative) amino acid divergence was higher in the *melanogaster* subgroup than in the *obscura* group and vice versa; we obtained similar counts for the comparison of (relative) amino acid divergence between the *melanogaster* subgroup with *D. willistoni* (Tables 2.5 and 2.6). For both the comparisons between the *melanogaster* subgroup and the *obscura* group and the comparison between the *melanogaster* subgroup and *D. willistoni*, there is a significant increase in the number of D-linked genes with higher estimates of (relative) amino acid divergence in the *obscura* group

or in *D. willistoni* than in the *melanogaster* subgroup when elements B or E are used as the baseline ($P < 0.017$, Fisher's exact test, all comparisons).

Table 2.5: Counts by Muller element of genes with estimates of (relative) amino acid divergence higher in *D. melanogaster* subgroup or *obscura group*. *P*-values are based on Fisher's exact test, using a 2x2 contingency table comparing the counts from element D with counts from each other Muller element.

Muller Element	Number of genes with $d_{N\,mel\ \text{group}} > d_{N\,obscura\ \text{group}}$	Number of genes with $d_{N\,mel\ \text{group}} < d_{N\,obscura\ \text{group}}$	*P*-value
A	408	460	0.98
B	501	440	**0.0054**
C	516	544	0.45
D	510	574	N/A
E	904	840	**0.013**

Table 2.6: Counts by Muller element of genes with estimates of (relative) amino acid divergence higher in *D. melanogaster* subgroup or *D. willistoni*. *P*-values are based on Fisher's exact test, using a 2x2 contingency table comparing the counts from element D with counts from each other Muller element.

Muller Element	Number of genes with $d_{N\,mel\ \text{group}} > d_{N\,willistoni}$	Number of genes with $d_{N\,mel\ \text{group}} < d_{N\,willistoni}$	*P*-value
A	410	458	0.22
B	466	475	**0.0023**
C	448	612	0.30
D	482	602	N/A
E	856	888	**0.017**

These results suggest that overall, the X chromosome may have an increased efficacy of positive selection relative to the autosomes. However these results are only tentative because they seem to differ based on which autosome is used as the baseline. These results are consistent with amino acid divergence on the X chromosome of the *obscura* group (both for the individual species and for the clade) being higher than the autosomes. In contrast, the weak support for faster-X in *D. willistoni* based on paired four-species comparisons coupled with the lack of significant difference in rates of amino acid divergence of X-linked and autosomal genes indicate that an increased efficacy of selection on the *D. willistoni* X is not sufficient to explain the results.

Rapidly evolving genes

An increased efficacy of positive selection on the X chromosome should be more readily detectable in genes that are potentially evolving under positive selection. To test this hypothesis, we compared estimates of ω for X-linked and autosomal genes on the subset of genes with evidence for positive selection based on the M7 versus M8 PAML comparison (see Materials and Methods). For the putative positively selected genes, when all of the autosomal genes are pooled together, estimates of ω are significantly higher for X-linked genes than for autosomal genes in *D. simulans* (median ω is 0.108 and 0.160 for the autosomes and X, respectively; P = 0.022, MWU). However, we do not see this pattern recapitulated for the individual chromosome arms (adjusted $P > 0.096$, all comparisons, MWU). This could be because of a lack of power, due to the small number of genes in the positively selected gene subset.

Because the M7 versus M8 comparison is rather stringent, and identifies genes that are evolving under positive selection across the entire phylogeny, we also examined the subset of genes that show lineage-specific accelerations in evolutionary rate (see Materials and Methods). When all autosomal genes are pooled together, estimates of ω are significantly increased for X-linked versus autosomal genes in the subset of genes with lineage-specific increases in evolutionary rate for *D. melanogaster* and *D. sechellia* (*D. melanogaster*: median ω is 0.257 and 0.350 for the autosomes and X, respectively; P = 0.039, MWU, *D. sechellia*: median ω is 0.399 and 0.468 for the autosomes and X, respectively; P = 0.043, MWU). Clade-specific analysis shows increased estimates of ω on the X in the *D. melanogaster* complex as well (median ω is 0.165 and 0.193 for the autosomes and X, respectively; P = 0.039,

MWU). Because of the small sample size in this gene subset, we cannot test each autosomal element against the X chromosome in each species.

The lack of consistent signal could be because many adaptively evolving genes only show evidence for positive selection at a small fraction of their sites (*DROSOPHILA* 12 GENOMES CONSORTIUM 2007). It could also be due to a lack of power: because we restricted the analysis to the set of genes with a single ortholog in the *melanogaster* group, there are fewer genes that show evidence of positive selection and/or have significantly accelerated rates of evolution. We therefore compared the proportion of X-linked and autosomal genes in each of these two gene subsets (*i.e*. positively selected gene subset and subset of genes with lineage-specific accelerations in evolutionary rate) to assess whether the X chromosome was particularly enriched for genes evolving rapidly. In *D. sechellia, D. simulans* and *D. erecta*, there is a marginally significant overabundance of X-linked genes among those genes that show evidence of positive selection across the entire phylogeny ($P < 0.089$, all comparisons, Fisher's exact test). In addition, within the *D. melanogaster* species complex, there is a clade-specific, marginally significant overrepresentation of X-linked genes in the positively selected gene subset ($P = 0.055$, Fisher's exact test). Similarly, the X chromosome appears to be enriched for genes with lineage-specific accelerations in rate of evolution specifically in both *D. sechellia* and *D. simulans* lineages ($P < 0.038$, both comparisons, Fisher's exact test); we see a similar clade-specific trend in the *melanogaster* species complex ($P = 0.0026$, Fisher's exact test). In *D. yakuba*, however, there is a dearth of X-linked genes with lineage-specific accelerations in evolutionary rate ($P = 0.023$, Fisher's exact test).

Thus, when our sample is enriched for genes evolving in a manner that is consistent with either positive selection across the phylogeny or lineage-specific increases in evolutionary rate, we do see a weak signal of increased efficacy of

positive selection on the X chromosome, particularly in the *D. melanogaster* species complex. This is evidenced not only by differences in the distributions of ω between the X and the autosomes, but also by the genomic distribution of rapidly and/or adaptively evolving genes.

Purifying selection

The efficacy of selection is expected to be increased on the X chromosome relative to the autosomes for both positive selection and purifying selection under the same conditions: assuming new deleterious mutations are on average recessive (CHARLESWORTH *et al.* 1987; Table 2.1).

Increased efficacy of purifying selection on the X

To test this hypothesis, we looked at levels of codon bias on the X chromosome relative to the autosomes in all 12 *Drosophila* species. Codon bias is the unequal usage of synonymous codons in protein coding sequences, and it is thought to be a consequence of selection-mutation-drift balance (AKASHI 1997; BULMER 1991; MCVEAN and CHARLESWORTH 1999; SHARP and LI 1986). Assuming that there is an optimal codon for each amino acid, corresponding to the most abundant tRNA, or the least error-prone amino-acyl tRNA charging reaction, then mutations away from these preferred codons might be weakly deleterious.

Codon bias is likely to be maintained by purifying selection because codon preferences appear conserved across the *Drosophila* phylogeny (*DROSOPHILA* 12 GENOMES CONSORTIUM 2007; VICARIO *et al.* 2007), suggesting that codon bias levels are likely near or at equilibrium in natural populations. For example, correlations between species in the frequency of optimal codons (FOP), are > 0.91 within the *D. melanogaster* subgroup (data not shown). Furthermore, preferred codons in

Drosophila are G- or C-ending (AKASHI 1995) yet the inferred mutational pressure is towards A and T in this system (PETROV and HARTL 1999; SINGH *et al.* 2006). Therefore, most novel mutations at synonymous sites will be away from preferred codons. Thus, we believe that codon bias is an appropriate measure for the efficacy of purifying selection because of the evidence that it is maintained by purifying selection.

Because of the presumed reduced efficacy of selection on the dot chromosome, we expect to see less codon bias on the dot. Except for *D. willistoni,* genes on the dot chromosome have significantly lower codon bias than genes on every other chromosome arm (adjusted $P << 0.0001$, all comparisons, MWU). In *D. willistoni,* Muller elements E and F have fused, and as a consequence, the F element in this species may behave differently from the dot chromosome of other species. However, even in *D. willistoni*, codon bias of genes on Muller element F is significantly lower than codon bias of genes on all other elements (adjusted $P < 0.001$, all comparisons, MWU) but element B. These data thus suggest that codon bias is a sensitive metric for evaluating the efficacy of purifying selection.

Previous reports suggest that codon bias of X-linked genes is significantly higher than codon bias of autosomal genes in *D. melanogaster* (COMERON *et al.* 1999; HAMBUCH and PARSCH 2005; SINGH *et al.* 2005) and *D. pseudoobscura* (SINGH *et al.* 2005), which is consistent with a greater efficacy of purifying selection on the X. We tested whether the increase in codon bias associated with X-linkage was consistent across the *Drosophila* phylogeny.

For all twelve species, estimates of FOP are significantly higher for genes on the X chromosome than for genes on the pooled autosomal chromosomes ($P << 0.0001$, all comparisons, MWU; Figure 2.1D, Table 2.3, Appendix Figure 2.1). In addition, FOP in genes on the neo-X chromosomes is significantly higher than FOP in genes on the autosomes for *D. pseudoobscura* and *D. persimilis* ($P < 0.0002$, both

comparisons, MWU). Moreover, FOP for genes on the ancestral X chromosome is significantly higher than FOP in genes on the neo-X chromosome in *D. willistoni*, *D. persimilis* and *D. pseudoobscura* ($P < 0.0001$, all comparisons, MWU).

Comparing FOP of X-linked genes versus genes on individual Muller elements yields similar results. In all species but *D. willistoni*, codon bias of ancestrally X-linked genes is significantly higher than codon bias of genes on every individual autosomal chromosome arm (adjusted $P < 0.0002$, all comparisons, MWU). In *D. willistoni*, codon bias on the ancestral X chromosome is significantly higher than codon bias of genes on all other elements (adjusted $P < 0.0005$, all comparisons, MWU) except for Muller element E. The consistent increase in codon bias on the X chromosome relative to the autosomes across the phylogeny is suggestive of an increased efficacy of purifying selection on the X chromosome.

There are other explanations for the increased levels of codon bias we see on the X chromosome than an increase in the efficacy of purifying selection. It is possible that the dosage problem due to the hemizygosity of the X chromosome in males, altered selection pressures, leading to higher levels of codon bias on the X chromosome, as has been suggested previously (SINGH *et al.* 2005). Furthermore, if codon bias were driven by positive selection rather than by purifying selection, the increase in codon bias associated with X-linkage could reflect the increased efficacy of positive selection on the X chromosome. We find this explanation to be unlikely, given the contrast between the consistency of the codon bias pattern across the phylogeny and the inconsistencies in X versus autosomal rates of protein evolution.

Conclusions and Future Directions

We took advantage of the complete sequencing of twelve eukaryotic *Drosophila* genomes to investigate potential differences in the efficacy of natural selection using rates of evolution between the X and the autosomes. We were able to

explore this on a genome-wide scale and for a large number of individual species and clades using both paired and unpaired approaches to test specifically for greater efficacy of positive selection on the X chromosome. We were able to directly ask whether positive selection is more efficient on the X than the autosomes by investigating a set of genes that are predicted to have experienced positive selection and genes that appear to be rapidly evolving on individual lineages. The latter genes are subject to bursts of substitutions and may be evolving adaptively.

Our results suggest a consistent increase in the efficacy of purifying selection on the X chromosome compared to the autosomes, because of the higher levels of codon bias on the X chromosomes (and neo-X chromosomes) relative to the autosomes across the phylogeny. The pattern is less clear for positive selection: we only find evidence for more efficient positive selection on the X chromosome in some lineages. The pattern of more efficient positive selection on the X chromosome is highly dependent on the metric used to measure adaptive substitutions and the lineage investigated. The lack of a consistently detectable effect across the *Drosophila* phylogeny may indicate that adaptive evolution from new mutations is not the dominant force that modulates evolutionary rate in these species.

There are two main hypotheses to explain this observation. One hypothesis is that the evolutionary rate for a gene is determined by the balance between the relative amount of purifying and positive selection a gene experiences; since most genes in *Drosophila* evolve under some evolutionary constraint, the resulting lower rates of subsitution may outweigh the effects of positive selection to increase the substitution rate. For genes which putatively evolve under positive selection, only an average of ~2% of codons within the gene experienced positive selection, while most of the rest of the codons evolved under selective constraint (*DROSOPHILA* 12 GENOMES CONSORTIUM 2007). Given the differences between species in geographic range, life

history characteristics and demographic history, this balance is not expected to necessarily be the same among species. This could explain why some species show evidence in support of an increased efficacy of selection on the X chromosome and other species do not. Alternatively, it may be that positive selection plays a large role in the evolution of protein-coding sequences in *Drosophila*, as has been suggested previously (BIERNE and EYRE WALKER 2004; *DROSOPHILA* 12 GENOMES CONSORTIUM 2007; SAWYER *et al.* 2003; SAWYER *et al.* 2007; WELCH 2006), but one or more of the underlying assumptions of the model have been violated. It could be that positive selection acts from standing variation rather than new mutations, or that selection for these mutations differs between the sexes, or that new positively selected mutations are not on average recessive. Any one of the instances would violate the theory underlying the models that say that selection should be more efficient on the X chromsome than the autosomes.

The observed differences in X-linked and autosomal substitution rates among species may reflect differences in life history traits or demographic history between species. The effective population sizes of species are likely to vary across the phylogeny: the sequenced species include both island endemics and cosmopolitan species. Insufficient polymorphism datasets in many of the species mean that little is known about their effective population sizes. However, polymorphism data suggest that the *D. sechellia* has a lower effective population size than *D. simulans* and *D. melanogaster* and nucleotide polymorphism data suggest that *D. melanogaster* has a smaller effective population size than *D. simulans* (*e.g*. MORIYAMA AND POWELL 1996). Finally, these results may be affected by the types of genes residing on the X chromosome versus the autosomes, and the variation among species with respect to relative rates of evolution of X-linked and autosomal genes may reflect interspecific differences in the types of genes that are X-linked versus autosomal. While the rate of

interchromosomal movement does appear to be quite low in *Drosophila* (BHUTKAR *et al.* 2007; RANZ *et al.* 2001; RICHARDS *et al.* 2005), the evolutionary depth of the phylogeny under study may be sufficiently great that species could differ in their distributions of different functional classifications of genes. Expression patterns can diverge rapidly between species, particularly for male-biased genes (MEIKLEJOHN *et al.* 2003), which could also alter the gene complements of the X and the autosomes in a lineage-specific manner. Finally, differences in the rates of recombination between species can also cause interspecific differences in the relative rates of X and autosome evolution. Crossover frequencies differ within the *D. melanogaster* species complex (TRUE *et al.* 1996), and there are suggestions that at least some regions in *D. pseudoobscura* and *D. simulans* may differ in their recombinational landscape relative to *D. melanogaster* (HAMBLIN and AQUADRO 1996; HAMBLIN and AQUADRO 1999; KULATHINAL *et al.* 2008), possibly associated with inversion polymorphism. Moreover, *D. pseudoobscura* appears to have higher rates of recombination than its sister species *D. persimilis*, though both of these species appear to have higher recombination rates than *D. melanogaster*. It is therefore possible that genic differences in local recombination rates between species can cause differences in the relative rate of evolution on the X chromosome and autosomes between species.

From our results, we can come to two clear conlusions. First, while the pattern is unclear for positive selection, our results suggest that the efficacy of purifying selection is higher on the X chromosome than on the autosomes across species, at least at synonymous sites. This suggests that deleterious synonymous mutations are partially recessive on average in *Drosophila*, and also indicates that purifying selection at synonymous sites acts predominantly on new mutations. Second, in some lineages there is support for an increase in the efficacy of positive selection on the X

chromosomes relative to the autosomes, suggesting that at least some new positively selected variants are on average at least partially recessive.

Overall, we believe these results are consistent with an increased efficacy of positive selection on the X chromosome relative to the autosomes. We suggest that while positive selection does contribute to rates and patterns of evolution, rates of adaptive evolution from new mutations are not sufficiently high to increase substitution rates on the X chromosome over the decrease in substitution rate that is expected from purifying selection.

This analysis has identified outstanding questions that we hope to investigate further in the future. Namely, the observation of reduced divergence at four-fold degenerate synonymous sites on the X chromosome of some species but not others may be suggestive of interspecific variation in sex-specific mutation rates and/or differences among species in life-history characteristics. Because our inferences of selection are confounded by these underlying processes, more sophisticated models are required to fully explain the lack of a consistent increase in the efficacy of positive selection on the X chromosome in *Drosophila*. Further work will be required to understand how the degree of variation in substitution rates between the X and the autosomes among species is affected by differences in lineage-specific patterns of positive selection, mutation rates and demography.

REFERENCES

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

AKASHI, H., 1995 Inferring Weak Selection from Patterns of Polymorphism and Sivergence at "Silent" Sites in *Drosophila* DNA. Genetics **139:** 1067-1076.

AKASHI, H., 1997 Codon bias evolution in *Drosophila*. Population genetics of mutation-selection drift. Gene (Amsterdam) **205:** 269-278.

AVERY, P. J., 1984 The population genetics of haplo-diploids and X-linked genes. Genetical Research **44:** 321-341.

BAUER DUMONT, V., and C. F. AQUADRO, 1997 Rates of DNA sequence evolution are not sex-biased in *Drosophila melanogaster* and *D. simulans*. Molecular Biology and Evolution **14:** 1252-1257.

BEGUN, D. J., and P. WHITLEY, 2000 Reduced X-linked nucleotide polymorphism in *Drosophila simulans*. Proceedings of the National Academy of Sciences of the United States of America **97:** 5960-5965.

BETANCOURT, A. J., Y. KIM and H. A. ORR, 2004 A pseudohitchhiking model of x vs. autosomal diversity. Genetics **168:** 2261-2269.

BETANCOURT, A. J., D. C. PRESGRAVES and W. J. SWANSON, 2002 A test for faster X evolution in *Drosophila*. Molecular Biology and Evolution **19:** 1816-1819.

BHUTKAR, A., S. RUSSO, T. F. SMITH and W. M. GELBART, 2006 Techniques for Multi-Genome Synteny Analysis to Overcome Assembly Limitations. Genome Informatics **17**.

BHUTKAR, A., S. M. RUSSO, T. F. SMITH and W. M. GELBART, 2007 Genome-scale analysis of positionally relocated genes. Genome Res **17:** 1880-1887.

BHUTKAR, A., S. W. SCHAEFFER, S. M. RUSSO, M. XU, T. F. SMITH *et al.*, 2008 Chromosomal rearrangement inferred from comparisons of 12 *Drosophila* genomes. Genetics **179:** 1657-1680.

BIERNE, N., and A. C. EYRE WALKER, 2004 The Genomic Rate of Adaptive Amino Acid Substitution in *Drosophila*. Molec. Biol. Evol. **21:** 1350-1360.

BULMER, M., 1991 The Selection-Mutation-Drift Theory of Synonymous Codon Usage. Genetics **129:** 897-908.

CHARLESWORTH, B., J. A. COYNE and N. H. BARTON, 1987 The relative rates of evolution of sex chromosomes and autosomes. American Naturalist **130:** 113-146.

CHINTAPALLI, V. R., J. WANG and J. A. T. DOW, 2007 Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. Nature Genetics **39:** 715-750.

COMERON, J. M., M. KREITMAN and M. AGUADE, 1999 Natural Selection on Synonymous Sites is Correlated with Gene Length and Recombination in *Drosophila*. Genetics **151:** 239-249.

COUNTERMAN, B. A., C. ORTIZ-BARRIENTOS and M. A. F. NOOR, 2004 Using Comparative Genomic Data to Test for Fast-X Evolution. Evolution **58:** 656-660.

DROSOPHILA 12 GENOMES CONSORTIUM, 2007 Evolution of Genes and Genomes on the *Drosophila* Phylogeny. Nature **450:** 203-218.

DROST, J. B., and W. R. LEE, 1995 Biological basis of germline mutation: comparisons of spontaneous germline mutation rates among *Drosophila*, mouse and human. Environmental and Molecular Mutagenesis **25** 48-64.

GARCIA-DORADO, A., C. LOPEZ-FANJUL and A. CABALLERO, 2004 Rates and effects of deleterious mutations and their evolutionary consequences, pp. 20-32 in

Evolution of Molecules and Ecosystems, edited by A. MOYA and E. FONT. Oxford University Press, London/New York/Oxford.

HAERTY, W., S. JAGADEESHAN, R. J. KULATHINAL, A. WONG, K. RAVI RAM *et al.*, 2007 Evolution in the fast lane: rapidly evolving sex-related genes in *Drosophila*. Genetics **177:** 1321-1335.

HAMBLIN, M. T., and C. F. AQUADRO, 1996 High nucleotide sequence variation in a region of low recombination in *Drosophila simulans* is consistent with the background selection model. Molecular Biology and Evolution **13:** 1133-1140.

HAMBLIN, M. T., and C. F. AQUADRO, 1999 DNA sequence variation and the recombinational landscape in *Drosophila pseudoobscura*: a study of the second chromosome. Genetics **153:** 859-869.

HAMBUCH, T. M., and J. PARSCH, 2005 Patterns of Synonymous Codon Usage in *Drosophila melanogaster* Genes with Sex-biased Expression. Genetics **170:** 1691-1700.

HARTL, D. L., and A. G. CLARK, 2007 *Principles of Population Genetics. Fourth Edition*. Sinauer Associates, Inc., Sunderland, Massachusetts.

HILL, W. G., and A. ROBERTSON, 1966 The effect of linkage on limits to artificial selection. Genet Res **8:** 269-294.

HOLM, S., 1979 A Simple Sequentially Rejective Bonferroni Test Procedure. Scandinavian Journal of Statistics **6:** 65-70.

HUTTER, S., H. LI, S. BEISSWANGER, D. DE LORENZO and W. STEPHAN, 2007 Distinctly Different Sex Ratios in African and European Populations of *Drosophila melanogaster* Inferred From Chromosome-wide Single Nucleotide Polymorphism Data. Genetics **177:** 469-480.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

KULATHINAL, R. J., S. M. BENNETT, C. L. FITZPATRICK and M. A. NOOR, 2008 Fine-scale mapping of recombination rate in *Drosophila* refines its correlation to diversity and divergence. Proc Natl Acad Sci U S A **105:** 10051-10056.

LARRACUENTE, A. M., T. B. SACKTON, A. J. GREENBERG, A. WONG, N. D. SINGH *et al.*, 2008 Evolution of protein-coding genes in *Drosophila*. Trends in Genetics **24:** 114-123.

LU, J., and C.-I. WU, 2005 Weak selection revealed by the whole-genome comparison of the X chromosome and autosomes of human and chimpanzee. Proc Natl Acad Sci U S A **102:** 4063-4067.

MACHADO, C. A., T. S. HASELKORN and M. A. F. NOOR, 2007 Evaluation of the genomic extent of effects of fixed inversion differences on intraspecific variation and interspecific gene flow in *Drosophila pseudoobscura* and *D. persimilis*. Genetics **175:** 1289-1306.

MCVEAN, G. A. T., and B. CHARLESWORTH, 1999 A population genetic model for the evolution of synonymous codon usage: Patterns and predictions. Genetical Research **74:** 145-158.

MEIKLEJOHN, C. D., J. PARSCH, J. M. RANZ and D. L. HARTL, 2003 Rapid evolution of male-biased gene expression in *Drosophila*. Proceedings of the National Academy of Sciences of the United States of America **100:** 9894-9899.

MORIYAMA, E. N., and J. R. POWELL, 1996 Intraspecific nuclear DNA variation in *Drosophila*. Molecular Biology and Evolution **13:** 261-277.

MUSTERS, H., M. A. HUNTLEY and R. S. SINGH, 2006 A Genomic Comparison of Faster-Sex, Faster-X, and Faster-Male Evolution Between *Drosophila melanogaster* and *Drosophila pseudoobscura*. Journal of Molecular Evolution **62:** 693-700.

ORR, H. A., and A. J. BETANCOURT, 2001 Haldane's Sieve and Adaptation From the Standing Genetic Variation. Genetics **157:** 875-884.

PANHUIS, T. M., N. L. CLARK and W. J. SWANSON, 2006 Rapid evolution of reproductive proteins in abalone and *Drosophila*. Philosophical Transactions of the Royal Society of London, Series B: Biological Sciences **361:** 261-268.

PARISI, M., R. NUTTALL, D. NAIMAN, G. BOUFFARD, J. MALLEY *et al.*, 2003 Paucity of genes on the *Drosophila* X chromosome showing male-biased expression. Science (Washington D C) **299:** 697-700.

PETROV, D. A., and D. L. HARTL, 1999 Patterns of nucleotide substitution in *Drosophila* and mammalian genomes. Proceedings of the National Academy of Sciences, USA **96:** 1475-1479.

POOL, J. E., and R. NIELSEN, 2007 Population size changes reshape genomic patterns of diversity. Evolution **61:** 3001-3006.

POWELL, J. R., and E. N. MORIYAMA, 1997 Evolution of codon usage bias in *Drosophila.* Proceedings of the National Academy of Sciences of the United States of America **94:** 7784-7790.

RANZ, J. M., F. CASALS and A. RUIZ, 2001 How malleable is the eukaryotic genome? Extreme rate of chromosomal rearrangement in the genus *Drosophila.* Genome Research **11:** 230-239.

RANZ, J. M., C. I. CASTILLO-DAVIS, C. D. MEIKLEJOHN and D. L. HARTL, 2003 Sex-dependent gene expression and evolution of the *Drosophila* transcriptome. Science **300:** 1742-1745.

RICE, W. R., 1984 Sex chromosomes and the evolution of sex dimorphism. Evolution **38:** 735-742.

RICHARDS, S., Y. LIU, B. R. BETTENCOURT, P. HRADECKY, LETOVSKY S *et al.*, 2005 Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. Genome Research **15:** 1-18.

SAWYER, S. A., R. J. KULATHINAL, C. D. BUSTAMANTE and D. L. HARTL, 2003 Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. Journal of Molecular Evolution **57:** S154-S164.

SAWYER, S. A., J. PARSCH, Z. ZHANG and D. L. HARTL, 2007 Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*. Proceedings of the National Academy of Sciences of the United States of America **104:** 6504-6510.

SCHOEFL, G., and C. SCHLOETTERER, 2004 Patterns of microsatellite variability among X chromosomes and autosomes indicate a high frequency of beneficial mutations in non-African *D. simulans*. Molecular Biology and Evolution **21:** 1384-1390.

SHARP, P. M., and W. H. LI, 1986 An evolutionary perspective on synonymous codon usage in unicellular organisms. Journal of Molecular Evolution **24:** 28-38.

SINGH, N. D., P. F. ARNDT and D. A. PETROV, 2006 Minor shift in background substitutional patterns in the *Drosophila saltans* and *willistoni* lineages is insufficient to explain GC content of coding sequences. BMC Biology **4:** doi:10.1186/1741-7007-1184-1137

SINGH, N. D., V. L. BAUER DUMONT, M. J. HUBISZ, R. NIELSEN and C. F. AQUADRO, 2007a Patterns of mutation and selection at synonymous sites in *Drosophila*. Mol Biol Evol **24:** 2687-2697.

SINGH, N. D., J. C. DAVIS and D. A. PETROV, 2005 X-linked genes evolve higher codon bias in *Drosophila* and *Caenorhabditis.* Genetics **171:** 145-155.

SINGH, N. D., J. M. MACPHERSON, J. D. JENSEN and D. A. PETROV, 2007b Similar levels of X-linked and autosomal nucleotide polymorphism in African and non-African strains of *Drosophila melanogaster*. BMC Evolutionary Biology **7**.

STOREY, J. D., and R. TIBSHIRANI, 2003 Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences, USA **98:** 7375-7379.

SWANSON, W. J., A. G. CLARK, H. M. WALDRIP-DAIL, M. F. WOLFNER and C. F. AQUADRO, 2001 Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. Proc Natl Acad Sci U S A **98:** 7375-7379.

SWANSON, W. J., and V. D. VACQUIER, 2002 The rapid evolution of reproductive proteins. Nature Reviews: Genetics **3:** 137-140.

THORNTON, K., D. BACHTROG and P. ANDOLFATTO, 2006 X chromosomes and autosomes evolve at similar rates in *Drosophila*: No evidence for faster-X protein evolution. Genome Research**:** gr.4447906.

THORNTON, K., and M. LONG, 2002 Rapid divergence of gene duplicates on the *Drosophila melanogaster* X chromosome. Molecular Biology and Evolution **19:** 918-925.

TRUE, J. R., J. M. MERCER and C. C. LAURIE, 1996 Differences in crossover frequency and distribution among three sibling species of *Drosophila*. Genetics **142:** 507-523.

VICARIO, S., E. N. MORIYAMA and J. R. POWELL, 2007 Codon usage in twelve species of *Drosophila.* BMC Evol Biol **7:** 226.

VICOSO, B., and B. CHARLESWORTH, 2006 Evolution on the X chromosome: unusual patterns and processes. Nature Reviews: Genetics **7:** 645-653.

WALL, J. D., P. ANDOLFATTO and M. PRZEWORSKI, 2002 Testing models of selection and demography in *Drosophila simulans*. Genetics **162:** 203-216.

WANG, K., L. KEAN, J. YANG, A. K. ALLAN, S. A. DAVIES *et al.*, 2004 Function-informed transcriptome analysis of *Drosophila* renal tubule. Genome Biology **5:** R69.

WELCH, J. J., 2006 Estimating the Genomewide Rate of Adaptive Protein Evolution in *Drosophila*. Genetics **173:** 821-837.

WONG, A., J. D. JENSEN, J. E. POOL and C. F. AQUADRO, 2007 Phylogenetic incongruence in the *Drosophila melanogaster* species group. Molecular Phylogenetics and Evolution **43:** 1138-1150.

YANAI, I., H. BENJAMIN, M. SHMOISH, V. CHALIFA-CASPI, M. SHKLAR *et al.*, 2005 Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification. Bioinformatics **21:** 650-659.

YANG, Z., 1998 Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Molecular Biology and Evolution **15:** 568-573.

CHAPTER 3[3]

TRANSLOCATION OF Y-LINKED GENES TO THE DOT CHROMOSOME IN *DROSOPHILA PSEUDOOBSCURA*

Introduction

The unique properties of the Y chromosome offer a distinct advantage to male-related genes residing there. Because Y-linked genes are only transmitted through males, they are protected from counter-selection in females. It therefore appears that Y-linkage would be selectively favored for genes with male-specific functions (FISHER 1930). Indeed, the 40 Mbp Y chromosome of *Drosophila melanogaster* is home to at least 12 protein-coding genes, all of which are putatively involved in male-related functions (CARVALHO *et al.* 2001; CARVALHO *et al.* 2000; VIBRANOVSKI *et al.* 2008). Males without a Y chromosome (X0) are sterile but otherwise phenotypically normal in *D. melanogaster* (ASHBURNER 1989; BRIDGES 1916). There are six male fertility factors on the *D. melanogaster* Y chromosome corresponding to regions that, when deleted, confer male sterility (KENNISON 1981).

However, Y-linkage also confers a notable disadvantage: males of most *Drosophila* species do not produce recombinant gametes. Empirical and theoretical evidence indicates that the absence of recombination and haploid transmission can significantly affect the evolutionary trajectory of Y-linked genes (BACHTROG 2003; BACHTROG 2004; CHARLESWORTH and CHARLESWORTH 2000). Selection acting anywhere on the Y chromosome can reduce the efficacy of selection elsewhere on the Y, causing an increased rate of fixation of deleterious alleles, reduced adaptation, and

[3] This chapter is currently under review at *Genetics* (A.M. Larracuente, M.A.F. Noor and A.G.Clark. Translocation of Y-linked genes to the dot chromosome in *Drosophila pseudoobscura*. *Genetics*. Submitted). M.A.F.N. set up and helped score male parent backcrosses and A.G.C. edited text.

subsequent degeneration of Y-linked loci (reviewed in CHARLESWORTH and CHARLESWORTH 2000).

A striking recent discovery about the Drosophila Y chromosome is that across the 12 species phylogeny, Y-linked genes experience a high rate of turnover (KOERICH *et al.* 2008). Despite the lability in gene content, a set of five genes appear to have been on the ancestral Y, prior to the split of the *Drosophila* and *Sophophora* subgenera (*kl-3, kl-2, ORY, PRY* and *PPr-Y*; KOERICH *et al.* 2008). The Y chromosomes of two species, *D. pseudoobscura* and *D. persimilis*, are unique in that they contain none of the ancestral Drosophila Y-linked genes (KOERICH *et al.* 2008). Instead, their Y chromosome may have originated from an X-Muller D fusion event that occurred in the ancestor of this species.

Several recent reports indicate that male fertility factors have an elevated tendency to change genomic location in *Drosophila* (CARVALHO and CLARK 2005; KOERICH *et al.* 2008; MASLY *et al.* 2006; RICHARDS *et al.* 2005). A striking example of male fertility factor movement in *Drosophila* is the translocation of five of the *D. melanogaster* Y-linked genes (*kl-3, ARY, kl-2, ORY* and *PPr-Y*), at least three of which may be male fertility factors (*kl-3, kl-2* and *ORY*), to an autosome in *D. pseudoobscura* (CARVALHO and CLARK 2005). These genes are also likely to be autosomal in *D. persimilis, D. miranda* (*pseudoobscura* subgroup), *D. affinis* and *D. azteca* (*affinis* subgroup) as they are not Y-linked in these species (CARVALHO and CLARK 2005). Nonetheless, these genes are highly conserved, and are transcribed and correctly spliced, and all but one of these genes have retained their testis-restricted expression (CARVALHO and CLARK 2005). This suggests that these likely functional genes may perform the same function as they do on the *D. melanogaster* Y. Interestingly, these genes underwent a drastic reduction in size: the introns of some of these genes reach megabases in length on the Y of *D. melanogaster* and *D. hydei,* (GATTI and

PIMPINELLI 1983; KUREK *et al.* 2000), whereas on the *D. pseudoobscura* autosome the introns are in the kilobase range (CARVALHO and CLARK 2005).

Mechanistically, a Y-autosome translocation could pose a problem with X-Y pairing in meiosis. In most Drosophila species, males do not undergo crossing over and so the autosomes pair during male meiosis at several homologous regions along much of their length (MCKEE 2004; VAZQUEZ *et al.* 2002). However, the ancestral Drosophila X and Y chromosomes do not have homologous sequences outside of the X-linked *Stellate* locus (with the Y-linked *Suppressor of Stellate* locus) and the rDNA repeats, which are typically found on both sex chromosomes in Drosophila (HENNIG *et al.* 1975; LOHE and ROBERTS 1990; ROY *et al.* 2005). It appears that the sex chromosomes pair at the repetitive non-transcribed intergenic spacer (IGS) region of the rDNA clusters, at least in species in the *melanogaster* subgroup (MCKEE 1996; MCKEE *et al.* 1992; AULT and RIEDER 1994; LOHE and ROBERTS 2000). A fusion of the ancestral Y to an autosome may mean that the ancestral Y chromosome rDNA cluster was also transferred, which raises a question about how the X and Y pair in this species.

To preserve the mechanism of X-Y pairing, the current Y chromosome would need to either acquire copies of the rDNA cluster or IGS repeats, or it could have evolved a novel mechanism for proper X-Y segregation. In order to understand how such a translocation would be tolerated and fixed in the population, the location of the rDNA must be determined. Previous results suggest X-linkage of the rDNA in *D. pseudoobscura*, by mapping a *bobbed* mutation, whose phenotype includes scutellar bristle defects and delayed development caused by a deficit of rDNA (STURTEVANT and NOVITSKI 1941; STURTEVANT and TAN 1937). Whether there are also copies of the rDNA and/or the IGS repeats on the *D. pseudoobscura* Y is a critical issue that had yet to be determined.

Here we report the mapping of the formerly Y-linked genes in *D. pseudoobscura* to the dot chromosome, suggesting that the heterochromatic environment of the dot chromosome may be important for the success of this Y chromosome translocation. We discover that the current Y chromosome contains no detectable rDNA genes, yet it has at least four large blocks of IGS repeats. Because such repeats are not observed on Muller's element D in other Drosophila species, the new Y chromosome of this species likely acquired and amplified the IGS, potentially to aid in X-Y pairing and normal disjunction.

Materials and Methods

Male Parent Backcrosses

The genic content of chromosome arms tends to be conserved in *Drosophila*; these conserved chromosome arms are called Muller elements A-F. *D. pseudoobscura* has three acrocentric autosomes (Muller B, C, and E) , a metacentric X chromosome (Muller A and D), a heterochromatic Y chromosome, and a dot chromosome (Muller F). We used two reciprocal male parent backcrosses between parental strains of *D. pseudoobscura*: *y*;*gl*;*or*;*inc* x Mather10 and Baja1 to map the *D. melanogaster* Y-linked genes in *D. pseudoobscura*. We used the following visible or molecular markers in the backcrosses: *glass* (*gl*) on the second chromosome (Muller E), *orange* (*or*) on the third chromosome (Muller C), *eyeless* (*ey*) on the fifth chromosome (Muller F or the dot) and SNPs detected in the formerly Y-linked genes *ORY* and *kl-3.* The fourth chromosome marker, *inc*, was not used because it is an unreliable marker with incomplete penetrance. We instead used microsatellites DPS4032 and DPS4033 on the fourth chromosome (Muller B; Ortiz-Barrientos et al. 2006). We scored 40 progeny of the male F_1 heterozygotes crossed to *y;gl;or;inc*. The genotype at *y*, *gl,* and *or* was determined by scoring phenotypes and in some cases confirmed using

microsatellite markers on the same chromosome**.** The genotypes at *ey, ORY* and *kl-3* were determined by PCR re-sequencing both strands of products containing a single base pair deletion in *ey* (forward 5' ACTTCACAGGTTGTACAGTAATGTGTACC; reverse 5' GTAGGTCGAGGCTATGAGGTCG; NOOR *et al.* 2001), a 5 bp deletion in *ORY* (forward primer 5' ATCGACTCGGCTATTGATGC and reverse primer 5' ACCATGAGCGTCTTTTTGCT) and an A/G SNP in *kl-3* (forward primer 5' TTTGGCGCTAGTAGCTGGTT and reverse primer 5' GGTCCCTTACCACGATCAGA). The Baja1 line contained a 97 bp deletion in the 5' upstream region of *ORY* (forward primer 5' CACCGACTCTACGTCGATGA and reverse primer 5' TTTTAGCCGAATCCCACATC) that was genotyped by visualizing PCR products on a gel. All PCR re-sequencing was done using BigDye chemistry and run on an ABI 3700 or ABI 3730XL DNA sequencer.

Female parent backcrosses

In an attempt to use recombination mapping to identify the location of the translocated genes on the dot chromosome, we scored the progeny of female parent backcrosses. We set up crosses using Baja1 and *y;gl;or;inc* flies to generate the F1 female heterozygotes that were then backcrossed to Baja1 males to get progeny for mapping. We scored a 5 bp deletion found in *ORY* and a single bp deletion in the intron of *ey* in 296 progeny using PCR re-sequencing methods described above.

Probes

In order to determine the location of the rDNA genes and their intergenic spacer (IGS) repeats, we performed fluorescent *in situ* hybridization using larval brain squashes. Probes were designed to 1,212 bp of the 18S rDNA gene (forward primer 5' TATCCGAGGCCCTGTAATTG and reverse primer 5' AATCCCAAGCATGAAAGTGG), 863 bp of the 28S rDNA gene (forward primer 5' GGGGAAAGAAGACCCTTTTG and reverse primer 5'

AACGGACGTAGCGTCATACC). Probes were designed to two of the IGS subrepeats (STAGE and EICKBUSH 2007; 226-bp IGS repeat forward primer 5' GTGGTCGTTTGTGGAACTTG and reverse primer 5' CTGTATTCATAATCAAATCATGCTCA; 267-bp IGS repeat forward primer 5' GAAAAGAAACTATTGTTAAGAGGCACT and reverse primer 5' AAATACACAGACATTGTCGGCTAA) using nick translation and biotinylated nucleotides (BioNick Labeling System, Invitrogen). The probes for both IGS subrepeats (267-bp and 226-bp IGS) and the probes for both 18S and 28S were combined before hybridization, unless stated otherwise. The probes used for *D. persimilis* (226-bp IGS, 267-bp IGS and 226-bp and 267-bp IGS combined), *D. affinis* (226-bp IGS, 267-bp IGS and 226-bp and 267-bp IGS combined), and *D. guanche* (18S and 28S combined and 267-bp IGS combined) are the same probes designed in and used for *D. pseudoobscura.*

Chromosome preparation

Brain squashes were carried out according to Pimpinelli *et al*. (2000) with some modification. Brains were dissected from third instar larvae, transferred to a hypotonic solution of 0.5% sodium citrate for 10 min then were fixed in a solution of 1.8% formaldehyde, 45% acetic and then squashed. The slides were frozen in liquid nitrogen and dehydrated in absolute ethanol, and kept dehydrated until use.

Fluorescence *in situ* hybridization (FISH)

FISH was carried out according to Pimpinelli *et al*. (2000) with some modification. Denaturation was achieved by placing slides on a heat block at 95° C for 6 min. Hybridization was performed at 30° C overnight. The slides were blocked in a 3% BSA solution then treated with Avidin-Rhodamine (Roche) for 30 min at 37° C

and washed in 4X SSC/0.1% Tween. Signal amplification was necessary in some cases to confirm the presence or absence of a weak signal. In these cases, chromosomes were fixed in a solution of 2.5% formaldehyde (rather than 1.8%), and 45% acetic acid. Amplification was achieved with an additional treatment with Avidin-Rhodamine after blocking in 10% normal goat serum in 1X PBS at 37°C then washed (4X SSC/0.1% Tween) and treated with biotinylated anti-avidin (Vector laboratories; for 30 min at 37°C). The slides were then washed again in 4X SSC/0.1% Tween, blocked in 3% BSA and treated with Avidin-Rhodamine again for 30 minutes both at 37°C and washed before mounting. The slides were mounted in Vectashield with DAPI (Vector laboratories) and visualized on an Olympus BX50 epifluorescence microscope. Images were taken with a QImaging camera (Retiga Ex*i* Fast 1394) at 200X with Metamorph imaging software, and then pseudo-colored and overlaid in Adobe Photoshop 7.0. FISH was done with larval brains pooled from several *D. pseudoobscura* lines collected from Mesa Verde National Park, CO by Steve Schaeffer and larval brains pooled from several *D. persimilis* lines collected from Santa Cruz Island, Channel Islands, CA by Luciano Matzkin. The *D. guanche* and *D. affinis* lines used for FISH were obtained through the Tucson *Drosophila* stock center: stock numbers 14011-0095.00 and 14012-0141.03, respectively.

Results

Mapping the Y translocation

The orthologs of the *D. melanogaster* Y-linked genes *kl-3, ARY, kl-2, ORY* and *PPr-Y* are located on 10 unmapped scaffolds from the Comparative Assembly Freeze 1 (CAF1) release of the *D. pseudoobscura* genome

(Figure 3.1; *kl-3*: Unknown group 52, Unknown group 114, Unknown group 358; *ARY* and *kl-2*: Unknown group 126; *ORY:* Unknown singleton 2759, Unknown singleton 626, Unknown group 378, Unknown group 301; and *PPr-Y*: Unknown group

301, Unknown group 499 and Unknown group 202; DROSOPHILA 12 GENOMES CONSORTIUM 2007), confirming the results of Carvalho and Clark (2005). In the closely-related species *D. persimilis*, all five genes are located on a single unmapped scaffold (Super 64) in the same order as they occur in *D. pseudoobscura* (DROSOPHILA 12 GENOMES CONSORTIUM 2007). An improved assembly of the Y-to-autosome translocated region was obtained by using an additional assembly from TIGR made with the Celera assembler in 2004 (CABA assembly; accession AAFS01000000). Scaffold 92961 (CH396212.1) links *kl-2* to *ORY*.

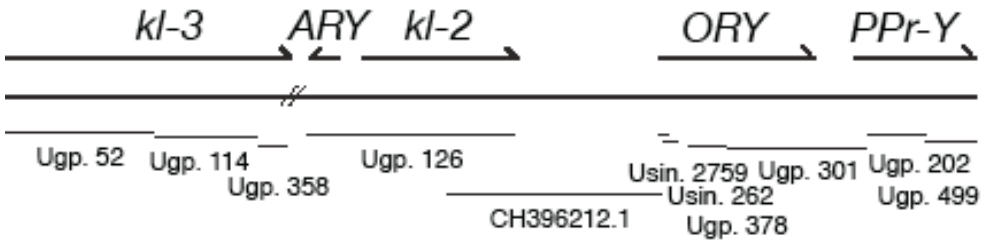


Figure 3.1. Organization of the Y-to-dot translocation. The five genes that were translocated from the Y chromosome to the dot chromosome are found on 10 scaffolds from the CAF1 assembly of the *D. pseudoobscura* genome (DROSOPHILA 12 GENOMES CONSORTIUM 2007). These scaffolds cover at least 158 kb, including estimated gap lengths, but this is likely to be a gross underestimate because it does not account for gaps between scaffolds. In *D. persimilis,* this region spans approximately 312 kb including estimated gaps within the single scaffold in which these genes are found. The scaffolds are all unmapped (Ugp.="unknown group" and Usin.= "unknown singleton"); the length of the lines drawn are proportional to the length of the scaffolds. *kl-3* does not overlap with the rest of the Y-to-dot genes, although in *D. persimilis* all of the genes, including *kl-3,* are contained on a single scaffold. *ARY* and *kl-2* can be linked to the first exons of *ORY* using a scaffold (CH396212.1) from the CABA assembly.

The formerly Y-linked genes *ARY, kl-2, ORY* and *PPr-Y* remain physically linked in *D. pseudoobscura* and can be ordered based on their locations in the CAF1 scaffolds (Figure 3.1; CARVALHO and CLARK 2005). The gene order in *D. persimilis* and inferred in *D. pseudoobscura* appears to be the same as the gene order in *D.*

melanogaster with the exception of at least one inversion involving the centromere (CARVALHO and CLARK 2005). We therefore only scored markers in *ORY* and *kl-3* to map the translocated region using two male parent backcrosses. Of the 26 progeny able to be scored at *ORY* for *y*;*gl*;*or*;*inc* x Mather10, all 26 *ORY* and *ey* markers co-segregated. Of the 21 progeny that were able to be scored at both *ORY* and *kl-3* for *y*;*gl*;*or*;*inc* x Baja1, all 21 *ORY, kl-3,* and *ey* markers co-segregated. In contrast, alleles on the other three autosomes were inherited independently of the markers on the ancestral Y and *ey*. These crosses indicate that the *D. pseudoobscura* orthologs of the *D. melanogaster* Y-linked genes are linked to the dot chromosome (chromosome 5) in *D. pseudoobscura*. These genes could not be mapped within the dot chromosome: we found no recombinant genotypes out of 296 progeny scored in our female parent backcrosses. This is consistent with the hypothesis that the dot chromosome of *D. pseudoobscura,* similar to *D. melanogaster* and *D. simulans*, undergoes little, if any recombination (ASHBURNER 1989; BRIDGES 1935; WANG *et al.* 2002; WANG *et al.* 2004). However, as yet, the physical distance between the Y-to-dot translocated genes and *ey* is unknown. BAC end sequence reads from *D. persimilis* indicate that the five genes *D. melanogaster* Y-linked genes are located on a scaffold that is linked to a known dot chromosome scaffold from the CAF1 assembly (Scaffold 51, *DROSOPHILA* 12 GENOMES CONSORTIUM 2007; Rod Wing, José Luis Goicoechea, *personal communication*), suggesting that this species possesses the same Y-to-dot translocation as *D. pseudoobscura.*

Identifying rDNA locations using in situ hybridizations

In situ hybridization of the 18S and 28S probes indicates the existence of an rDNA cluster on the left arm of the X chromosome, supporting the mapping of *bobbed* in *D. pseudoobscura* (STURTEVANT and NOVITSKI 1941; STURTEVANT and TAN 1937). The hybridization pattern suggests that few, if any, copies of 18S and 28S rDNA exist

on the Y chromosome, as no signal is detected in *D. pseudoobscura* (Figure 3.2). *In situ* hybridizations of the 226-bp and 267-bp IGS subrepeat probes map the IGS to the X chromosome, at a position that appears coincident to the location of the rDNA. We also find at least four additional blocks of bright staining on the Y chromosome using the IGS probes (Figure 3.2).

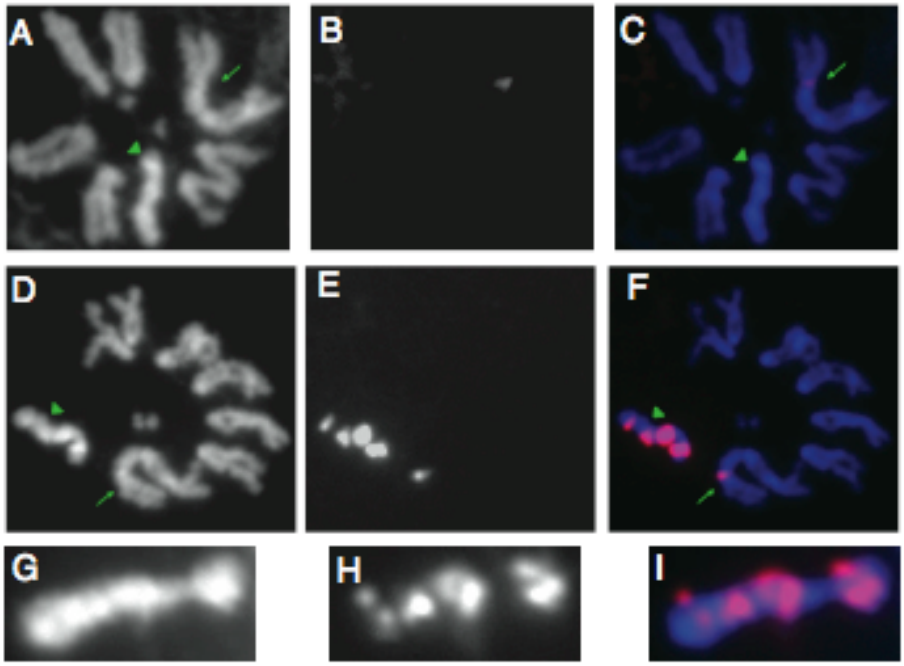
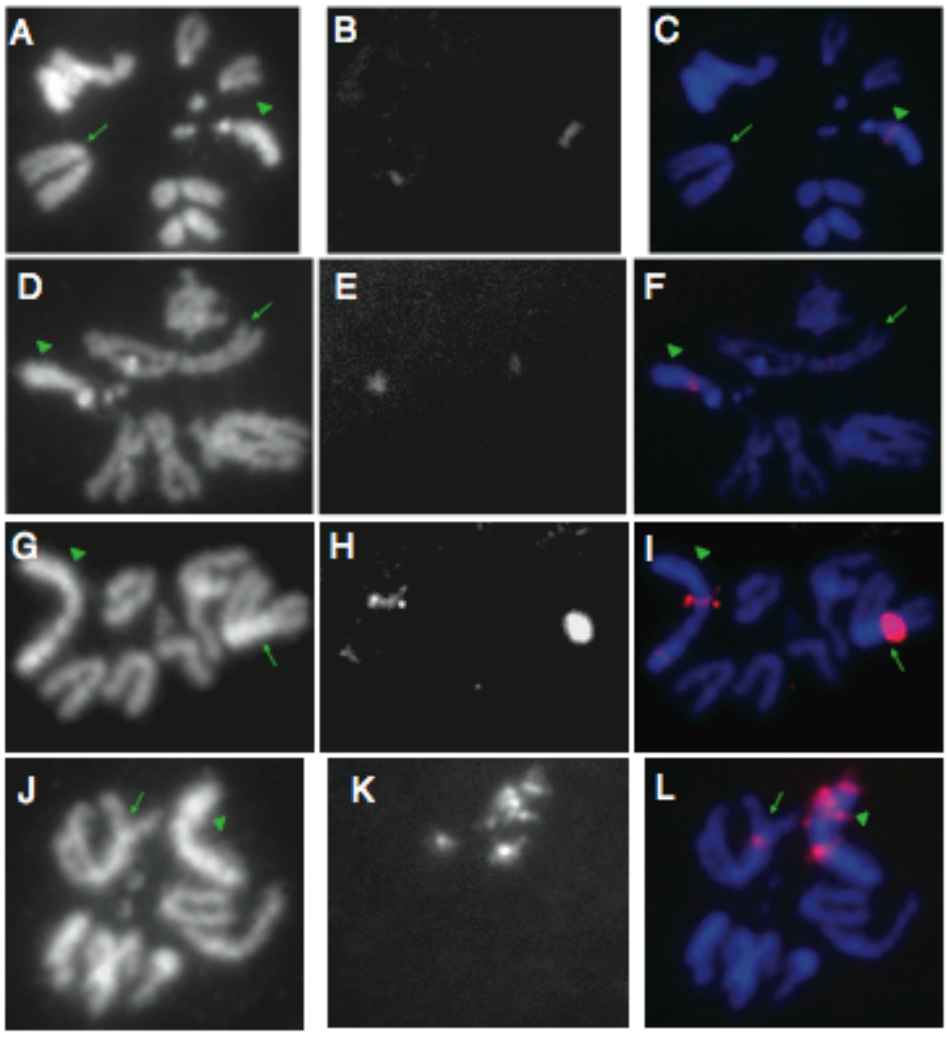


Figure 3.2. Hybridization of the rDNA probes (18S and 28S) and IGS probes to *D. pseudoobscura* mitotic chromosomes from larval brains suggest that the rDNA repeats are exclusively X-linked in *D. pseudoobscura* and the rDNA IGS spacer region is found on the X and in multiple clusters on the Y. The green arrow points to the X chromosome and the green arrowhead points to the Y chromosome. (A) DAPI staining of *D. pseudoobscura* mitotic chromosomes. (B) Signal for the 18S and 28S probes. (C) Overlay of A and B with DAPI staining colored in blue and the probe staining colored in red. (D,E, and F) DAPI DNA staining, probe hybridization and overlay for the IGS probes, respectively. (G,H, and I) DAPI DNA staining, IGS probe hybridization and overlay for just the Y chromosome, respectively. The rDNA genes 18S and 28S map exclusively to the X chromosome (shown as green arrow in C); no signal is seen on the Y chromosome, supporting the mapping of *bobbed* (STURTEVANT and NOVITSKI 1941; STURTEVANT and TAN 1937). The IGS subrepeats are present in at least four clusters on the Y chromosome (shown as green arrowhead in F) and in one cluster on the X chromosome (shown as green arrow in F). There is no signal from the dot chromosome.

Interestingly, it appears that only one subrepeat, the 267-bp IGS, is distributed widely across the current Y chromosome (Appendix Figure 3.1), whereas the 226-bp IGS subrepeat has few, if any copies on this chromosome (Appendix Figure 3.1). We see a very similar pattern on the *D. persimilis* Y: four clusters of IGS (mostly consisting of 267-bp repeats) are on the Y chromosome and one small band on the X chromosome (Appendix Figure 3.2). However, the 226-bp subrepeat occurs in two small clusters on the Y chromosome of the *D. persimilis* strains surveyed (Appendix Figure 3.2). *D. persimilis* also differs from *D. pseudoobscura* in rDNA repeat distribution; two small clusters of rDNA genes (18S and 28S) appear on the current Y chromosome in addition to the rDNA repeats on the X chromosome (Figure 3.3). *D. affinis*, an *obscura* group species that has the X-D fusion and a Y translocation, has rDNA genes both on the ancestral X and current Y chromosomes as evidenced by FISH signals (Figure 3.3).

Figure 3.3. FISH in *D. affinis* and *D. persimilis* using *D. pseudoobscura* probe shows that the current Y chromosomes acquired rDNA genes and spacers. The green arrow points to the X chromosome and the green arrowhead points to the Y chromosome. (A) DAPI staining of *D. affinis* mitotic chromosomes. (B) Signal for the 18S and 28S probes. (C) Overlay of B and C with DAPI staining colored in blue and the probe staining colored in red. (D, E and F) DAPI DNA staining of *D. affinis* mitotic chromosomes, probe hybridization and the overlay for the IGS probes, respectively. Both the rDNA genes 18S and 28S and the IGS repeats map to the X and Y chromosomes in *D. affinis* (shown as a green arrow and a green arrowhead for the X and Y, respectively in C and F). The IGS repeats are not tandemly repeated on the Y chromosome in *D. affinis*. (G) DAPI staining of *D. persimilis* mitotic chromosomes. (H) Signal for the 18S and 28S probes. (I) Overlay of G and H with DAPI staining colored in blue and the probe staining colored in red. Signal amplification (described in the Materials and Methods) was required to detect the signal in *D. persimilis* panels H and I. (J, K and L) DAPI DNA staining of *D. persimilis* mitotic chromosomes, probe hybridization and the overlay for the IGS probes, respectively. The rDNA genes 18S and 28S map to both the X chromosome and the current Y chromosome in *D. persimilis* (shown as a green arrow and a green arrowhead for the X and Y in I). The IGS repeats occur in at least four clusters on the current *D. persimilis* Y chromosome, similar to *D. pseudoobcura* (L).


A
B
C
D
E
F
G
H
I
J
K
L

This species does not have the clusters of IGS along the length of its current Y. Instead, it appears that they only have one block of IGS repeats on the X chromosome and one block of IGS repeats on the current Y, each that appear coincident with the rDNA loci on these chromosomes. In a related *obscura* group species that does not have the Y-A translocation or X-D fusion, *D. guanche,* FISH suggests that the rDNA repeats occur on both the X and the Y chromosomes, as well as the IGS sequences (Figure 3.4).

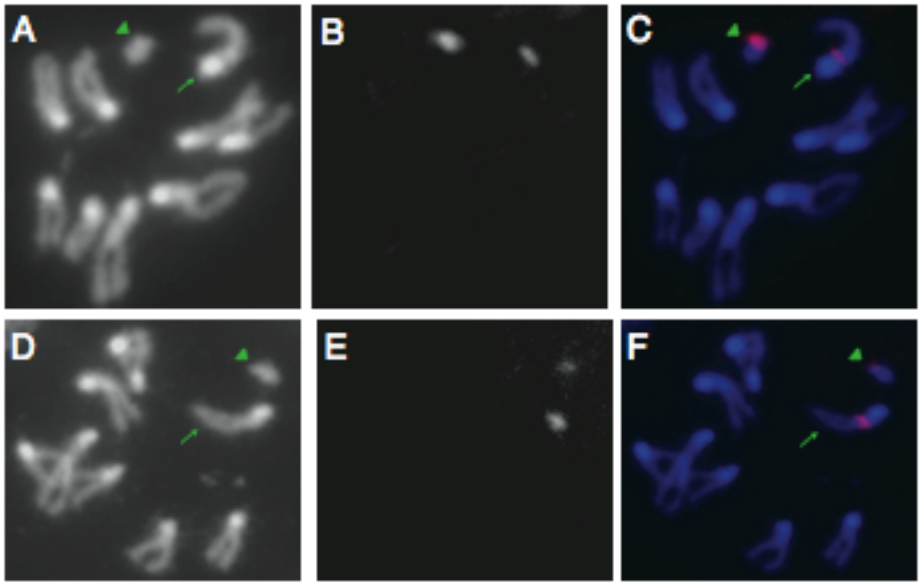


Figure 3.4. FISH in *D. guanche* using *D. pseudoobcura* probes suggests that the ancestral locations of the rDNA for *D. pseudoobscura* were likely on the X and Y chromosomes. The green arrow points to the X chromosome and the green arrowhead points to the Y chromosome. (A) DAPI staining of *D. guanche* mitotic chromosomes. (B) Signal for the 18S and 28S probes. (C) Overlay of B and C with DAPI staining colored in blue and the probe staining colored in red. (D, E and F) DAPI DNA staining of *D. guanche* mitotic chromosomes, probe hybridization and the overlay for the IGS probes, respectively. Both the rDNA genes 18S and 28S and the IGS repeats map to the X and Y chromosomes in *D. guanche* (shown as a green arrow and a green arrowhead for the X and Y, respectively in C and F). The IGS repeats are not tandemly repeated on the Y chromosome in this species, indicating that the current Y-linked IGS arrays in *D. pseudoobscura* are derived.

Discussion

We identified the location of five (*kl-3, ARY, kl-2, ORY,* and *PPr-Y*) genes present on the Y chromosome of *D. melanogaster* on the dot chromosome of *D. pseudoobscura* using male parent backcrosses. Four of these genes are a subset of the five genes hypothesized to be on the ancestral *Drosophila* Y chromosome (KOERICH *et al.* 2008). Thus it appears that most of the genic content of the ancestral Y chromosome translocated to the dot in this species. The only ancestral Y-linked gene that is not on the dot chromosome in *D. pseudoobscura* (*PRY*) is X-linked in this species (KOERICH *et al.* 2008), and was transferred to the X chromosome independently of the Y-to-dot translocation. The conserved order of the genes implies that the genes moved in a single translocation event, rather than individually.

The origin of the *D. pseudoobscura* current Y chromosome is unknown, however it is hypothesized to be derived from a Muller D element that originated in an X-Muller D fusion event that occurred in an ancestor of the species (CARVALHO and CLARK 2005). The current Y chromosome of *D. pseudoobscura* is necessary for male fertility (MORGAN *et al.* 1930), yet has no known genes homologous to *D. melanogaster* Y-linked genes. While it is possible that the current Y chromosome acquired a gene essential for male fertility from an autosome, given the high traffic of autosomal genes to the Y chromosome (CARVALHO *et al.* 2001; CARVALHO *et al.* 2000; KOERICH *et al.* 2008), it is also possible that this fertility factor was on the ancestral Y and fused to the current Y.

Interestingly, the *obscura* group species *D. affinis* has the X-D fusion and has a Y translocation we presume to be the same as in *D. pseudoobscura*, and yet X0 males of this species are fertile (VOELKER and KOJIMA 1971), indicating that the current Y chromosome of this species lacks male fertility factors. *D. affinis* may represent the ancestral state after the Y-to-dot translocation, and is X0-male fertile because all the

male fertility factors translocated from the Y chromosome. However, it is perhaps more likely that the ancestral state either retained or acquired a Y-linked male fertility factor that *D. affinis* subsequently lost, since most other *obscura* group species are X0 sterile (VOELKER and KOJIMA 1971).

Since males do not undergo crossing over in most *Drosophila* species, the ancestral *Drosophila* X and Y chromosomes pair at the repetitive intergenic spacer (IGS) region of the rDNA, at least in *D. melanogaster* (MCKEE 1996; MCKEE *et al.* 1992) and likely in *D. simulans* (AULT and RIEDER 1994; LOHE and ROBERTS 2000). If the entire ancestral Y chromosome translocated to the dot chromosome in *D. pseudoobscura*, this could disturb meiosis because the dot chromosome would then have elements causing it to pair with the X. To assure normal disjunction of the Y from the X, the current Y must have either acquired rDNA or IGS repeats, or developed a new mechanism for X-Y pairing. While the rDNA cluster maps exclusively to the X chromosome (Figure 3.2; STURTEVANT and NOVITSKI 1941; STURTEVANT and TAN 1937), we found that the IGS is present in at least four locations on the Y chromosome in addition to the small region on the X chromosome (Figure 3.2). A similar situation exists in *D. simulans*, where there is a large tandem array of 240-bp IGS repeats found on the tip of the Y chromosome, yet the Y has no detectable rDNA repeats (LOHE and ROBERTS 1990). It is hypothesized that the IGS was retained on the Y chromosome of *D. simulans* and thereby its function in X-Y pairing was conserved (AULT and RIEDER 1994; LOHE and ROBERTS 2000).

The conspicuous absence of the rDNA from the current *D. pseudoobscura* Y chromosome appears to be derived. Both the rDNA repeats (18S and 28S) and the IGS repeats map to the X and Y chromosomes in an *obscura* group species, *D. guanche* (Figure 3.4), which does not have either the X-Muller D fusion or the Y-Autosome translocation (CARVALHO and CLARK 2005). The ancestor of *D. affinis* and *D.*

persimilis likely acquired rDNA repeats on their current Y chromosomes: we found two small clusters of rDNA genes on the Y chromosome in addition to the X chromosome (Figure 3.3) in *D. persimilis* and one cluster of rDNA on the X and on the current Y of *D. affinis*. The rDNA clusters on the *D. affinis* Y chromosome may not be transcribed, however, since the Y chromosome was previously found not to have an effect on X-linked *bobbed* phenotypes (STURTEVANT 1940). While the location and copy number of the rDNA evolve very rapidly in short periods of time (EICKBUSH and EICKBUSH 2007; LOHE and ROBERTS 2000), our results from *D. guanche* coupled with the known locations of the rDNA in other *Drosophila* species (HENNIG *et al.* 1975; LOHE and ROBERTS 1990; ROY *et al.* 2005) suggests that the rDNA repeats are ancestrally X- and Y-linked (Figure 3.5, Figure 3.6A).

It is possible that the absence/presence of rDNA may be segregating within *D. pseudoobscura* and/or *D. persimilis*. We used larvae from several lines from only a single population each of *D. pseudoobscura* and *D. persimilis* and did not see any variation in the staining pattern. The Y chromosomes of these species are known to vary in size and morphology between populations (DOBZHANSKY 1935; DOBZHANSKY 1937). It is thus possible that the absence of rDNA is segregating in *D. pseudoobscura,* given its close relationship with *D. persimilis* and it is absent in the population we investigated from Mesa Verde. We are currently exploring this possibility.

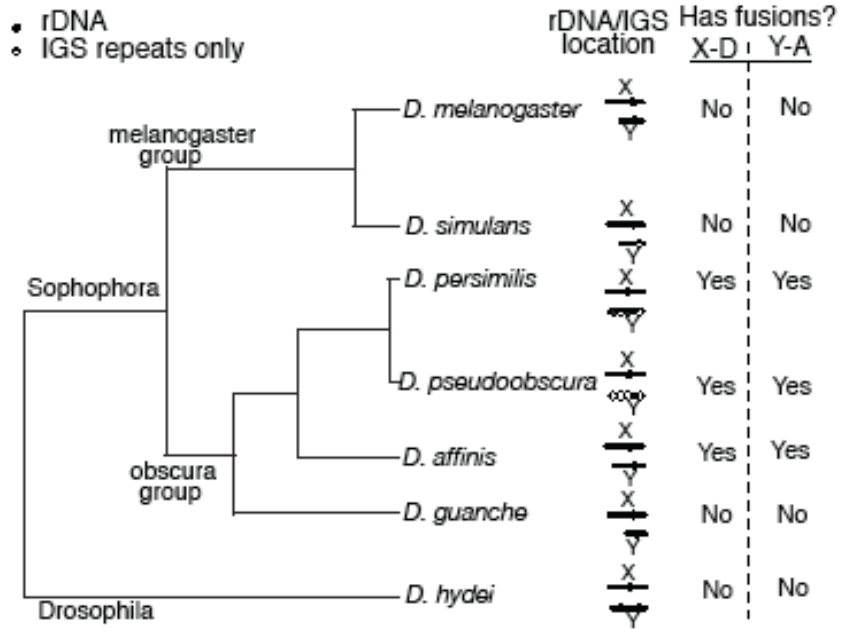
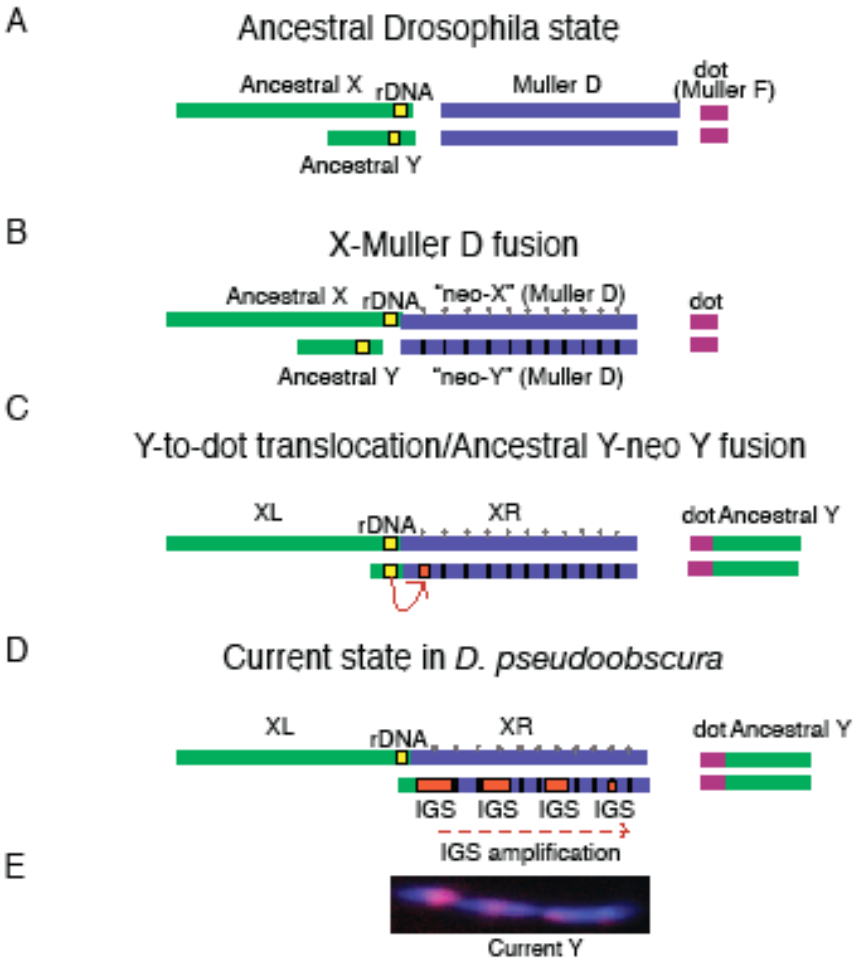


Figure 3.5. The location of the rDNA in the *melanogaster* group, *obscura* group and *D. hydei* in the *Drosophila* subgenus suggest that the ancestral locations of the rDNA are on the X and Y chromosomes. *D. simulans* presents an exception where the ancestral Y-linked rDNA locus was lost after the amplification of acquired IGS repeats on the Y chromosome (LOHE and ROBERTS 1990). *D. hydei* has two clusters of rDNA repeats on the Y chromosome in addition to the X (HENNIG *et al.* 1975). The three species that have the X-Muller D fusion (X-D) and the Y-A translocation (Y-A), *D. pseudoobscura*, *D. persimilis* and *D. affinis*, appear to have a more recent acquisition, followed by the amplification of the IGS repeats on the current Y chromosomes of *D. pseudoobscura* and *D. persimilis* (Figures 3.2 and 3.4). *D. pseudoobscura* appears to have lost its ancestral Y-linked rDNA locus (Figure 3.2), whereas *D. persimilis* and *D. affinis* retained rDNA on their current Y chromosomes.

Figure 3.6.We propose that there was a Y-to-dot translocation in *D. pseudoobscura*, and that the current Y chromosome originated from an X-D fusion, followed by acquisition of IGS sequences. (A) The Drosophila ancestral state with respect to the sex chromosomes and rDNA repeats. (B) An X-Muller D fusion occurred between 11 and 18 Myr ago. The homolog of the fused element (neo-Y) was transmitted as a Y. The neo-Y eventually degenerates and, in response, the neo-X becomes dosage compensated. (C) The IGS spacer is transferred to the neo-Y chromosome either from the ancestral X or as diagrammed: from the ancestral Y chromosome. Some fraction of the ancestral Y containing the rDNA may have remained free and fused with the current Y. (D) The ancestral Y chromosome translocated to the dot chromosome and the current Y chromosome is a degenerated neo-Y chromosome that originated from the X-D fusion event. The IGS spacer acquired by the current Y chromosome is amplified, producing at least four clusters and the ancestral Y-linked rDNA locus is lost. We illustrate this as a translocation of most of the ancestral Y to the dot with subsequent fusion of the remaining ancestral Y with the current Y rather than a complete fusion, although both scenarios are possible. The left arm of the X (XL) in *D. pseudoobscura* corresponds to the ancestral X chromosome and the right arm (XR) is the Muller D element. (E) The IGS spacer repeats (in red) occur in at least four large clusters on the Y chromosome (DAPI staining).


A
Ancestral Drosophila state
Ancestral X
rDNA
Muller D
dot
(Muller F)
Ancestral Y
B
X-Muller D fusion
Ancestral X
rDNA
"neo-X" (Muller D)
dot
Ancestral Y
"neo-Y" (Muller D)
C
Y-to-dot translocation/Ancestral Y-neo Y fusion
XL
rDNA
XR
dot Ancestral Y
D
Current state in D. pseudoobscura
XL
rDNA
XR
dot Ancestral Y
IGS
IGS
IGS
IGS
IGS amplification
E
Current Y

Model for the Y-dot translocation

Several evolutionary scenarios are consistent with the current state in *D. pseudoobscura*, where the ancestral Y chromosome is now on the dot chromosome and the rDNA repeats are exclusively X-linked while distinct clusters of IGS repeats occur on the current Y chromosome. Here we propose a model to explain the evolution of the Y chromosome and rDNA repeats in *D. pseudoobscura*. After the well-documented fusion between an autosome (Muller element D) and the X chromosome (X-D fusion; Figure 3.6B), the ancestor of *D. pseudoobscura,* (also *D. persimilis, D. affinis* and *D. azteca*) went through the transitional phase in which both the ancestral Y and the former Muller D autosome pair with the X-D fusion (Figure 3.6B). The Muller D arm that is now fused to the X has all the transmission properties of an X and is referred to as a "neo-X." The homolog of the fused Muller D is only passed through males and segregates like a Y chromosome, and thus is referred to as a "neo-Y" chromosome. This neo-Y chromosome followed a seemingly inevitable trajectory for Y chromosomes, including accumulation of loss-of-function mutations, which led to the evolution of dosage compensation on the neo-X (CHARLESWORTH 1978). The mechanics of such X-Autosome fusions are currently unknown: It is possible that the centromere of this new metacentric X chromosome is homologous to the Muller D centromere, the ancestral X centromere, or is composed of both.

Rather than the neo-Y chromosome being lost, as was previously thought to have happened in *D. pseudoobscura* (WHITE 1973), or the fusion of the neo-Y and current Y, which has happened in other *Drosophila* species including *D. albomicans* (YU *et al.* 1999), the ancestral Y translocated to the dot chromosome. The entire ancestral Y chromosome may have fused to the dot chromosome, however it is more likely that the majority of the ancestral Y translocated to the dot chromosome, but some fraction, including the rDNA (and possibly other unidentified genes including

fertility factors), remained free and later fused with the neo-Y. It is possible that the neo-Y chromosome acquired copies of the rDNA IGS from the X chromosome. It has been hypothesized that one possible mechanism for transfer of X-linked 240-bp IGS repeats to the Y chromosome of *D. simulans* could occur through extrachromosomal circular chromosomes (LOHE and ROBERTS 1990), which have been found in *D. melanogaster* embryos with many copies of the 240-bp IGS repeats (COHEN *et al.* 2003; PONT *et al.* 1987). The *D. pseudoobscura* neo-Y could have acquired the rDNA IGS in a similar manner and could also have acquired a gene essential for male fertility from an autosome. The ancestral Y, however could instead be the source of the IGS on the neo-Y if the two chromosomes fused after a substantial fraction of the ancestral Y translocated to the dot chromosome (Figure 3.6C, 3.6D). The IGS was amplified to form what appear to be tandem arrays (Figure 3.6E), and the ancestral Y-linked rDNA locus was lost. This hypothesis gains plausibility because *D. affinis* and *D. persimilis* have rDNA genes including the IGS repeats on their current Y chromosomes. Moreover, if the rDNA were never transferred to the dot, this would avoid problems with dot chromosome spuriously pairing with the X in meiosis.

This model suggests that the current Y chromosome is the degenerated neo-Y that originated from the Muller D element in the X-D fusion. In fact, 10 of the 15 identified genes on the current Y chromosome arose from two segmental duplications from the Muller D (CARVALHO and CLARK 2005), however it is yet to be determined whether these date back to the X-D fusion or whether they represent more recent duplications. *D. affinis* and *D. persimilis* may represent intermediate steps in the evolution of rDNA in *D. pseudoobscura;* the ancestor of *D. pseudoobscura, D. persimilis* and *D. affinis* may have acquired rDNA repeats on the neo-Y chromosome. *D. affinis* may represent the state following the acquisition of the rDNA by the neo-Y

(Figure 3.6C) and *D. persimilis* may represent the state after the amplification of the IGS on the neo-Y but this species has not lost its Y-linked rDNA.

It is also tempting to speculate that X-Y pairing in *D. pseudoobscura* is mediated, as in *D. melanogaster* and likely *D. simulans*, by the IGS repeats. It is unknown whether one of the ancestral functions of the IGS subrepeat is in X-Y pairing, however the Y-specific IGS amplification in *D. pseudoobscura* seems to parallel *D. simulans,* suggesting that perhaps its function is conserved. After amplification of the IGS on the *D. pseudoobscura* current Y, there may not have been a need for the rDNA in the ancestral Y location and so the ancestral Y-linked rDNA locus could be lost and the pairing function could rest with the IGS array on the current Y.

Features of the dot chromosome

The Y chromosome is an extreme outlier in the *Drosophila* genome due to its heterochromatic content, however 80% of the dot chromosome is highly heterochromatic in *D. melanogaster*. The dot chromosome in *D. pseudoobscura* appears to be very similar to the *D. melanogaster* dot chromosome both cytologically (SCHAEFFER *et al.* 2008; SLAWSON *et al.* 2006) and in its gene content (*DROSOPHILA* 12 GENOMES CONSORTIUM 2007; SCHAEFFER *et al.* 2008). The dot also has other peculiar features that distinguish it from the major autosomes: it is the only autosome that is haplosufficient across its length, triplo-dot flies are viable, and the addition of a dot chromosome shifts 2X:3A intersexes toward females (reviewed in LARSSON and MELLER 2006; RIDDLE and ELGIN 2006). It has even been hypothesized that the dot originated from a dosage-compensated X chromosome (reviewed in LARSSON and MELLER 2006; RIDDLE and ELGIN 2006). *Painting of Fourth* (POF), a dot-specific protein reminiscent of the MSL dosage compensation complex was recently shown to regulate the expression of dot-linked genes in *D. melanogaster* (JOHANSSON *et al.*

2007a; JOHANSSON *et al.* 2007b). It is possible that this protein is involved in regulating the expression of the formerly Y-linked genes on the dot chromosome of *D. pseudoobscura* since POF binds the dot chromosome in this species as well (LARSSON *et al.* 2004), although it is currently unknown whether it binds to the formerly Y-linked region.

Initially, the translocation of male-related Y-linked genes to the dot chromosome may have offered some fitness advantage to males bearing the translocation. It is unknown what this advantage could have been, however it is possible that it was related to the doubling in dosage of the genes following translocation and/or the mechanics of gene expression on the dot chromosome. The heterochromatic nature of the dot may contribute to the success of the translocation; translocation to the dot chromosome would still require the ability to function when in heterochromatin. For some heterochromatic genes, the proximity to heterochromatin is a requirement for proper expression (CORRADINI *et al.* 2007; WAKIMOTO and HEARN 1990), and recent work has implicated a role for transposable elements and repetitive sequences in this process (YASUHARA *et al.* 2005; YASUHARA and WAKIMOTO 2006).

If these genes possessed *cis*-regulatory sequences that directed testes-specific expression on the Y chromosome, there should be no adverse effects of moving to the dot chromosome from an expression standpoint. However, the expression of genes on the Y chromosome of *D. melanogaster* seems to be controlled at the level of chromosome-wide decondensation during spermatogenesis, so any gene that is translocated to the Y acquires this particular pattern of expression. If the genes lost their male-restricted expression on transfer to the dot, they would be exposed to female counter-selection, therefore these loci may face selection pressures from preventing mis-expression in females.

In the long term, a Y-autosome translocation would enjoy an increased efficacy of selection due to a larger effective population size and a higher rate of recombination. The effective population size of these genes increased four fold on transferring to the dot, resulting in lower frequencies of slightly deleterious variation in mutation-selection balance as well as augmenting the response to weakly beneficial mutations.

As in the *melanogaster* subgroup species, the dot chromosome of *D. pseudoobscura* may experience little to no meiotic crossing over (ASHBURNER 1989; BRIDGES 1935; WANG *et al.* 2002; WANG *et al.* 2004). We found no recombinant genotypes in our screen, although the sample was small so that one can infer only that recombination is at most rare. But even a small amount of recombination, as perhaps by gene conversion, could be highly beneficial in increasing the efficacy of natural selection (FISHER 1930). In conjunction with the larger effective population size, this in the presence of a specific male advantage associated with the translocation may be the reason this translocation fixed in the ancestor to *D. pseudoobscura* and that the drastic reduction in intron size across the region was evolutionarily possible. The discovery that the ancestral Y chromosome moved to the dot chromosome in *D. pseudoobscura* is an exciting addition to the parallels between the dot and the sex chromosomes.

REFERENCES

ASHBURNER, M., 1989 *Drosophila: A Laboratory Handbook*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

AULT, J. G., and C. L. RIEDER, 1994 Meiosis in *Drosophila* males. I. The question of separate conjunctive mechanisms for the XY and autosomal bivalents. Chromosoma **103:** 352-356.

BACHTROG, D., 2003 Adaptation shapes patterns of genome evolution on sexual and asexual chromosomes in *Drosophila*. Nat Genet **34:** 215-219.

BACHTROG, D., 2004 Evidence that positive selection drives Y-chromosome degeneration in *Drosophila miranda*. Nat Genet **36:** 518-522.

BRIDGES, C. B., 1916 Non-Disjunction as Proof of the Chromosome Theory of Heredity (Concluded). Genetics **1:** 107-163.

BRIDGES, C. B., 1935 The mutants and linkage data of chromosome four of *Drosophila melanogaster*. Biol Zh **4:** 401-420.

CARVALHO, A. B., and A. G. CLARK, 2005 Y chromosome of *D. pseudoobscura* is not homologous to the ancestral *Drosophila* Y. Science **307:** 108-110.

CARVALHO, A. B., B. A. DOBO, M. D. VIBRANOVSKI and A. G. CLARK, 2001 Identification of five new genes on the Y chromosome of *Drosophila melanogaster*. Proc Natl Acad Sci U S A **98:** 13225-13230.

CARVALHO, A. B., B. P. LAZZARO and A. G. CLARK, 2000 Y chromosomal fertility factors *kl-2* and *kl-3* of *Drosophila melanogaster* encode dynein heavy chain polypeptides. Proc Natl Acad Sci U S A **97:** 13239-13244.

CHARLESWORTH, B., 1978 Model for evolution of Y chromosomes and dosage compensation. Proc Natl Acad Sci U S A **75:** 5618-5622.

CHARLESWORTH, B., and D. CHARLESWORTH, 2000 The degeneration of Y chromosomes. Philos Trans R Soc Lond B Biol Sci **355:** 1563-1572.

COHEN, S., K. YACOBI and D. SEGAL, 2003 Extrachromosomal circular DNA of tandemly repeated genomic sequences in *Drosophila*. Genome Res **13:** 1133-1145.

CORRADINI, N., F. ROSSI, E. GIORDANO, R. CAIZZI, F. VERNI *et al.*, 2007 *Drosophila melanogaster* as a model for studying protein-encoding genes that are resident in constitutive heterochromatin. Heredity **98:** 3-12.

DROSOPHILA 12 GENOMES CONSORTIUM, 2007 Evolution of genes and genomes on the *Drosophila* phylogeny. Nature **450:** 203-218.

EICKBUSH, T. H., and D. G. EICKBUSH, 2007 Finely orchestrated movements: evolution of the ribosomal RNA genes. Genetics **175:** 477-485.

FISHER, R. A., 1930 *The Genetical Theory of Natural Selection*. Oxford University Press, Oxford.

GATTI, M., and S. PIMPINELLI, 1983 Cytological and genetic analysis of the Y chromosome of *Drosophila melanogaster*. I. Organization of the fertility factors. Chromosoma **88:** 349-373.

HENNIG, W., B. LINK and O. LEONCINI, 1975 The location of the nucleolus organizer regions in *Drosophila hydei*. Chromosoma **51:** 57-63.

JOHANSSON, A. M., P. STENBERG, C. BERNHARDSSON and J. LARSSON, 2007a Painting of fourth and chromosome-wide regulation of the 4th chromosome in *Drosophila melanogaster*. EMBO J **26:** 2307-2316.

JOHANSSON, A. M., P. STENBERG, F. PETTERSSON and J. LARSSON, 2007b POF and HP1 bind expressed exons, suggesting a balancing mechanism for gene regulation. PLoS Genet **3:** e209.

KENNISON, J. A., 1981 The Genetic and Cytological Organization of the Y Chromosome of *Drosophila melanogaster*. Genetics **98:** 529-548.

KOERICH, L. B., X. WANG, A. G. CLARK and A. B. CARVALHO, 2008 Low conservation of gene content in the *Drosophila* Y chromosome. Nature **456:** 949-951.

KUREK, R., A. M. REUGELS, U. LAMMERMANN and H. BUNEMANN, 2000 Molecular aspects of intron evolution in dynein encoding mega-genes on the heterochromatic Y chromosome of *Drosophila sp*. Genetica **109:** 113-123.

LARSSON, J., and V. H. MELLER, 2006 Dosage compensation, the origin and the afterlife of sex chromosomes. Chromosome Res **14:** 417-431.

LARSSON, J., M. J. SVENSSON, P. STENBERG and M. MAKITALO, 2004 Painting of fourth in genus *Drosophila* suggests autosome-specific gene regulation. Proc Natl Acad Sci U S A **101:** 9728-9733.

LOHE, A. R., and P. A. ROBERTS, 1990 An unusual Y chromosome of *Drosophila simulans* carrying amplified rDNA spacer without ribosomal-RNA genes. Genetics **125:** 399-406.

LOHE, A. R., and P. A. ROBERTS, 2000 Evolution of DNA in heterochromatin: the *Drosophila melanogaster* sibling species subgroup as a resource. Genetica **109:** 125-130.

MASLY, J. P., C. D. JONES, M. A. NOOR, J. LOCKE and H. A. ORR, 2006 Gene transposition as a cause of hybrid sterility in *Drosophila.* Science **313:** 1448-1450.

MCKEE, B. D., 1996 The license to pair: identification of meiotic pairing sites in *Drosophila.* Chromosoma **105:** 135-141.

MCKEE, B. D., 2004 Homologous pairing and chromosome dynamics in meiosis and mitosis. Biochim Biophys Acta **1677:** 165-180.

MCKEE, B. D., L. HABERA and J. A. VRANA, 1992 Evidence That intergenic spacer repeats of *Drosophila melanogaster* ribosomal-RNA genes function as X-Y pairing sites in male meiosis, and a general model for achiasmatic pairing. Genetics **132:** 529-544.

MORGAN, T. H., C. B. BRIDGES and J. SCHULTZ, 1930 The constitution of the germinal material in relation to heredity. Carnegie Inst. Washington Publ. **29:** 352-360.

NOOR, M. A., K. L. GRAMS, L. A. BERTUCCI, Y. ALMENDAREZ, J. REILAND *et al.*, 2001 The genetics of reproductive isolation and the potential for gene exchange between *Drosophila pseudoobscura* and *D. persimilis* via backcross hybrid males. Evolution **55:** 512-521.

PIMPINELLI, S., S. BONACCORSI, L. FANTI and M. GATTI, 2000 Preparation and Analysis of *Drosophila* mitotic chromosomes in *Drosophila Protocols*, edited by W. SULLIVAN, ASHBURNER, M., HAWLEY, R.S. Cold Spring Harbor Laboratory Press, New York.

PONT, G., F. DEGROOTE and G. PICARD, 1987 Some extrachromosomal circular DNAs from *Drosophila* embryos are homologous to tandemly repeated genes. J Mol Biol **195:** 447-451.

RICHARDS, S., Y. LIU, B. R. BETTENCOURT, P. HRADECKY, S. LETOVSKY *et al.*, 2005 Comparative genome sequencing of *Drosophila pseudoobscura*: Chromosomal, gene, and *cis*-element evolution. Genome Research **15:** 1-18.

RIDDLE, N. C., and S. C. ELGIN, 2006 The dot chromosome of *Drosophila*: insights into chromatin states and their change over evolutionary time. Chromosome Res **14:** 405-416.

ROY, V., L. MONTI-DEDIEU, N. CHAMINADE, S. SILJAK-YAKOVLEV, S. AULARD *et al.*, 2005 Evolution of the chromosomal location of rDNA genes in two

Drosophila species subgroups: *ananassae* and *melanogaster*. Heredity **94:** 388-395.

SCHAEFFER, S. W., A. BHUTKAR, B. F. MCALLISTER, M. MATSUDA, L. M. MATZKIN *et al.*, 2008 Polytene chromosomal maps of 11 *Drosophila* species: the order of genomic scaffolds inferred from genetic and physical maps. Genetics **179:** 1601-1655.

SLAWSON, E. E., C. D. SHAFFER, C. D. MALONE, W. LEUNG, E. KELLMANN *et al.*, 2006 Comparison of dot chromosome sequences from *D. melanogaster* and *D. virilis* reveals an enrichment of DNA transposon sequences in heterochromatic domains. Genome Biol **7:** R15.

STAGE, D. E., and T. H. EICKBUSH, 2007 Sequence variation within the rRNA gene loci of 12 *Drosophila* species. Genome Research **17:** 1888-1897.

STURTEVANT, A. H., 1940 Genetic data on *Drosophila affinis*, with a discussion of the relationships in the subgenus *Sophophora*. Genetics **25:** 337-353.

STURTEVANT, A. H., and E. NOVITSKI, 1941 The homologies of the chromosome elements in the genus *Drosophila*. Genetics **26:** 517-541.

STURTEVANT, A. H., and C. C. TAN, 1937 The comparative genetics of *Drosophila pseudoobscura* and *D. melanogaster*. J. Genet. **34:** 415-437.

VAZQUEZ, J., A. S. BELMONT and J. W. SEDAT, 2002 The dynamics of homologous chromosome pairing during male *Drosophila* meiosis. Curr Biol **12:** 1473-1483.

VIBRANOVSKI, M. D., L. B. KOERICH and A. B. CARVALHO, 2008 Two new Y-linked genes in *Drosophila melanogaster*. Genetics.

VOELKER, R. A., and K. I. KOJIMA, 1971 Fertility and fitness of X0 males in *Drosophila* 1. qualitative study. Evolution **25:** 119-128.

WAKIMOTO, B. T., and M. G. HEARN, 1990 The effects of chromosome rearrangements on the expression of heterochromatic genes in chromosome 2L of *Drosophila melanogaster*. Genetics **125:** 141-154.

WANG, W., K. THORNTON, A. BERRY and M. LONG, 2002 Nucleotide variation along the *Drosophila melanogaster* fourth chromosome. Science **295:** 134-137.

WANG, W., K. THORNTON, J. J. EMERSON and M. LONG, 2004 Nucleotide variation and recombination along the fourth chromosome in *Drosophila simulans*. Genetics **166:** 1783-1794.

WHITE, M. J. D., 1973 *Animal Cytology and Evolution*. Cambridge University Press, Cambridge.

YASUHARA, J. C., C. H. DECREASE and B. T. WAKIMOTO, 2005 Evolution of heterochromatic genes of *Drosophila*. Proc Natl Acad Sci U S A **102:** 10958-10963.

YASUHARA, J. C., and B. T. WAKIMOTO, 2006 Oxymoron no more: the expanding world of heterochromatic genes. Trends Genet **22:** 330-338.

YU, Y. C., F. J. LIN and H. Y. CHANG, 1999 Stepwise chromosome evolution in *Drosophila albomicans*. Heredity **83 (Pt 1):** 39-45.

CHAPTER 4

SIGNATURES OF SELECTION ON THE DOT CHROMOSOME OF *DROSOPHILA PSEUDOOBSCURA*

Introduction

The dot chromosome is a small autosome with sufficiently unusual properties to warrant a disproportionate share of research interest. Although it has little to no recombination *via* crossing over (ASHBURNER 1989; BRIDGES 1935; WANG *et al.* 2002; WANG *et al.* 2004), it nevertheless harbors many important genes. While most of the chromosome is heterochromatic, about 20% of the chromosome has a gene density similar to euchromatin, occurring in domains of interspersed heterochromatin and euchromatin (SUN *et al.* 2000). Many properties of this chromosome led some to believe that it has origins in an ancient X chromosome (LARSSON and MELLER 2006; RIDDLE and ELGIN 2006). *Painting of Fourth* (POF) is a protein that specifically recognizes and binds to the dot chromosome (LARSSON *et al.* 2001) to regulate the expression of its genes (JOHANSSON *et al.* 2007a; JOHANSSON *et al.* 2007b). The binding of POF to the dot chromosome is reminiscent of the male-specific-lethal (MSL) complex binding to the male X chromosome. The dot chromosome is the only autosome that has such a specific chromatin protein. Furthermore, the dot can exist in a copy number other than two without consequence (the dot is haplosufficient and triplo-dot flies are viable), and the addition of a dot chromosome to 2X:3A intersexes shifts the flies toward females, which are all properties of X chromosomes (reviewed in LARSSON and MELLER 2006; RIDDLE and ELGIN 2006). Regardless of its origins, the unique features of the dot chromosome have made this chromosome a topic of research for many years.

In *D. pseudoobscura*, the dot chromosome is even more exceptional. A Y-to-autosome (Y-A) translocation event in the *D. pseudoobscura* lineage (CARVALHO and

CLARK 2005) moved the ancestral *Drosophila* Y chromosome to the dot chromosome in this species (LARRACUENTE submitted). The Y is also on the dot chromosome in *D. persimilis* (LARRACUENTE submitted) and is likely to be on the dot in *D. affinis* and *D. miranda*, as the *D. melanogaster* Y-linked genes are present in both sexes in these species (CARVALHO and CLARK 2005). These species likely contain the same Y-A translocation as *D. pseudoobscura,* because they are closely related. The current Y chromosome of *D. pseudoobscura* is not homologous to the *D. melanogaster* Y chromosome, but is necessary for male fertility. This Y chromosome may be a degenerated neo-Y chromosome that originated in an X-Muller D fusion event that occurred in the lineage of *D. pseudoobscura* (CARVALHO and CLARK 2005; LARRACUENTE submitted).

A Y-A translocation means that the formerly Y-linked genes that were passed exclusively through males for many millions of years, are now passed through both sexes. Genes residing on the Y in *D. melanogaster* are expressed in the testes, presumably because the entire chromosome is decondensed during spermatogenesis, but it is currently unknown whether the Y also has *cis*-regulatory elements that would drive the expression in the testes. Four of the five Y-A translocated genes retained their exclusive testes-restricted expression on the *D. pseudoobscura* dot chromosome; one of the genes has a small amount of expression in female soma (CARVALHO and CLARK 2005). Because of the essentially haploid transmission and the complete lack of recombination on the Y, there is a reduced efficacy of selection on the Y that leads to an accumulation of deleterious mutations and reduced adaptation (reviewed in CHARLESWORTH and CHARLESWORTH 2000). This reduced efficacy of selection allows for the expansion of Y chromosomal introns up to megabases in size in *D. melanogaster* and *D. hydei* (GATTI and PIMPINELLI 1983; KUREK *et al.* 2000). Perhaps the most interesting feature of these genes in *D. pseudoobscura* is that they shrank at

least 10-fold after translocating from the Y chromosome to the dot chromosome (CARVALHO and CLARK 2005). The transcription of megabase-sized introns in an autosomal background with high gene density, like the dot chromosome, may be disadvantageous to a cell (CARVALHO and CLARK 1999), and may disrupt the nuclear compartmentalization of chromatin (PRACHUMWAT *et al.* 2004). A drastic size reduction of this scale may be driven by positive selection favoring the shortening of introns and intergenic regions over time.

In regions of low recombination like the dot chromosome, selection at one site will interfere with selection at linked sites; this effect is referred to as Hill-Robertson Interference (FELSENSTEIN 1974; HILL and ROBERTSON 1966). Therefore, if a deletion on one dot chromosome is favored and swept to fixation, it will affect levels of neutral variability at linked loci. The extent of the reduced variability caused by the sweep depends on the strength of selection at the site and the level of recombination (SMITH and HAIGH 1974). Since recombination is hypothesized to be low or non-existent on the *D. pseudoobscura* dot chromosome (LARRACUENTE submitted), as it is in *D. melanogaster* and *D. simulans* (ARGUELLO *et al.* submitted; ASHBURNER 1989; JENSEN *et al.* 2002; WANG *et al.* 2002; WANG *et al.* 2004), it is expected that recurrent selective sweeps will reduce variability and skew the frequency spectrum of mutations toward rare variants across the whole *D. pseudoobscura* dot chromosome. Variation in the dot chromosomal locus *eyeless* in *D. pseudoobscura*, showed very low levels of diversity (θ_w=0.00054 and π=0.00021), suggesting the action of selective sweeps (MACHADO and HEY 2003). An alternate hypothesis is that background selection (CHARLESWORTH *et al.* 1993) reduced the level of variation on the dot of *D. pseudoobscura* (MACHADO and HEY 2003). Background selection is capable of reducing variation on the dot chromosome: as deleterious mutations are purged from large populations by purifying selection, the effective population size (N_e) of a

population with no recombination under background selection is reduced to $f_0 N_e$, where f_0 is the fraction of chromosomes free from deleterious mutations (CHARLESWORTH *et al.* 1995). If the selection against these deleterious mutations is relatively weak, the frequency spectrum of mutations is also expected to be skewed toward rare variants (CHARLESWORTH *et al.* 1995).

In this paper, we conducted a survey of polymorphism across the dot chromosome of *D. pseudoobscura* in order to investigate the evolutionary forces that have shaped it. Our hypothesis is that the Y-to-dot translocated region has been affected by positive selection favoring the shortening of introns. The resulting recurrent selective sweeps would have resulted in recurrent reductions in size of the translocated region over time. We found that the *D. pseudoobscura* dot chromosome has significantly less diversity than expected under both a neutral model and under a model of population expansion that accounts for levels of autosomal diversity. These results are consistent with a simple model of recurrent selective sweeps, where the time since the most recent selective sweep was approximately 228,000 years ago. But even in this extreme case, alternatives such as background selection cannot be rejected.

Materials and Methods

Fly Strains

We surveyed dot chromosome variation in 64 lines of *D. pseudoobscura* from nine different populations, spanning the geographic range of the species (Table 4.1). All *D. pseudoobscura* lines were iso-female lines (except for the Flagstaff line) that were inbred for 10 or 11 generations by Richard P. Meisel. Genomic DNA was isolated from multiple male flies using a phenol chloroform DNA extraction.

Table 4.1. Strain information for 64 lines of *D. pseudoobscura* surveyed. [a] All lines were inbred for either 10 or 11 generations by Richard P. Meisel while in Steve Schaeffer's lab

STRAIN ID	SAMPLING LOCATION	YEAR	COLLECTED BY[a]
MV1	Mesa Verde National Park, CO	2005	S.W. Schaeffer
MV2	Mesa Verde National Park, CO	2005	S.W. Schaeffer
MV5	Mesa Verde National Park, CO	2005	S.W. Schaeffer
MV6	Mesa Verde National Park, CO	2005	S.W. Schaeffer
MV7	Mesa Verde National Park, CO	2005	S.W. Schaeffer
MV8	Mesa Verde National Park, CO	2005	S.W. Schaeffer
MV10	Mesa Verde National Park, CO	2005	S.W. Schaeffer
MV11	Mesa Verde National Park, CO	2005	S.W. Schaeffer
MV15	Mesa Verde National Park, CO	2005	S.W. Schaeffer
MV19	Mesa Verde National Park, CO	2005	S.W. Schaeffer
MV21	Mesa Verde National Park, CO	2005	S.W. Schaeffer
MV23	Mesa Verde National Park, CO	2005	S.W. Schaeffer
MV25	Mesa Verde National Park, CO	2005	S.W. Schaeffer
MV26	Mesa Verde National Park, CO	2005	S.W. Schaeffer
MV28	Mesa Verde National Park, CO	2005	S.W. Schaeffer
MV30	Mesa Verde National Park, CO	2005	S.W. Schaeffer
MV32	Mesa Verde National Park, CO	2005	S.W. Schaeffer
TU1	Tucson, AZ	2006	R.P. Meisel/S.W. Schaeffer
TU2	Tucson, AZ	2006	R.P. Meisel/S.W. Schaeffer
TU4	Tucson, AZ	2006	R.P. Meisel/S.W. Schaeffer
TU8	Tucson, AZ	2006	R.P. Meisel/S.W. Schaeffer
TU11	Tucson, AZ	2006	R.P. Meisel/S.W. Schaeffer
TU12	Tucson, AZ	2006	R.P. Meisel/S.W. Schaeffer
TU15	Tucson, AZ	2006	R.P. Meisel/S.W. Schaeffer
TU16	Tucson, AZ	2006	R.P. Meisel/S.W. Schaeffer
TU17	Tucson, AZ	2006	R.P. Meisel/S.W. Schaeffer
TU19	Tucson, AZ	2006	R.P. Meisel/S.W. Schaeffer
TU20	Tucson, AZ	2006	R.P. Meisel/S.W. Schaeffer
TU21	Tucson, AZ	2006	R.P. Meisel/S.W. Schaeffer
TU22	Tucson, AZ	2006	R.P. Meisel/S.W. Schaeffer
BMC1	Bosque de Apache NWF, NM	2006	S. Sheeley/B. McAllister
BMC2	Bosque de Apache NWF, NM	2006	S. Sheeley/B. McAllister
BMC4	Bosque de Apache NWF, NM	2006	S. Sheeley/B. McAllister
BMC5	Bosque de Apache NWF, NM	2006	S. Sheeley/B. McAllister
BMC6	Bosque de Apache NWF, NM	2006	S. Sheeley/B. McAllister
BMC9	Bosque de Apache NWF, NM	2006	S. Sheeley/B. McAllister
BMC10	Bosque de Apache NWF, NM	2006	S. Sheeley/B. McAllister
BMC11	Bosque de Apache NWF, NM	2006	S. Sheeley/B. McAllister
BMC12	Bosque de Apache NWF, NM	2006	S. Sheeley/B. McAllister
KB1	Kaibab National Forest, AZ	2005	S.W. Schaeffer
KB2	Kaibab National Forest, AZ	2005	S.W. Schaeffer
KB3	Kaibab National Forest, AZ	2005	S.W. Schaeffer
KB4	Kaibab National Forest, AZ	2005	S.W. Schaeffer
KB5	Kaibab National Forest, AZ	2005	S.W. Schaeffer
KB6	Kaibab National Forest, AZ	2005	S.W. Schaeffer

Table 4.1 (Continued)

KB9	Kaibab National Forest, AZ	2005	S.W. Schaeffer
KB10	Kaibab National Forest, AZ	2005	S.W. Schaeffer
KB12	Kaibab National Forest, AZ	2005	S.W. Schaeffer
SPE123-01	San Pablo Etla, Oaxaca, MEX	2003	T.A. Markow
SPE123-02	San Pablo Etla, Oaxaca, MEX	2003	T.A. Markow
SPE123-04	San Pablo Etla, Oaxaca, MEX	2003	T.A. Markow
SPE123-05	San Pablo Etla, Oaxaca, MEX	2003	T.A. Markow
SPE123-06	San Pablo Etla, Oaxaca, MEX	2003	T.A. Markow
SPE123-07	San Pablo Etla, Oaxaca, MEX	2003	T.A. Markow
SPE123-08	San Pablo Etla, Oaxaca, MEX	2003	T.A. Markow
SC02	Santa Cruz Is., Channel Is., CA	2004	L. Matzkin
SC12	Santa Cruz Is., Channel Is., CA	2004	L. Matzkin
SC13	Santa Cruz Is., Channel Is., CA	2004	L. Matzkin
GOLD14B	Goldendale, WA	1996	M.A.F. Noor
GOLD47	Goldendale, WA	1996	M.A.F. Noor
GOLD108	Goldendale, WA	1996	M.A.F. Noor
CHENY66	Cheney, WA	1996	M.A.F. Noor
CHENY75B	Cheney, WA	1996	M.A.F. Noor
FLAGSTAFF	Flagstaff, AZ	1993	M.A.F. Noor

Sequencing

Primers were designed using the CAF1 *D. pseudoobscura* assembly (Table 4.2; *DROSOPHILA* 12 GENOMES CONSORTIUM 2007). Each inbred line was subjected to PCR re-sequencing of 20 PCR products spanning the dot chromosome: 11 from the Y-to-dot translocated region and nine from the rest of the dot. The PCR conditions for each reaction were at least 40 cycles of 95°C for 30 seconds, 55°C for 45 seconds and 72°C for 1 minute. Unincorporated nucleotides were removed from PCR reactions using Exonuclease I/Shrimp alkaline phosphatase clean up prior to the sequencing reaction. PCR re-sequencing was done using the ABI Prism Big Dye cycle sequencing kit according to manufacturer's protocol and sequencing reactions were purified using a Sephadex column.

Table 4.2. Primer sequences and amplicon length for 20 loci surveyed.

FRAGMENT	Amplicon length	Forward (5'→3')	Reverse (5'→3')
Y-to-dot			
kl-3 ex4	792	TGGAAAGGTTTAACGCTCTG	GTGTCAATGACGATCGCAAC
kl-3 ex17	879	ACAAGGCTTACCAACGGATG	ATTCGCCTTATGCGTTTCTG
YA19-23	449	CTGAACGCCTGGCTATAACG	GCCGAATTCCGGATTTAATG
ARY ex1	322	AAGCGCAGGACTATTTGACC	ACACCTGGAACGCTATGGAG
kl-2 ex2	851	TCGGCGTGACTTGATAACTG	TGGAAGCCTCGGATACATTC
kl-2 ex5	835	AACGGCAGTGGCTTTATTTG	ACATCTCGGCCATGAATCTC
ORY mm3mm4	644	CACCGACTCTACGTCGATGA	TTTTAGCCGAATCCCACATC
YA5-7	545	AAAGCAACGGGAGGTTCATA	TGCGCAATCTGAAGTTTTTC
ORY 4 seq	529	CACCCACTCATGAGCAACAC	TTGAGGTCCCTCGAATTCAC
Ppr-Y 4 seq	589	GCAAGATGCATATCGTGGTG	ACAGCAGAAAAGGGCTGATG
YA10-16	581	TATGGGACAAAAAGGGATCG	GCGTGTCGCAATTCTATCCT
Dot, non-Y			
GA27948	650	TTCCCAGACCACCAAGTAGC	CAATTGCGTCAATGAGTTGG
GA10714	571	AACGCAGTTGGCTTAGATGC	ATATAGCCCATGCCCTTGTG
GA10734	573	AGAGCGTGTTCCTAGCCAAA	CGGTTCCTTGTGGATTCAGT
GA14409	592	CGAAAATCTTCGGGTGTTGT	GTACACCGGAAAGGCAAAAA
GA14323	615	ATTTCGCAATCTCCATCACC	CCCAGTGCTATGTGTGGTTG
GA15199	586	TTTCCGGAGAACGATACTGG	GTCTGGCTTTACCCCCTTTC
GA15170	572	TCAAGGCCAGAAACCAATTC	TGCTGGTGCTGCAATTATTC
GA13377	626	AAAGCCGTTGACGTATGGAG	CATTTGTCGGATCACTGTGG
ey	661	CTCTCGAGGAAATGGCTCAC	TCCGCCTAGTCCACTACCAC

Both the forward and reverse strand of each PCR product were sequenced using an ABI 3730 automated sequencer. A total of 10,593 bp were sequenced in each of 64 lines of *D. pseudoobscura*: 5,898 bp were from the Y-to-dot translocated region and 4,695 were from the rest of the dot chromosome. Traces were edited and aligned using Sequencher version 4.7 (Gene Codes, Ann Arbor, MI). Sequences were exported, concatenated (for each analysis) and formatted using PERL scripts.

Polymorphism analysis

We estimated θ_w (population mutation rate per silent nucleotide site), π (measure of nucleotide diversity per silent site), θ_h (measure of diversity per site that depends on high frequency variants), Tajima's *D*, haplotype diversity using the

program "compute" in the analysis version 0.7.2 package associated with the libsequence version 1.6.6 library (THORNTON 2003) for each locus sampled. To estimate recombination rate parameters and to analyze the amount of linkage disequilibrium across the sampled regions, the individual loci were concatenated. Because not every line has complete sequence for each locus, we imputed the missing data using fastPHASE (SCHEET and STEPHENS 2006). The concatenated dataset included 19 loci (*kl-3* exon 17 was excluded), totaling 10,007 bp, and 40 SNPs; sites with multiple segregating mutations and gaps were eliminated. The frequency of optimal codons (FOP) was estimated using the CodonW software (http://codonw.sourceforge.net/). The relative location of each fragment was determined using the CAF1 Whole Genome Shotgun assembly. For the Y-to-dot translocated region, gaps with estimated length were factored into the distance, however gaps of unknown length were not (*e.g.* gaps between scaffolds). Therefore the distances in the Y-to-dot translocated region and thus the total length of the dot chromosome are approximations and should be considered to be minimum distances.

Polymorphism on the autosomes (HAMBLIN and AQUADRO 1999a; MACHADO *et al.* 2002; SCHAEFFER *et al.* 2003; SCHAEFFER *et al.* 2001) and X chromosome (KOVACEVIC and SCHAEFFER 2000; MACHADO *et al.* 2002) was summarized from the literature. For the third chromosome data borrowed from Schaeffer *et al.* (SCHAEFFER *et al.* 2003), we averaged within-inversion summary statistics across five inversion types (AR, PP, ST, CH and TL).

Divergence

The orthologous sequences for all 20 dot-linked loci in *D. miranda* were obtained by blasting a *D. miranda* Illumina short-read sequence assembly (Doris Bachtrog and Zhou Qi, *personal communication*). Divergence between *D. pseudoobscura* and *D. miranda* was calculated as the average number of nucleotide

substitutions per site using DNAsp version 4.10.3. Divergence between *D. pseudoobscura* and *D. miranda* for the X chromosome and autosomes was summarized from the literature (HAMBLIN and AQUADRO 1999b; KOVACEVIC and SCHAEFFER 2000; MACHADO *et al.* 2002; SCHAEFFER *et al.* 2003). We used the codeml program in PAML version 3.15 (YANG 1997) to estimate the nonsynonymous substitution rate per nonsynonymous site (d_N) and synonymous substitution rate per synonymous site (d_S) and their quotient (ω) for a tree with three species: *D. pseudoobscura, D. persimilis,* and *D. miranda.* However, the loci we examined here did not have sufficient nonsynonymous sites or divergence to glean much information from this analysis.

Recombination

All analyses of recombination and linkage disequilibrium were performed on the concatenated dataset both with missing and imputed data, and no large difference between the results were found; only the results with the imputed dataset are reported. We estimated the minimum number of recombination events using the RecMin software (http://www.stats.ox.ac.uk/~myers/RecMin/) as R_m (HUDSON and KAPLAN 1985) and R_h (MYERS and GRIFFITHS 2003). Pairwise estimates of r^2 and D' were obtained using the genetics package version 1.3.2 in R. We also estimated ρ ($4Nr$; N is effective population size and r is the per generation per base rate of crossing over) and the rate of gene conversion to crossovers, f, (g/ρ; where g is the probability per generation of a gene conversion at a particular site) using the MaxHap software (HUDSON 2001; http://home.uchicago.edu/~rhudson1/source/maxhap.html). MaxHap uses a composite likelihood method, searching over a grid of values of ρ and f. We ran MaxHap without gene conversion to determine whether a model with crossing over but without gene conversion fits our data better than with gene conversion and crossing over. A grid of 500 ρ values ranging from $\rho = 0.000001$ to $\rho = 0.1$ and 800

values of f from 0 to 1000 was searched, with points equally spaced on a log scale. We did this for mean conversion tract lengths ranging from 50 to 600 base pairs, incrementing by 50 base pairs.

Neutral coalescent simulations

Neutral coalescent simulations were performed using a custom C++ script that uses the libsequence version 1.6.6 library (THORNTON 2003). We simulated 10,000 genealogies using θ_w drawn from a random uniform distribution corresponding to the range of the observed autosomal θ_w at silent sites (θ_w ~U(0.0037,0.0359)) summarized from the literature (HAMBLIN and AQUADRO 1999b; MACHADO *et al.* 2002; SCHAEFFER *et al.* 2003; SCHAEFFER *et al.* 2001).

Fitting demographic models

An approximate Bayesian computation (ABC) method (PRITCHARD *et al.* 1999; PRZEWORSKI 2003; THORNTON and ANDOLFATTO 2006) was used to infer the past demographic history of *D. pseudoobscura.* We simulated data sets under a model of population bottlenecks with exponential recovery, population expansion or constant population size with custom C++ scripts that use the libsequence version 1.6.6 library (THORNTON 2003). A rejection sampling technique was used to obtain m samples from the joint posterior distribution of θ_w, π, ρ and parameters in the demographic model (illustrated in Figure 4.1).

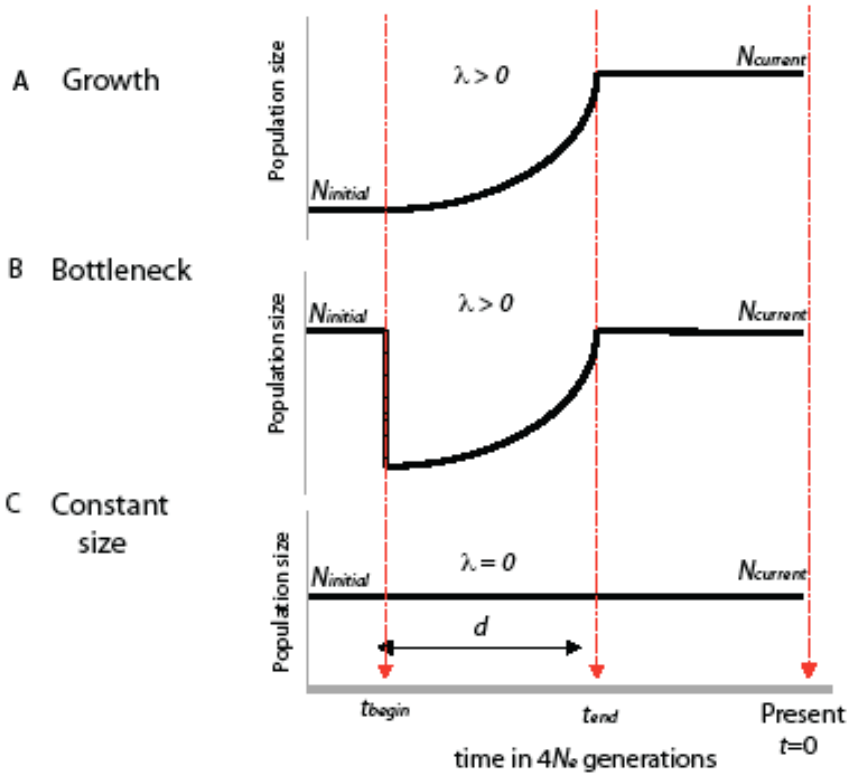


Figure 4.1. Demographic model parameters. The three models tested are each special cases of a bottleneck model. (A) A model of exponential population growth where the population begins growing at t_{begin} $4N_e$ generations in the past at a rate of λ, and reaches its current population size (stops growing) at t_{end} $4N_e$ generations in the past. (B) A model of a simple population bottleneck where the population bottlenecks at t_{begin} $4N_e$ generations in the past and recovers at a rate of λ, to the current population size at t_{end} $4N_e$ generations in the past. (C) A model of constant population size where the growth rate is simply set to $\lambda = 0$.

The rejection sampling algorithm is as follows:

1. Draw θ_w, ρ, t_{end}, d (duration of the event), λ (if growth or bottleneck), and $N_{initial}$ (if model is not a bottleneck) from prior distributions.
2. Simulate genealogies for 6 independent loci using the coalescent under demographic model based on empirical sample size for each locus.
3. Calculate summary statistics for simulated genealogies.

3. Accept or reject chosen parameter values conditional Po(S,$4\theta\tau$)/Po(S,S), and $|\pi_{observed}-\pi_{simulated}|\leq\varepsilon$ and $|S_{1observed}-S_{1simulated}|\leq\varepsilon$, where S is the number of segregating sites, π is the pairwise nucleotide diversity per silent site and S_1 is the number of singletons.
4. Return to step 1 and continue simulations until m desired samples from the joint posterior probability distribution are collected.

Each model is a special case of the bottleneck with exponential recovery model as diagrammed in Figure 4.1 and has a t_{end} parameter, which describes when the population stops growing and returns to its current size in terms of $4N_e$ generations. The duration (d) parameter describes the length of time the population is changing size, in terms of $4N_e$ generations. The t_{begin} parameter describes when the exponential growth begins and is calculated from $t_{end}+ d =t_{begin}$. The initial size parameter ($N_{initial}$) is the size of the population at the end of the simulation at t_{begin} $4N_e$ generations in the past.

For model choice, we compared acceptance rates because these models are nested and the acceptance rates are proportional to the marginal likelihoods when the prior distributions of parameters are equal, as they are in this case. In the first program, the prior distribution on the growth rate is split 50:50 between a model of constant population size (λ=0) and population expansion ($\lambda > 0$). In the second program, the prior distribution on the initial size parameter ($N_{initial}$) is split 50:50 between a model of population bottlenecks with exponential recovery (N varies between 0 and 2 $4N_e$ generations ago) and population expansion (N=-1 means that N is whatever the population shrank to and is governed by duration and λ). The model with the highest acceptance rate was considered to have a superior fit to the data over the other models.

The prior distributions used were: $\theta_w \sim \gamma(1,0.1)$, $t_{end} \sim$ U(0,2), $d \sim$ U(0,2), $\lambda \sim$ U(0,100) and $N_{initial} \sim$ U(0.0001,2). For simulations reported here, the tolerance parameter, ε, was set to 50% of the observed π and S_1, and m was set to 1,000. The data used in the simulations were summary statistics from non-coding sites from six loci on the second, third and fourth chromosome of *D. pseudoobscura* (*2001,2002, 2003,3002, 4002,4003*; MACHADO *et al.* 2002).

To assess the significance of the reduction in variation on the dot chromosome and the skew in the frequency spectrum of mutations, we generated data under the model of exponential growth with parameters of the growth model directly from the joint posterior distribution. Mutations were placed on the genealogy using an infinite sites model with θ_w drawn from a uniform prior distribution ($\theta_w \sim$U(0.0037,0.0359); $\theta_w \sim \gamma(1,0.1)$ gave the same results). These simulations were performed 10,000 times and an empirical cumulative probability function (ecdf) on the distribution of the summary statistics π and Tajima's *D* were used to calculate P values in R. The false discovery rate (FDR) was calculated using the p.adjust function in R (BENJAMINI 1995).

Modeling selective sweeps

We estimated the time since the most recent selective sweep using an ABC with rejection sampling, similar to the one described above (PRITCHARD *et al.* 1999; PRZEWORSKI 2003). We simulated selective sweeps in an expanding population: a population was expanding and at some time, t_{sweep} $4N_e$ generations in the past, all remaining lineages were coalesced. The population expansion part of the simulation was performed by simulating directly from the posterior distribution of the growth model that accounted for levels of autosomal variability, described in the previous section. Selective sweep simulations were performed using a custom C++ script that uses the libsequence version 1.6.6 library (THORNTON 2003). For these simulations,

we assumed that there is no recombination between the surveyed loci in the Y-to-dot translocation. A rejection sampling technique was used to obtain *m* samples from the joint posterior distribution of θ_w, π, and parameters in the expansion with sweep model.

The rejection sampling algorithm is as follows:

1. Draw θ_w and t_{sweep} from prior distributions.
2. Simulate genealogies for the concatenated Y-to-dot chromosome dataset using the coalescent under a population expansion with selective sweep model based on empirical sample size.
3. Calculate summary statistics for simulated genealogies.
4. Accept or reject chosen parameter values conditional $\text{Po}(S,4\theta\tau)/\text{Po}(S,S)$, and $|\pi_{observed}-\pi_{simulated}|\leq\varepsilon$, $|DTaj_{observed}-DTaj_{simulated}|\leq\varepsilon$, and $|S_{1observed}-S_{1simulated}|\leq\varepsilon$, where S is the number of segregating sites, π is the pairwise nucleotide diversity at silent sites, *DTaj* is Tajima's *D,* and S_1 is the number of singletons.
5. Return to step 1 and continue simulations until *m* desired samples from the joint posterior probability distribution are collected.

The prior distributions used were $\theta \sim \gamma(1,0.1)$ and $t_{sweep} \sim U(1\text{x}10^{-7},2)$. For simulations reported here, ε was set to 10% of the observed π, Tajima's *D*, and S_1 and *m* was set to 1,000. The data used were silent sites from the concatenated dataset of Y-to-dot chromosome loci sampled in this paper. These simulations were also performed using silent sites from the whole dot chromosome concatenated dataset and the concatenated dataset for just the part of the dot chromosome not involved in the translocation.

Results

Reduced diversity on the dot

We surveyed polymorphism and divergence at 20 dot-linked loci spanning the dot chromosome in 64 lines of *D. pseudoobscura* from nine different geographic locations. Eleven regions are in the Y-to-dot translocation and nine are on the rest of the dot chromosome (Table 4.3).

Table 4.3. Summary statistics for the 20 loci surveyed from the dot chromosome. The table shows the name of the fragment and corresponding region of the dot chromosome, either a Y-to-dot translocated gene or a part of the dot, not originating from the Y (dot, non-Y). The alignment length is the region considered for the polymorphism analysis. Also shown is S (the total number of segregating sites), and θ_w and π are shown for the total number of sites and silent sites and Tajima's D. The silent sites consider are both non-coding and synonymous sites. The significance of π and Tajima's D are indicated with a * for a significant value ($P < 0.05$) under the standard neutral model with constant population size and $^{\#}$ for a significant value under the model of population expansion described in the Results. † indicates that this value has a FDR of < 0.05.

Fragment	Region	Alignment Length	S	θ_w all (silent)	π all (silent)	D_{Taj}
Y-to-dot						
kl-3 exon 4	*kl-3*	583	0	0 (0)	0 (0*$^{\#}$)	NA
kl-3 exon 17	*kl-3*	684	3	0.00166 (0.00438)	0.00041 (0.00107*$^{\#\dagger}$)	-1.484*
YA19-23	*kl-3*	384	0	0 (0)	0 (0*$^{\#}$)	NA
ARY exon 1	*ARY*	289	1	0.00076 (0.00339)	0.00037 (0.00165*)	-0.675
kl-2 exon 2	*kl-2*	626	1	0.00036 (0.0017)	0.00014 (0.00063*$^{\#\dagger}$)	-0.860
kl-2 exon 5	*kl-2*	736	1	0.00029 (0)	0.00005 (0*$^{\#\dagger}$)	-1.088*$^{\#}$
ORY mm3mm4	*ORY*	610	2	0.00085 (0.00064)	0.00039 (0.00011*$^{\#\dagger}$)	-0.934
YA5-7	*ORY*	546	1	0.00039 (0.00126)	0.00037 (0.00119*)	-0.0687

Table 4.3 (Continued)

ORY 4 seq	*ORY*	458	7	0.00334 (0.00446)	0.00058 (0.00078*#†)	-2.104*#†
Ppr-Y 4 seq	*Ppr-Y*	505	5	0.00265 (0.00242)	0.00041 (0.00037*#†)	-1.971*#
YA10-16	*Ppr-Y*	477	0	0 (0)	0 (0*#†)	NA
Dot, non-Y						
GA27948	Dot, non-Y	578	2	0.00074 (0.00093)	0.00042 (0.00014*#†)	-0.749
GA10714	Dot, non-Y	450	2	0.00095 (0.00459)	0.00015 (0.00070*#†)	-1.442*
GA10734	Dot, non-Y	498	0	0 (0)	0 (0*#†)	NA
GA14409	Dot, non-Y	540	4	0.00178 (0.00238)	0.00032 (0.00043*#†)	-1.862*#
GA14323	Dot, non-Y	521	1	0.00043 (0.00075)	0.00008 (0.00013*#†)	-1.103
GA15199	Dot, non-Y	475	1	0.00046 (0)	0.00014 (0*#†)	-0.891
GA15170	Dot, non-Y	518	3	0.00135 (0.00358)	0.00145 (0.00155*)	0.153
GA13377	Dot, non-Y	569	2	0.00075 (0.0013)	0.00012 (0.0002*#†)	-1.442*
EY	Dot, non-Y	546	6	0.00239 (0.00242)	0.00221 (0.00224*#)	-0.192

Average overall θ_w and π for the dot chromosome are 0.00096 and 0.00038, respectively, which is significantly lower than variation reported on the autosomes and X chromosome of *D. pseudoobscura.* Several loci had no variation, both in the Y-to-dot translocated region (kl3-ex4, YA19-23, YA10-16; Table 4.3) and the part of the dot not associated with the translocation (dot, non-Y; GA10734; Table 4.3). The most variable fragment we surveyed in the Y-to-dot translocated region was in *ORY* (*ORY* 4 seq overall $\theta_w = 0.00446$, silent site $\theta_w = 0.00334$ Table 4.3). Interestingly, this region had 97 bp, 69 bp, 7 bp and 1 bp insertion/deletion polymorphisms segregating, at a frequency of 35.6%, 3.39%, 45.76% and 15.25%, respectively. The most variable of the dot chromosome loci not involved in the translocation (dot, non-Y) was *ey* (overall

θ_w = 0.00239, silent θ_w = 0.00242), which is nearly an order of magnitude higher than was previously estimated at this locus (MACHADO and HEY 2003). The Y-to-dot translocation does not significantly differ from the rest of the dot chromosome in levels of variation ($P_{\theta w}$=0.8300, P_{π}=0.8173, Mann-Whitney U-test, MWU; Table 4.4), supporting the idea that there is little recombination on this chromosome.

Table 4.4. Mean and median diversity estimates θ_w and π for all sites (total) and at silent sites. [a] 10 loci from second chromosome (HAMBLIN and AQUADRO 1999b; MACHADO *et al.* 2002); 9 loci from third chromosome (MACHADO *et al.* 2002; SCHAEFFER *et al.* 2003); 4 loci from fourth chromosome (MACHADO *et al.* 2002; SCHAEFFER *et al.* 2001). [b] includes 11 loci from the Y-to-dot translocation and 9 loci from the rest of the dot, outside of the translocation (dot, non-Y).

	n	**Total θ_w mean (median)**	**Total π mean (median)**	**Silent θ_w mean (median)**	**Silent π mean (median)**	**References**
Autosomes[a]	23	0.01466 (0.0128)	0.0134 (0.0102)	0.0147 (0.0128)	0.0134 (0.0102)	(HAMBLIN and AQUADRO 1999b; MACHADO *et al.* 2002; SCHAEFFER *et al.* 2003; SCHAEFFER *et al.* 2001)
X	8	0.0228 (0.0179)	0.0256 (0.019)	0.0228 (0.0179)	0.0154 (0.013)	(KOVACEVIC and SCHAEFFER 2000; MACHADO *et al.* 2002)
Dot[b]	20	0.00096 (0.00075)	0.00038 (0.00023)	0.00171 (0.00128)	0.00056 (0.00029)	This study
Y-to-dot	11	0.00094 (0.00039)	0.00025 (0.00037)	0.00166 (0.00126)	0.00053 (0.00037)	This study
Dot, non-Y	9	0.00098 (0.00075)	0.00054 (0.00015)	0.00177 (0.0013)	0.00060 (0.00020)	This study

The dot chromosome has significantly less diversity than the autosomes overall ($P_{\theta w}$ =1.061x10^{-6} ,P_{π}=2.524x10^{-8}, MWU; Table 4.4), including when comparing genes just in the Y-to-dot translocation ($P_{\theta w}$=1.043x10^{-5} P_{π}=3.4x10^{-6}, MWU; Table 4.4) and the

non-Y parts of the dot ($P_{\theta w}$=3.127x10^{-5} , P_{π}=1.9x10^{-5}, MWU; Table 4.4). The dot also has significantly less diversity than the X chromosome ($P_{\theta w}$=4.844x10^{-5} , P_{π}=4.84x10^{-5}, Table 4.4). These patterns are mimicked at silent sites: the Y-to-dot translocation and the rest of the dot have similar levels of silent variation ($P_{\theta w}$=0.7289, P_{π}=0.8173, MWU; Table 4.4; Figure 4.2). Silent site diversity on the dot chromosome is significantly less than the autosomes ($P_{\theta w}$=1.061x10^{-7} , P_{π}=2.524x10^{-8}, MWU; Table 4.4; Figure 4.2), and the X chromosome ($P_{\theta w}$=4.84x10^{-5}, P_{π}=4.84x10^{-5}, MWU; Table 4.4; Figure 4.2).

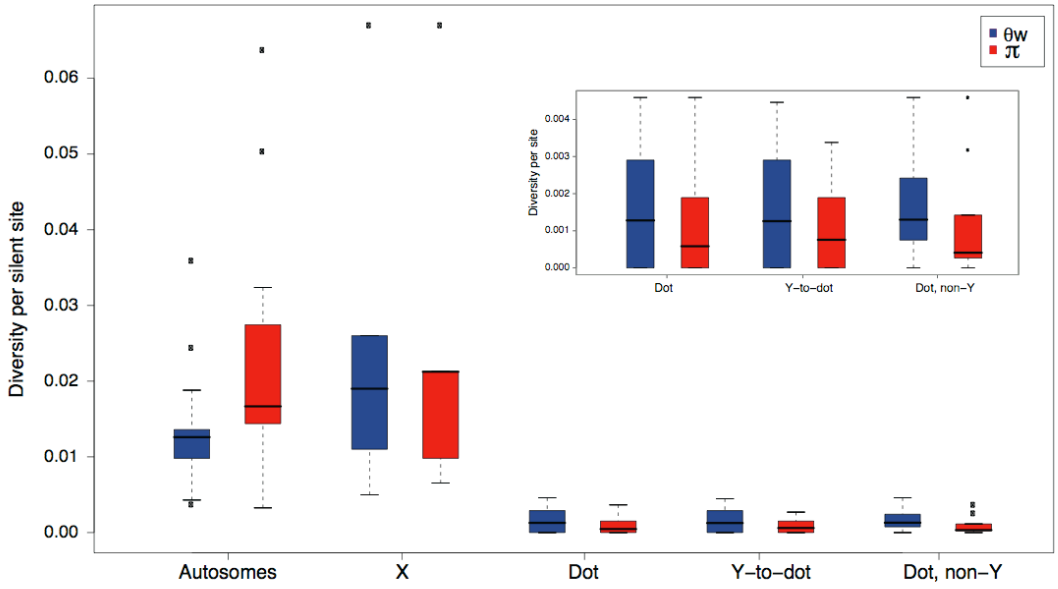


Figure 4.2. Boxplot of diversity per silent site in *D. pseudoobscura*. Shown are the medians and interquartile ranges of θ_w and π for the autosomes, the X chromosome, and the dot chromosome. The dot is further broken up into the Y-to-dot translocated region and the rest of the dot, not involved in the translocation (dot, non-Y). Just the dot chromosome and its components are shown in the inset.

The level of silent polymorphism surveyed on the dot chromosome is lower than expected based on reported levels of silent autosomal polymorphism under a neutral coalescent model ($\pi_{obs}<\pi_{sim}$, P=0.0002). Silent site divergence on the dot chromosome

is not significantly different than divergence on the autosomes (P=0.5485, MWU) and X chromosome (P=0.8694, MWU) in *D. pseudoobscura* (Table 4.5; Figure 4.3), indicating that the lower level of polymorphism on the dot chromosome is not likely due to a lower neutral mutation rate.

Table 4.5. Summary of average pairwise divergence per silent site between *D. pseudoobscura* and *D. miranda* for the autosomes, X, dot chromosome, Y-to-dot and dot, non-Y sequences. [a]10 loci from second chromosome (HAMBLIN and AQUADRO 1999b; MACHADO *et al.* 2002); 9 loci from third chromosome (MACHADO *et al.* 2002; SCHAEFFER *et al.* 2003); 4 loci from fourth chromosome (MACHADO *et al.* 2002; SCHAEFFER *et al.* 2001). [b] includes 11 loci from the Y-to-dot translocation and 9 loci from the rest of the dot, outside of the translocation (dot, non-Y).

	Divergence mean (median)	**References**
Autosomes[a]	0.0296 (0.0242)	(HAMBLIN and AQUADRO 1999b; MACHADO *et al.* 2002; SCHAEFFER *et al.* 2003; SCHAEFFER *et al.* 2001)
X	0.0474 (0.0343)	(KOVACEVIC and SCHAEFFER 2000; MACHADO *et al.* 2002)
Dot[b]	0.0322 (0.0303)	This study
Y-to-dot	0.03613 (0.0430)	This study
Dot, non-Y	0.0273 (0.0274)	This study

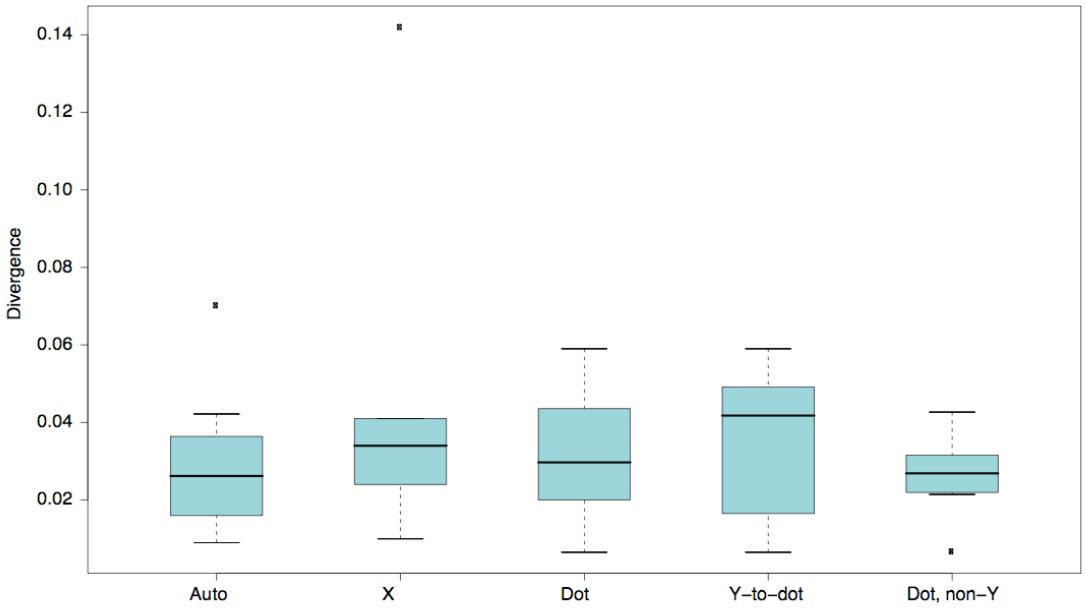


Figure 4.3. Divergence on the autosomes, X and dot chromosome of *D. pseudoobscura.* Boxplot showing the median and interquartile ranges of pairwise divergence between *D. pseudoobscura* and *D. miranda* on the 23 autosomal loci, 8 X-linked loci, and 20 dot chromosome loci. The dot chromosome is further broken down into the 11 fragments from the Y-to-dot translocation and those on the part of the dot chromosome not involved in the translocation (dot, non-Y).

Evidence for recombination

The dot chromosome has long been regarded as a non-recombining chromosome: in thousands of meioses, no crossover events were observed. More recently, exchange between homologous dot chromosomes has been recorded in *D. melanogaster,* although exactly how these events are resolved is unknown (HUGHES *et al.* 2009). Population genetic analyses in *D. melanogaster* and *D. simulans* have documented historical recombination events, leading to the conclusion that the dot chromosome shows evidence for very low levels of recombination by crossing over and gene conversion (JENSEN *et al.* 2002; WANG *et al.* 2002; WANG *et al.* 2004) (ARGUELLO *et al.* submitted). To determine whether the *D. pseudoobscura* dot

chromosome has evidence of ancestral recombination events, we calculated the minimum number of recombination events as R_m (HUDSON and KAPLAN 1985) and R_h (MYERS and GRIFFITHS 2003). Our values of R_m and R_h were 2 and 7, respectively, indicating that there is evidence for recombination on the *D. pseudoobscura* dot chromosome. However, the frequency of recombination events found in this sample are low: the R_m density is 0.0031/kb/chromosome (2 events per 10kb in 64 chromosomes) with a lower bound of 1.918×10^{-5} (2 events per 1630 kb in 64 chromosomes) and the R_h density is 0.0194 (7 events per 10kb in 64 chromosomes) with a lower bound of 6.71×10^{-5} (7 events per 1630 kb in 64 chromosomes). The R_m density found in this sample of *D. pseudoobscura* is lower than R_m density in *D. simulans* (0.0451 R_m/kb/chromosome, WANG *et al.* 2004 and a lower bound of 0.024 R_m/kb/chromosome ARGUELLO *et al.* submitted) and *D. melanogaster* dot chromosomes (0.005 R_m/kb/chromosome, WANG *et al.* 2002 and a lower bound of 0.010 R_m/kb/chromosome ARGUELLO *et al.* submitted). Because R_m depends on sample size, and our sample size differs from that of other studies, it is difficult to compare to other regions of the *D. pseudoobscura* genome and other species.

The regions of the dot sampled show a high level of linkage disequilibrium (LD), with 57.6% pairs showing very high LD (449/780 pairs with $0.9 < |D'| < 1$; Figure 4.4A).

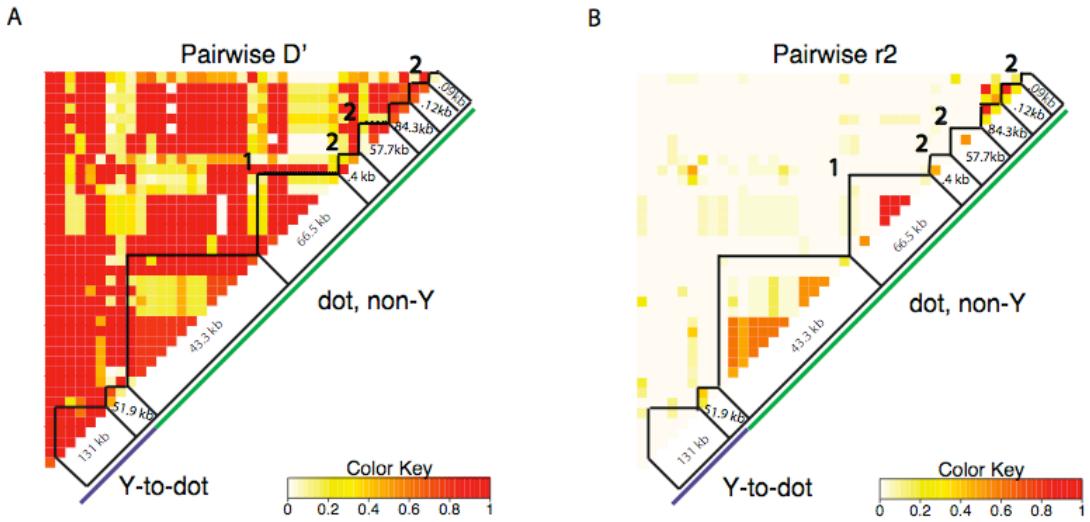


Figure 4.4. Linkage disequilibrium on the dot chromosome. Plot shows amount of linkage disequilibrium assessed by: (A) D' and (B) r^2. The boxed areas correspond to predicted haplotypes calculated in RecMin (see Materials and Methods). The length of each haplotype block ranges from 88 bp to 131 kb. The inferred location and number of recombination events is found above the boxed in haplotype blocks. All of these inferred events are predicted to be in the region of the dot chromosome not involved in the Y-to-dot translocation.

Only 16% (125/780) of pairwise comparisons between SNPs rejected the null hypothesis of independence, however (Figure 4.4B). The product moment correlation between distance between SNPs and r^2 (-0.2728) is statistically significant (P=8.9x10^{-15}; Figure 4.5), which supports this chromosome experiencing recombination events as recombination is expected to break down LD over long distances.

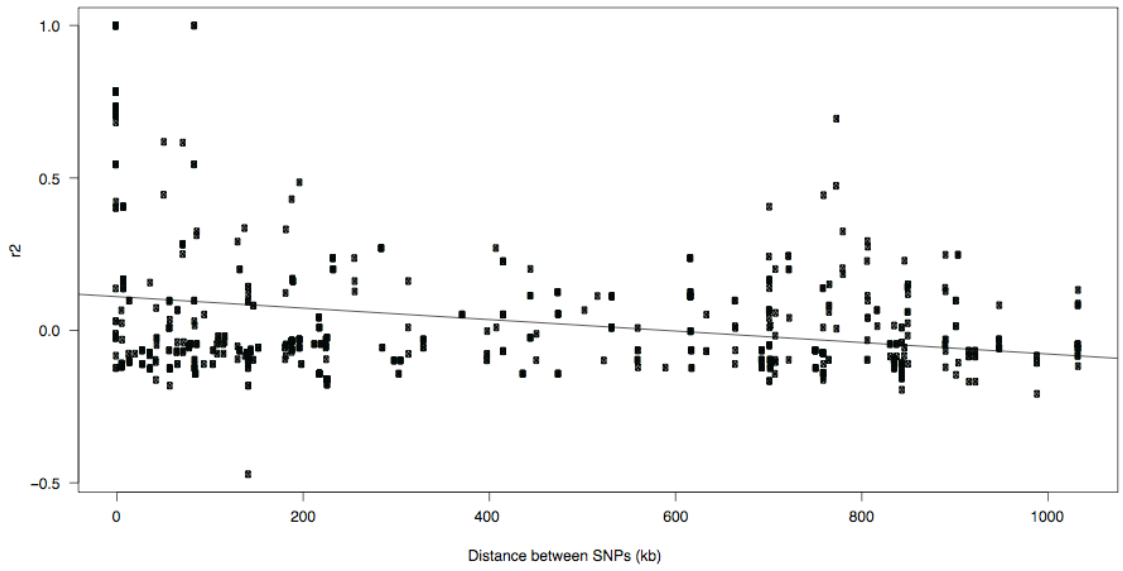


Figure 4.5. The relationship between distance between SNPs and r^2. The distance between SNPs and r^2 are significantly negatively correlated (-0.272), indicating that this chromosome has experienced recombination.

We find both very long (between 43 kb and 130 kb; Figure 4.4) and very short haplotype blocks (88-124 bp; Figure 4.4). The very long recombination tracts are unlikely to be explained by gene conversion, since we estimate an average tract length of 150 bp in *D. pseudoobscura*.

In order to examine whether the observed recombination events can be attributed to crossover events or gene conversion, we estimated the population level recombination parameter $\rho = 4Nr$, where N is the effective population size and r is the rate of recombination per base pair per generation, and the relative rate of gene conversion, $f = g/\rho$, where g is the probability of a gene conversion event per base pair. We used a composite likelihood approach to search over a grid of values of ρ and f, for several different conversion tract lengths. The maximum composite likelihood for the different conversion tract lengths was at 150 bp, giving an estimate $\rho = 0.000092$ and $f = 170.2$ (Table 4.6), indicating that there is support for a model involving both

recombination and gene conversion. In all cases, the composite likelihood for the model assuming no gene conversion was lower than a model with both gene conversion and recombination for all tract lengths examined (Table 4.6). Our estimates of ρ and f are similar to estimates on the dot chromosome of *D. simulans* (WANG *et al.* 2004) and slightly higher than estimates observed in *D. melanogaster* (ARGUELLO *et al.* submitted). This is consistent with *D. pseudoobscura* having a larger effective population size (RILEY *et al.* 1989; SCHAEFFER *et al.* 1987; SCHAEFFER and MILLER 1992b) and overall higher rates of recombination (HAMBLIN and AQUADRO 1999b) than *D. melanogaster.* The average ρ of autosomal loci ranges from 0 (*4002*) to 0.1633 (*4003*) with an average ρ of 0.0516 (MACHADO *et al.* 2002). On the X chromosome, ρ varies from 0 (*X010*; MACHADO *et al.* 2002) to 0.107 (*Est-5*; KOVACEVIC and SCHAEFFER 2000) with an average ρ of 0.0474. The average ρ on the dot chromosome is estimated at three orders of magnitude lower than the X and autosomes, however the estimates from the X and autosomes do not consider a model with gene conversion. These results indicate that the dot chromosome of *D. pseudoobscura* has a very low rate of recombination, with the majority of events being resolved as gene conversions.

Table 4.6. Composite likelihood analysis of recombination rate (ρ) and the ratio of crossovers to gene conversion (f). The maximum composite likelihood estimate of ρ and f are given for different average gene conversion tract lengths. The first row is a model with no gene conversion. The model with the highest likelihood is in bold.

Mean tract length	ρ	f	lnL
0 (no gene conversion)	0.000139	0	-3795.031585
50	0.000092	470.588235	-3785.959551
100	0.000092	242.803504	-3785.767135
150	**0.000092**	**170.212766**	**-3785.709478**
200	0.000092	132.665832	-3785.742768
250	0.000092	111.389237	-3785.807385
300	0.000092	106.382979	-3785.907192
350	0.000090	100.125156	-3786.022639
400	0.000090	87.609512	-3786.136024
450	0.000090	78.848561	-3786.242966
500	0.000090	71.339174	-3786.344305
550	0.000090	67.584481	-3786.442007
600	0.000090	61.326658	-3786.536484

Modeling the demographic history of D. pseudoobscura

For all but one of the loci sampled (GA15170), θ_w is greater than π and Tajima's D is negative (Figure 4.6). Of the 16 loci where segregating sites were found, seven of these loci have a significantly negative Tajima's D (D_{Taj}), as does the concatenated dataset (D_{Taj}=-2.13; P=0.001) under a standard neutral model. Thus it appears that the frequency spectra of loci on the dot chromosome are significantly skewed towards rare variants.

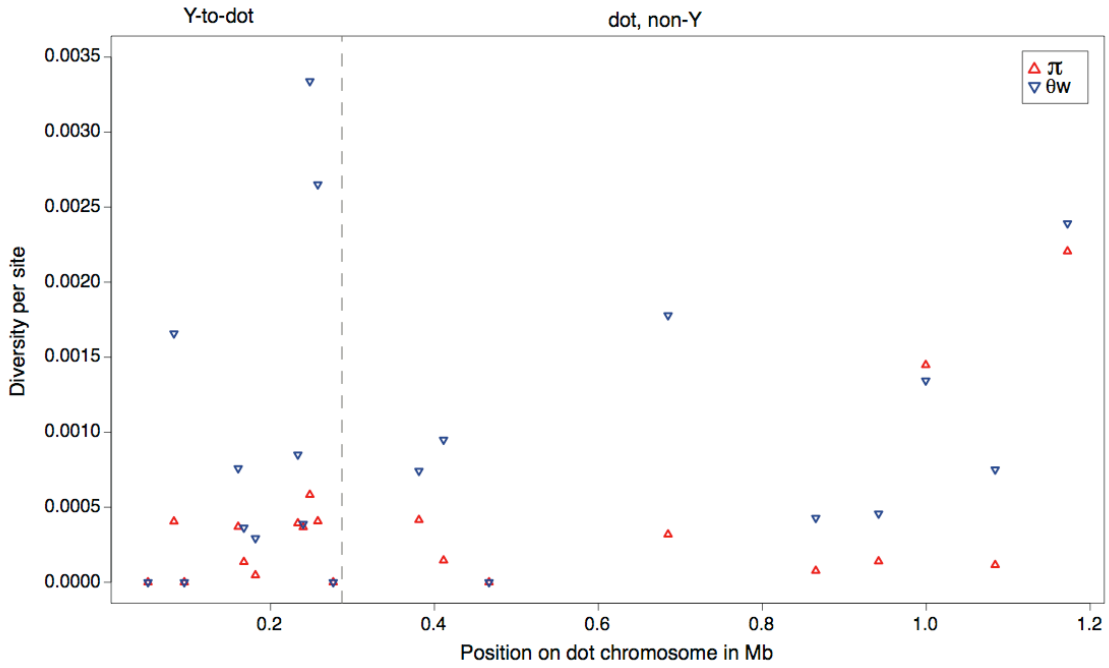


Figure 4.6. Diversity per site for each gene fragment. The estimates of π and θ_w are plotted at the position in Mb of each gene fragment on the dot chromosome. The first 11 loci are from the Y-to-dot translocated region (left of the dotted line) and the last 9 are from the rest of the dot chromosome (dot, non-Y to the right of the dotted line).

An excess of rare variants segregating in a population is consistent with several evolutionary scenarios. Selective sweeps can both reduce levels of neutral variation through the rapid fixation of a beneficial allele sweeping out linked neutral variability and skew the frequency spectrum of mutations towards rare alleles (TAJIMA 1989). Background selection also leads to reduced levels of neutral variability and a skew in the frequency spectrum towards rare alleles, especially for deleterious mutations with a weaker effect (CHARLESWORTH *et al.* 1995). Recently it was shown that mutations with a larger selection coefficient can also skew the frequency spectrum towards rare variants in regions of low recombination (KAISER and CHARLESWORTH 2009). The demographic history of a species can also have a strong effect on the frequency spectrum of mutations: rapid population expansion and recovery from a recent

population bottleneck are both expected to lead to an excess of rare variants. A skew in the frequency spectra of nearly all loci sampled in *D. pseudoobscura* suggests that this species has a history of population expansion (HAMBLIN and AQUADRO 1999a; KOVACEVIC and SCHAEFFER 2000; MACHADO *et al.* 2002; SCHAEFFER *et al.* 2003).

We attempted to model the demographic history of *D. pseudoobscura* by comparing nested models of constant population size and simple models of exponential population growth with population bottlenecks using statistics summarized from non-coding regions of six autosomal loci (MACHADO *et al.* 2002). We found that a demographic model of exponential population growth has an order of magnitude higher acceptance rate (0.00688) than one of constant population size (0.00054), suggesting that *D. pseudoobscura* has a history of population expansion. We were unable to distinguish between models of population expansion and population bottlenecks based on acceptance rate alone (bottleneck acceptance rate was 0.00673), however the joint posterior distribution of the population size before the bottleneck and the growth rate indicate that the severity of the bottleneck was so low ($\sim 3.67 \times 10^{-25}$) that the most appropriate bottleneck model actually converges on a model of simple population expansion. Our simulations indicate that the maximum *a posteriori* estimate (MAP) of the time the population stopped growing was 0.0835 $4N_e$ generations ago (95% t_{end} C.I. 0.0074 - 1.415; Figure 4.7) and we have little information about the time the population started growing (MAP $t_{begin} = 1.574\ 4N_e$ generations ago, 95% C.I. t_{begin} 0.2344 - 2.6577; Figure 4.8A; MAP λ=72.61, 95% C.I. λ 8.2288 - 97.6384; Figure 4.8B). The confidence intervals are quite wide, likely due to the small number of loci used for this inference.

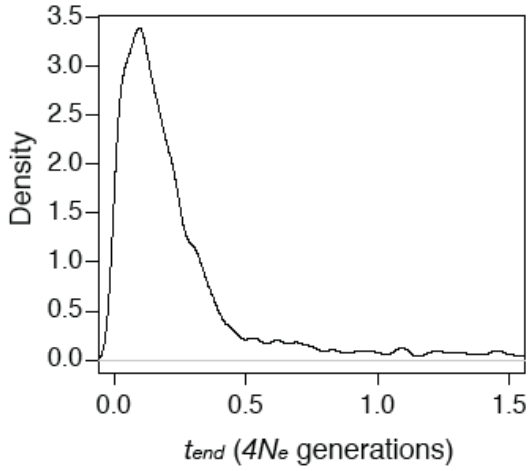


Figure 4.7. The marginal posterior distribution for the time the population stopped growing (t_{end}) in $4N_e$ generations.

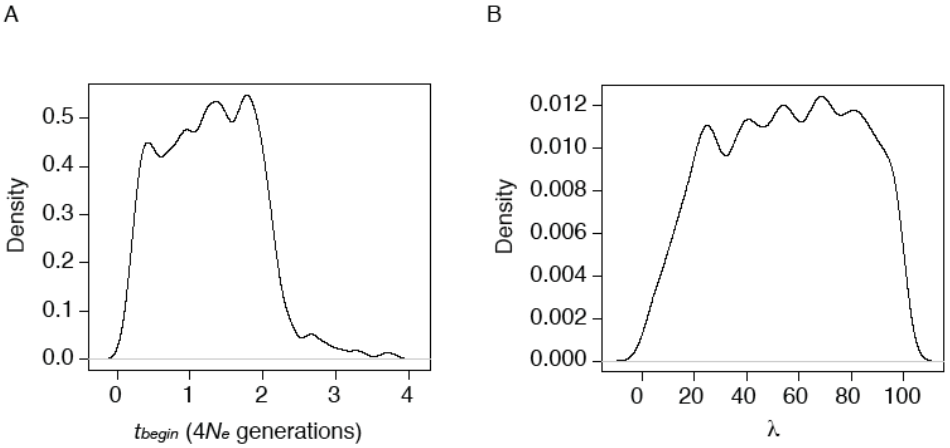


Figure 4.8. Marginal posterior distributions of t_{begin} and λ. (A) Marginal posterior distribution on the time the population started growing (t_{begin}) $4N_e$ generations in the past. (B) Marginal posterior distribution on the population growth rate (λ).

Assuming an N_e of 4.5×10^6 for *D. pseudoobscura* (SCHAEFFER 1995), and 5 generations per year, the population expansion ended approximately 300 thousand years ago.

It is not our intention to infer an accurate model for the demographic history of this species because of the limited amount of polymorphism data available in *D. pseudoobscura*. Instead, our intention is to determine whether a demographic model involving population expansion can explain the reduction in variation and skew in the frequency spectrum that we observe on the dot chromosome. We generated simulated data from the joint posterior distribution of the parameters of the exponential growth model that fit to the autosomal data. In all cases, the summary statistics of the autosomal loci used for the inference (see Materials and Methods) were well within the range simulated. Using the data simulated from the joint posterior distribution of the exponential growth model, we determined that the dot chromosome (using the concatenated dataset) shows a significant reduction in variability even under a model of population expansion that fits the autosomal data used in the simulation (π_{dot}=0.00062, $P < 0.0265$). Because a relatively broad range of demographic parameters fit the autosomal *D. pseudoobscura* data under our rejection sampling scheme, this is a conservative test. The degree of the skew in the frequency spectrum towards rare alleles on the dot chromosome (Tajima's D_{dot}= -2.13) is also not expected under the model of population expansion ($P = 0.0014$). This suggests that variability on the dot chromosome has been affected by more than just the demographic history of the species, and further implicates the action of natural selection.

Evidence for selection

There is no evidence for positive selection in the protein coding regions of any of the loci surveyed here: McDonald-Kreitman tests (McDONALD and KREITMAN 1991) for the equality of the ratio of silent to replacement polymorphism and the ratio of silent to replacement divergence between species were not significant for any gene region (Appendix Table 4.1). An analysis of the constraint on protein evolution

reveals that the gene regions sampled here are constrained at the protein level and do not show an excess of nonsynonymous divergence. These tests have little power in our dataset, however, because of the low number of nonsynonymous sites (and corresponding low number of segregating sites) and low nonsynonymous divergence between *D. pseudoobscura* and *D. miranda.*

Our hypothesis is that the Y-to-dot translocated region may have experienced recurrent selective sweeps as the introns of the region shrank in size, and so we do not necessarily expect to see signatures of selection at the protein levels. To test our hypothesis, we simulated selective sweeps in a population that is expanding exponentially, according to the demographic model that fit the autosomal data described above. These simulations assume that there is no recombination at the sampled locus, therefore we only use the concatenated Y-to-dot translocation dataset. Genealogies were simulated under a population expansion with sweep model and were accepted conditional on the number of segregating sites, Tajima's *D,* and the number of singletons. The acceptance rate for the summary statistics from silent sites of the concatenated Y-to-dot chromosome dataset was 5.45×10^{-5}. Using the approximate Bayesian computation method and rejection sampling conditional on summaries of the Y-to-dot frequency spectrum, we estimate the time since the last selective sweep to be 0.0632 (95% C.I. t_{sweep} 0.01288 – 1.914) $4N_e$ generations (Figure 4.9).

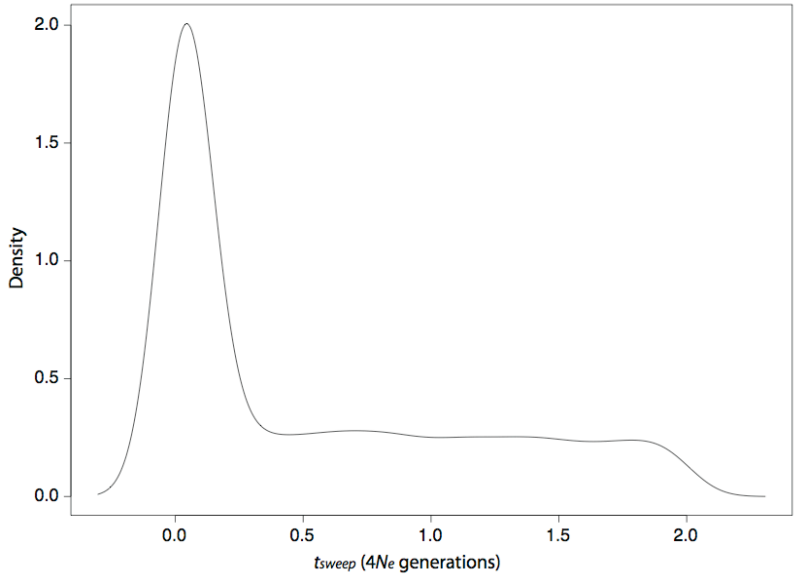


Figure 4.9. The marginal posterior distribution for the time since the last selective sweep (t_{sweep}) in $4N_e$ generations.

Assuming an N_e of 4.5 million (SCHAEFFER 1995) and 5 generations per year, the last sweep occurred approximately 228,000 years ago. One should not place too much confidence in the estimates of time in millions of years due to the uncertainty in the effective size of *D. pseudoobscura* (see HAMBLIN and AQUADRO 1999b) and the number of generations per year. These simulations assume complete linkage among the Y-to-dot chromosome loci. By restricting ourselves to just the dot chromosome loci in the Y-to-dot translocation, where there are no detected recombination events, we improve our estimate of the time since the last selective sweep over including all of the dot chromosome loci (Appendix Figure 4.1). We have detected ancestral recombination events on the *D. pseudoobscura* dot chromosome, all in the region of the dot not associated with the translocation (Figure 4.4), however the overall rate of recombination on this chromosome is estimated to be very low (see below). By our estimation, this sweep occurred since the split of *D. pseudoobscura* and *D. persimilis*.

Purifying selection on the dot

Hill-Robertson interference in regions of low recombination like the dot chromosome is expected to interfere with selection on synonymous sites and thus lead to lower levels of codon bias (BETANCOURT AND PRESGRAVES 2002; HADDRILL *et al.* 2007, LARRACUENTE *et al.* 2008). Indeed, purifying selection appears to be less efficient on the dot chromosome (HADDRILL *et al.* 2007; SINGH *et al.* 2008), and it is possible that this is interference is due to selective sweeps. It was previously shown that the dot chromosome of *D. pseudoobscura* has significantly less codon bias (measured by the frequency of optimal codons, or FOP) than the autosomes (mean $FOP_{dot} = 0.2921$, median $FOP_{dot} = 0.2945$; mean $FOP_{auto} = 0.5370$, median $FOP_{auto} = 0.538$; MWU $P = 1.821 \times 10^{-6}$; SINGH *et al.* 2008) in *D. pseudoobscura*. As expected, this is also true of the loci sampled here (mean $FOP_{dot} = 0.2842$, median $FOP_{dot} = 0.2760$; MWU $P = 1.526 \times 10^{-5}$).

Background selection and models of local adaptation are expected to increase Fst where there are low levels of variation in a structured population (CHARLESWORTH *et al.* 1997; STEPHAN *et al.* 1998). All loci surveyed here have low Fst values between populations (Appendix Table 4.2), indicating that populations are relatively homogenous. Population homogeneity is a signature of positive selection sweeping selected variants across subpopulations, however there is considerable gene flow in *D. pseudoobscura* (KOVACEVIC and SCHAEFFER 2000; RILEY et al. 1989; SCHAEFFER et al. 2003; SCHAEFFER and MILLER 1992a), which also leads to population homogeneity. Therefore, based on population structure and levels of codon bias, we cannot exclude the possibility that background selection is responsible for the reduction in variation on the dot. However, because there are relatively few genes in the translocated region, there is a relatively small target of purifying selection. Moreover, the overall relaxed constraints suggest that purifying selection is less

efficacious on the dot chromosome, and may not be sufficient to generate enough background selection to result in the patterns of variability that we observe. We are currently exploring this possibility.

Discussion

Patterns of variation and the inference of recombination on the dot

The dot chromosome was thought to be completely non-recombining in *D. melanogaster* (ASHBURNER 1989), although more recent empirical studies have identified that recombination (in the form of gene conversion and possibly a low level of crossing over) has affected the dot chromosomes of *D. melanogaster* and *D. simulans* historically (ARGUELLO *et al.* submitted; JENSEN *et al.* 2002; WANG *et al.* 2002; WANG *et al.* 2004). Polymorphism on the dot chromosome of *D. pseudoobscura* indicates that recombination has affected this chromosome, as it has in the *melanogaster* subgroup. We estimate the recombination rate to be very low (ρ = 0.000092) and gene conversion may be a more frequent occurrence, with as many as f = 170.2 conversion events for each crossover. The low level of recombination on the dot is expected to affect the ability of natural selection to work efficiently on this chromosome. Indeed, the efficacy of selection appears reduced on the dot chromosome as the dot, in general, has higher nonsynonymous divergence than the other chromosomes and lower levels of codon bias (ARGUELLO *et al.* submitted; BETANCOURT *et al.* 2009; HADDRILL *et al.* 2007; SINGH *et al.* 2008). We see the same patterns in the loci we surveyed here.

Selection on the dot

The dot chromosome of *D. pseudoobscura* is unique in that it has a translocation of a large fraction of the ancestral *Drosophila* Y chromosome (LARRACUENTE submitted). The formerly Y-linked genes are now passed through both males and females and present in twice the dosage than they were on the Y

chromosome. The most intriguing aspect of this translocation is the 10-fold reduction in intron size (CARVALHO and CLARK 2005). Because of its male-male transmission and lack of recombination, the Y chromosome has a reduced efficacy of selection (reviewed in CHARLESWORTH and CHARLESWORTH 2000), allowing the expansion of introns to the Mb size range. It is thought that transcribing megabase pair-long introns may come at a cost to the cell (CARVALHO and CLARK 1999), especially in an active, gene-dense part of the genome (PRACHUMWAT *et al.* 2004). While the size of the introns on the Y chromosome may have escaped selection, the new location on the dot chromosome, with four times the N_e and possibly a very low level of recombination, allowed for the shrinking of introns.

We find significantly reduced levels of variation on the dot chromosome and a skew in the frequency spectra towards rare alleles. Though the demographic history of *D. pseudoobscura* likely has involved population expansion, this event alone does not seem to explain the reduction in variation (Table 4.3), indicating that natural selection is likely involved. There is no evidence that the loci surveyed in this study experience positive selection at the protein level, as indicated by an analysis of the ratio of non-synonymous to synonymous divergence and McDonald Kreitman tests comparing the ratios of silent to replacement polymorphism to silent to replacement divergence. One plausible explanation for the patterns of variation on the dot of *D. pseudoobscura* is that positive selection favoring the deletions in introns cause recurrent selective sweeps acting to shorten the region over time, reducing levels of neutral variability. If this chromosome were subject to strong selective sweeps due to the shortening of introns, the lack of evidence of positive selection at dot-linked loci may be expected, due to interference generated by the sweeps. This model of selective sweeps also predicts a significant skew in the frequency spectra toward rare alleles, which we find for the concatenated dot chromosome dataset and for 7 out of 16 loci

(where nucleotide variation was found) under a standard neutral model and 4 out of 16 loci under a population expansion model (Table 4.3). All loci surveyed on the dot chromosome have a significant reduction in silent variation under both a standard neutral model and one of population expansion (Table 4.3). Selective sweeps can explain the patterns of variation on the dot chromosome. If this region experienced selective sweeps, we estimate the time to the most recent selective sweep as 0.0632 (95% C.I. t_{sweep} 0.01288 – 1.914) $4N_e$ generations ago, or approximately 228,000 years ago.

Recurrent selective sweeps do not offer the only explanation for patterns of diversity we see on the dot chromosome of *D. pseudoobscura*. A model of temporally fluctuating selection can reduce levels of genetic diversity, however this is not expected to result in a large skew in the frequency spectrum towards rare variants. Background selection does offer a viable option: purifying selection against strongly deleterious mutations can reduce levels of neutral variability (CHARLESWORTH *et al.* 1995) and skew the frequency spectrum toward rare alleles in regions of low recombination (KAISER and CHARLESWORTH 2009). While patterns of low Fst on the dot chromosome support a model of selective sweeps better than a model of background selection, Fst is only expected to be elevated under background selection if the population is structured (CHARLESWORTH *et al.* 1997; STEPHAN *et al.* 1998). Extensive gene flow among *D. pseudoobscura* populations has been described at other autosomal loci and so this assumption may be violated.

The effect of background selection decreases as the rate of recombination approaches the strength of selection against a deleterious mutation; therefore in regions of low recombination, segregating weakly deleterious mutations can significantly affect the frequency spectrum of mutations (CHARLESWORTH *et al.* 1995). Pervasive selection against weakly deleterious mutations can also reduce levels of

neutral variability and skew the frequency spectrum toward rare alleles. One possible source of these deleterious mutations is transposable elements (TEs). Selection against TEs is probably very weak (*sh* is approximately $2x10^{-4}$; CHARLESWORTH *et al.* 1995) and could play a role on the dot chromosome. Upon moving to the dot chromosome, the Y chromosome brings a high density of transposable elements among other repetitive elements. However, many of these transposable elements on the Y chromosome are truncated and degenerated forms of relic transposable elements (KUREK *et al.* 2000). Nonetheless, repetitive elements such as TEs tend to accumulate in heterochromatic regions like the dot chromosome (DIMITRI 1997; JUNAKOVIC *et al.* 1998) and could potentially affect the evolutionary trajectory of the chromosome.

Intron evolution in the Y-to-dot translocated region

The translocation of the Y chromosome to the dot chromosome occurred between 12.7 and 20.8 million years ago: both the X-D fusion and Y-A translocation are found in *D. affinis* and *D. azteca* (clade split from *D. pseudoobscura* ancestor between 12.7 and 14.9 Myr ago; GAO *et al.* 2007), but are not found in the *obscura* group species *D. bifasciata* and *D. guanche* (CARVALHO and CLARK 2005; *D. bifiasciata* ancestor split from *D. pseudoobscura* ancestor between 17.8 and 20.8 Myr ago; GAO *et al.* 2007). The large introns of the Y-to-dot translocated genes shrank ~10-fold in 12.7 to 20.8 million years. The megabase-sized introns on the *D. melanogaster* and *D. hydei* Y chromosomes form three large lampbrush loop-like structures consisting mostly of satellite DNA and TE repeats (KUREK *et al.* 2000). The lambrush loop-like structures result from the transcription of the large introns and are clearly visible in primary spermatocytes. Interestingly, cytological evidence in *D. pseudoobscura* primary spermatocyte shows three lampbrush-loop like structures (PIERGENTILI 2007). It is unknown what the sources of these loops are. They could be from new male fertility factors on the *D. pseudoobscura* Y chromosome, which may

be derived from a degenerated Muller D element. Another possibility is that are from the formerly Y-linked genes the translocated to the dot chromosome. The former hypothesis may be more likely because the introns of the formerly Y-linked genes are 10-fold smaller on the dot chromosome, and one of the genes which forms lampbrush loop-like structures on the *D. melanogaster* and *D. hydei* Y chromosomes, *kl-5*, is on the second chromosome of *D. pseudoobscura.* It appears that *kl-5* translocated to the Y chromosome twice, independently and that the ancestral location of this gene is autosomal (KOERICH *et al.* 2008). Furthermore, different introns are expanded in *D. melanogaster* and *D. hydei* for these genes (KUREK *et al.* 2000), highlighting the dynamic properties of heterochromatin expansion on the Y chromosome. Indeed, there is rapid turnover of repetitive sequences on the Y of *Drosophila: D. melanogaster* and *D. simulans,* show different TE families and repeats on their Y chromosomes (JUNAKOVIC *et al.* 1998), despite their close relationship.

REFERENCES

ARGUELLO, R., T. KADO, H. INNAN, W. WANG, M. LONG, Recombination yet inefficient selection along the *D. melanogaster* subgroup's fourth chromosome. Genome Res**.** Submitted.

ASHBURNER, M., 1989 *Drosophila: A Laboratory Handbook*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

BENJAMINI, Y., HOCHBERG, Y., 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B **57:** 289-300.

BETANCOURT, A. J., J. J. WELCH and B. CHARLESWORTH, 2009 Reduced effectiveness of selection caused by a lack of recombination. Curr Biol **19:** 655-660.

BETANCOURT, A. J., and D. C. PRESGRAVES, 2002 Linkage limits the power of natural selection in *Drosophila*. Proc Natl Acad Sci U S A **99:** 13616-13620.

BRIDGES, C. B., 1935 The mutants and linkage data of chromosome four of *Drosophila melanogaster*. Biol Zh **4:** 401-420.

CARVALHO, A. B., and A. G. CLARK, 1999 Intron size and natural selection. Nature **401:** 344.

CARVALHO, A. B., and A. G. CLARK, 2005 Y chromosome of *D. pseudoobscura* is not homologous to the ancestral *Drosophila* Y. Science **307:** 108-110.

CHARLESWORTH, B., and D. CHARLESWORTH, 2000 The degeneration of Y chromosomes. Philos Trans R Soc Lond B Biol Sci **355:** 1563-1572.

CHARLESWORTH, B., M. T. MORGAN and D. CHARLESWORTH, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289-1303.

CHARLESWORTH, B., M. NORDBORG and D. CHARLESWORTH, 1997 The effects of local selection, balanced polymorphism and background selection on

equilibrium patterns of genetic diversity in subdivided populations. Genet Res **70:** 155-174.

CHARLESWORTH, D., B. CHARLESWORTH and M. T. MORGAN, 1995 The pattern of neutral molecular variation under the background selection model. Genetics **141:** 1619-1632.

DIMITRI, P., 1997 Constitutive heterochromatin and transposable elements in *Drosophila melanogaster*. Genetica **100:** 85-93.

DROSOPHILA 12 GENOMES CONSORTIUM, 2007 Evolution of Genes and Genomes on the *Drosophila* Phylogeny. Nature **450:** 203-218.

FELSENSTEIN, J., 1974 The evolutionary advantage of recombination. Genetics **78:** 737-756.

GAO, J. J., H. A. WATABE, T. AOTSUKA, J. F. PANG and Y. P. ZHANG, 2007 Molecular phylogeny of the *Drosophila obscura* species group, with emphasis on the Old World species. BMC Evol Biol **7:** 87.

GATTI, M., and S. PIMPINELLI, 1983 Cytological and genetic analysis of the Y chromosome of *Drosophila melanogaster*. I. Organization of the fertility factors. Chromosoma **88:** 349-373.

HADDRILL, P. R., D. L. HALLIGAN, D. TOMARAS and B. CHARLESWORTH, 2007 Reduced efficacy of selection in regions of the *Drosophila* genome that lack crossing over. Genome Biol **8:** R18.

HAMBLIN, M. T., and C. F. AQUADRO, 1999a DNA sequence variation and the recombinational landscape in *Drosophila pseudoobscura*: a study of the second chromosome. Genetics **153:** 859-869.

HAMBLIN, M. T., and C. F. AQUADRO, 1999b DNA sequence variation and the recombinational landscape in *Drosophila pseudoobscura*: a study of the second chromosome. Genetics **153:** 859-869.

HILL, W. G., and A. ROBERTSON, 1966 The effect of linkage on limits to artificial selection. Genet Res **8:** 269-294.

HUDSON, R. R., 2001 Two-locus sampling distributions and their application. Genetics **159:** 1805-1817.

HUDSON, R. R., and N. L. KAPLAN, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147-164.

HUGHES, S. E., W. D. GILLILAND, J. L. COTITTA, S. TAKEO, K. A. COLLINS *et al.*, 2009 Heterochromatic threads connect oscillating chromosomes during prometaphase I in *Drosophila* oocytes. PLoS Genet **5:** e1000348.

JENSEN, M. A., B. CHARLESWORTH and M. KREITMAN, 2002 Patterns of genetic variation at a chromosome 4 locus of *Drosophila melanogaster* and *D. simulans*. Genetics **160:** 493-507.

JOHANSSON, A. M., P. STENBERG, C. BERNHARDSSON and J. LARSSON, 2007a Painting of fourth and chromosome-wide regulation of the 4th chromosome in *Drosophila melanogaster*. EMBO J **26:** 2307-2316.

JOHANSSON, A. M., P. STENBERG, F. PETTERSSON and J. LARSSON, 2007b POF and HP1 bind expressed exons, suggesting a balancing mechanism for gene regulation. PLoS Genet **3:** e209.

JUNAKOVIC, N., A. TERRINONI, C. DI FRANCO, C. VIEIRA and C. LOEVENBRUCK, 1998 Accumulation of transposable elements in the heterochromatin and on the Y chromosome of *Drosophila simulans* and *Drosophila melanogaster*. J Mol Evol **46:** 661-668.

KAISER, V. B., and B. CHARLESWORTH, 2009 The effects of deleterious mutations on evolution in non-recombining genomes. Trends Genet **25:** 9-12.

KOERICH, L. B., X. WANG, A. G. CLARK and A. B. CARVALHO, 2008 Low conservation of gene content in the *Drosophila* Y chromosome. Nature **456:** 949-951.

KOVACEVIC, M., and S. W. SCHAEFFER, 2000 Molecular population genetics of X-linked genes in *Drosophila pseudoobscura*. Genetics **156:** 155-172.

KUREK, R., A. M. REUGELS, U. LAMMERMANN and H. BUNEMANN, 2000 Molecular aspects of intron evolution in dynein encoding mega-genes on the heterochromatic Y chromosome of *Drosophila sp*. Genetica **109:** 113-123.

LARRACUENTE, A. M., M.A.F. NOOR, A.G. CLARK, Translocation of Y-linked genes to the dot chromosome in *Drosophila pseudoobscura*. Genetics. submitted.

LARSSON, J., J. D. CHEN, V. RASHEVA, A. RASMUSON-LESTANDER and V. PIRROTTA, 2001 *Painting of fourth*, a chromosome-specific protein in *Drosophila*. Proc Natl Acad Sci U S A **98:** 6273-6278.

LARSSON, J., and V. H. MELLER, 2006 Dosage compensation, the origin and the afterlife of sex chromosomes. Chromosome Res **14:** 417-431.

MACHADO, C. A., and J. HEY, 2003 The causes of phylogenetic conflict in a classic *Drosophila* species group. Proc Biol Sci **270:** 1193-1202.

MACHADO, C. A., R. M. KLIMAN, J. A. MARKERT and J. HEY, 2002 Inferring the history of speciation from multilocus DNA sequence data: the case of *Drosophila pseudoobscura* and close relatives. Mol Biol Evol **19:** 472-488.

MCDONALD, J. H., and M. KREITMAN, 1991 Adaptive protein evolution at the Adh locus in *Drosophila*. Nature **351:** 652-654.

MYERS, S. R., and R. C. GRIFFITHS, 2003 Bounds on the minimum number of recombination events in a sample history. Genetics **163:** 375-394.

PIERGENTILI, R., 2007 Evolutionary conservation of lampbrush-like loops in drosophilids. BMC Cell Biol **8:** 35.

PRACHUMWAT, A., L. DEVINCENTIS and M. F. PALOPOLI, 2004 Intron size correlates positively with recombination rate in *Caenorhabditis elegans*. Genetics **166:** 1585-1590.

PRITCHARD, J. K., M. T. SEIELSTAD, A. PEREZ-LEZAUN and M. W. FELDMAN, 1999 Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol Biol Evol **16:** 1791-1798.

PRZEWORSKI, M., 2003 Estimating the time since the fixation of a beneficial allele. Genetics **164:** 1667-1676.

RIDDLE, N. C., and S. C. ELGIN, 2006 The dot chromosome of *Drosophila*: insights into chromatin states and their change over evolutionary time. Chromosome Res **14:** 405-416.

RILEY, M. A., M. E. HALLAS and R. C. LEWONTIN, 1989 Distinguishing the forces controlling genetic variation at the Xdh locus in *Drosophila pseudoobscura*. Genetics **123:** 359-369.

SCHAEFFER, S. W., 1995 Population genetics in *Drosophila pseudoobscura:* a synthesis based on nucleotide sequence data for the Adh gene. , pp. 329-352 in *Genetics of natural populations: the continuing importance of Theodosius Dobzhansky*, edited by L. LEVINE. Columbia University Press, New York.

SCHAEFFER, S. W., C. F. AQUADRO and W. W. ANDERSON, 1987 Restriction-map variation in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. Mol Biol Evol **4:** 254-265.

SCHAEFFER, S. W., M. P. GOETTING-MINESKY, M. KOVACEVIC, J. R. PEOPLES, J. L. GRAYBILL *et al.*, 2003 Evolutionary genomics of inversions in *Drosophila pseudoobscura*: evidence for epistasis. Proc Natl Acad Sci U S A **100:** 8319-8324.

SCHAEFFER, S. W., and E. L. MILLER, 1992a Estimates of gene flow in *Drosophila pseudoobscura* determined from nucleotide sequence analysis of the alcohol dehydrogenase region. Genetics **132:** 471-480.

SCHAEFFER, S. W., and E. L. MILLER, 1992b Molecular population genetics of an electrophoretically monomorphic protein in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. Genetics **132:** 163-178.

SCHAEFFER, S. W., C. S. WALTHOUR, D. M. TOLENO, A. T. OLEK and E. L. MILLER, 2001 Protein variation in Adh and Adh-related in *Drosophila pseudoobscura*. Linkage disequilibrium between single nucleotide polymorphisms and protein alleles. Genetics **159:** 673-687.

SCHEET, P., and M. STEPHENS, 2006 A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet **78:** 629-644.

SINGH, N. D., A. M. LARRACUENTE and A. G. CLARK, 2008 Contrasting the efficacy of selection on the X and autosomes in Drosophila. Molecular Biology and Evolution **25:** 454-467.

SMITH, J. M., and J. HAIGH, 1974 The hitch-hiking effect of a favourable gene. Genet Res **23:** 23-35.

STEPHAN, W., L. XING, D. A. KIRBY and J. M. BRAVERMAN, 1998 A test of the background selection hypothesis based on nucleotide data from *Drosophila ananassae.* Proc Natl Acad Sci U S A **95:** 5649-5654.

SUN, F. L., M. H. CUAYCONG, C. A. CRAIG, L. L. WALLRATH, J. LOCKE *et al.*, 2000 The fourth chromosome of *Drosophila melanogaster*: interspersed euchromatic and heterochromatic domains. Proc Natl Acad Sci U S A **97:** 5340-5345.

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

THORNTON, K., 2003 Libsequence: a C++ class library for evolutionary genetic analysis. Bioinformatics **19:** 2325-2327.

THORNTON, K., and P. ANDOLFATTO, 2006 Approximate Bayesian inference reveals evidence for a recent, severe bottleneck in a Netherlands population of *Drosophila melanogaster*. Genetics **172:** 1607-1619.

WANG, W., K. THORNTON, A. BERRY and M. LONG, 2002 Nucleotide variation along the *Drosophila melanogaster* fourth chromosome. Science **295:** 134-137.

WANG, W., K. THORNTON, J. J. EMERSON and M. LONG, 2004 Nucleotide variation and recombination along the fourth chromosome in *Drosophila simulans*. Genetics **166:** 1783-1794.

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

CHAPTER 5

CONCLUSIONS AND FUTURE DIRECTIONS

Drosophila comparative genomics and the evolution of protein-coding genes

Many evolutionary questions have benefited from doing comparative genomics in the context of a phylogeny. Among the many applications of this approach is the study of protein-coding genes and what governs their evolutionary rate. We have learned that many factors impact the way a gene evolves in *Drosophila,* which may be different than the current view in yeast, where expression level is thought to be the main factor governing rates of protein evolution (DRUMMOND *et al.* 2006). At the center of many evolutionary biologists' interests is what causes a gene to evolve adaptively. It is difficult to predict which genes will evolve adaptively based on genic features such as expression level, breadth of expression, gene length, intron number and length, although narrowly-expressed genes appear to experience more positive selection than broadly-expressed genes. While certainly groups of genes are known for their rapid evolution and frequently experience positive selection (*e.g.* reproductive genes, HAERTY *et al.* 2007 and genes involved in immunity, SACKTON *et al.* 2007), in general it is not true that genes with similar function experience similar levels of positive selection even though they experience similar levels of selective constraint (*DROSOPHILA* 12 GENOMES CONSORTIUM 2007). Instead, the rates of protein evolution and adaptive evolution are influenced by complex interactions between genic features and the genomic context in which a gene evolves (*e.g.* local recombination rate and location in the genome).

The future of *Drosophila* comparative genomics and the study of protein-coding gene evolution is bright. As the cost of sequencing a whole genome goes down, it will become more feasible to sequence additional species. A proper analysis of protein evolutionary rates on a phylogeny is done by judicious choice of species so

that there are clades sequenced with varying divergence times. For example, in the *melanogaster* group, sequenced species include the very closely related *melanogaster* species complex *D. melanogaster, D. simulans* and *D. sechellia*, with the closely related species *D. yakuba* and *D. erecta.* In this clade, there is sufficient divergence to estimate evolutionary rates with confidence but not so much that there is saturation at synonymous sites. For species such as *D. pseudoobscura* and *D. persimilis,* there is no outgroup species divergent enough to confidently estimate evolutionary rates without saturation at synonymous sites. Very recently, *D. miranda,* a species closely related to *D. pseudoobscura* was sequenced using Illumina's short read sequencing (KULATHINAL *et al.* 2009). This presents an opportunity to measure evolutionary rates with more confidence, however a species that offers even more divergence, such as *D. affinis,* would further improve our ability to estimate evolutionary rates and make inferences about adaptive evolution. Filling in the tree with these types of species with an intermediate amount of divergence will improve our ability to detect adaptive evolution on a genome-wide scale with comparative genomics. Studies of protein-coding gene evolution will also greatly benefit from the sequencing of multiple members of the same species. Combining polymorphism data with divergence data (population genomics) is an excellent tool with which to ask evolutionary questions. Such projects have been done (BEGUN *et al.* 2007) and several more are already underway.

Efficacy of selection on the X chromosome

While it appears clear that the efficacy of purifying selection is increased on the X chromosomes relative to the autosomes, the patterns of positive selection are not so clear. There is some support for the efficacy of positive selection being greater on the X compared to the autosomes, but we seem to get conflicting results depending on which metric is used and which lineages we look in. While there is hypothesized to be

a lot of positive selection in *Drosophila* (BIERNE and EYRE-WALKER 2004; *DROSOPHILA* 12 GENOMES CONSORTIUM 2007; SAWYER *et al.* 2003; SAWYER *et al.* 2007; WELCH 2006), we don't see a signal genome-wide. While there are several scenarios that can generate these results (*e.g.* positive selection acts on segregating variation, new positively selected variants are not on average recessive *etc.*), I suggest that the reason is mainly due to the balance between positive and purifying selection across an individual gene. We estimated that for positively selected genes, just 2% of codons showed evidence of positive selection whereas the rest of the codons in these genes evolve under evolutionary constraint (*DROSOPHILA* 12 GENOMES CONSORTIUM 2007). Purifying selection is more pervasive than positive selection, and affects a greater number of sites. If positive selection were more efficient on the X chromosome, it would be difficult to detect this using the rate of nonsynonymous substitutions because increased efficacy of purifying selection could drive this rate down. The signatures of positive versus purifying selection are perhaps best detected by combining within species polymorphism data with between species divergence data.

Y-to-dot translocation in Drosophila pseudoobscura

The most fascinating sex chromosome rearrangement described in *Drosophila* to date is the translocation of the ancestral *Drosophila* Y chromosome to the dot chromosome of *D. pseudoobscura.* These formerly Y-linked genes had been passed exclusively through males for millions of years but are now passed through both sexes. Very little is known about the current Y chromosome of *D. pseudoobscura,* although we hypothesize that it originated from an X-Muller D fusion event that occurred in the ancestor of the species. While the ancestral Y appears to have lost its rDNA, the current Y chromosome of *D. pseudoobscura* acquired and amplified copies of the intergenic spacer (IGS) repeats of the rDNA, which are responsible for X-Y

pairing in *D. melanogaster*. We hypothesize that these elements are responsible for maintaining pairing between the current Y and the X chromosome in the absence of the rDNA. This hypothesis is difficult to test explicitly since the IGS occurs on a very large fraction of the *D. pseudoobscura* Y chromosome. It is not feasible to simply delete these repeats and see whether pairing occurs or not. We can make observations that support this hypothesis, however. One experiment that can be done is observing X-Y pairing in meiosis and seeing whether these chromosome associate at the location of the rDNA on the X and one of the IGS clusters on the Y chromosome. I attempted these experiments but encountered technical problems, which likely can be resolved in the future.

Some details of the mechanics of the X-Muller D fusion and Y-dot translocation are currently unexplored: it is unknown what happened to the centromeres after the ancestral X and Muller D element fused and the ancestral Y translocated to the dot chromosome. Immediately following the fusion/translocation events, some flies would be heterozygous for the X-D fusion and Y-to-dot translocation. The fusion and translocation must have offered some benefit to the flies bearing the rearrangments and became fixed in the population fairly quickly, however the mechanistic details of this transitionary phase are unknown. An important question left unanswered is how the X-D fusion paired with the ancestral Y and the unfused Muller D element segregating in males immediately following the fusion, during the multiple-sex chromosome phase (Figure 3.6). The solution to this problem would reveal details of chromosome mechanics that would afford us a great opportunity to understand large-scale genomic rearrangements that occur relatively frequently in insects.

We hypothesize that the Y-to-dot translocated region was subject to recurrent selective sweeps as the gigantic introns of the ancestral Y shrank 10-fold. While we

show that patterns of polymorphism across the *D. pseudoobscura* dot are consistent with this hypothesis, we are currently unable to rule out the action of background selection. Since there are only five known genes in this translocated region, the target for purifying selection is fairly small, challenging the idea that background selection could create the patterns of diversity we observe on the dot chromosome. We are currently exploring whether background selection can account for the reduction in diversity on the dot chromosome: efforts to simulate the reduction in diversity under a background selection model and compare this to our empirical data are currently underway.

We are currently taking steps to sequence the Y-to-dot translocated region and flanking areas in *D. persimilis* using 454 sequencing. Because this region appears heterochromatic and thus was not assembled, there are large gaps of unknown size in the whole genome shotgun (WGS) assembly. It is our hope that sequencing BACs containing the Y-to-dot translocation will fill in the gaps in the WGS assembly and physically link the Y-to-dot translocated region with other dot-linked scaffolds. With these sequences, we will be able to determine the composition of the translocation by identifying genes missed in the WGS assembly, and determining the distribution of repetitive sequences such as transposable elements. It will be interesting to see whether there are certain types of repeats that were preferentially lost from the introns. With these data, we will also be able to estimate the size of introns in the region to more accurately quantify the magnitude of the reduction in intron size.

REFERENCES

BEGUN, D. J., A. K. HOLLOWAY, K. STEVENS, L. W. HILLIER, Y. P. POH *et al.*, 2007 Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biol **5:** e310.

BIERNE, N., and A. EYRE-WALKER, 2004 The genomic rate of adaptive amino acid substitution in *Drosophila*. Mol Biol Evol **21:** 1350-1360.

DROSOPHILA 12 GENOMES CONSORTIUM, 2007 Evolution of genes and genomes on the *Drosophila* phylogeny. Nature **450:** 203-218.

DRUMMOND, D. A., A. RAVAL and C. O. WILKE, 2006 A single determinant dominates the rate of yeast protein evolution. Mol Biol Evol **23:** 327-337.

HAERTY, W., S. JAGADEESHAN, R. J. KULATHINAL, A. WONG, K. RAVI RAM *et al.*, 2007 Evolution in the fast lane: rapidly evolving sex-related genes in *Drosophila.* Genetics **177:** 1321-1335.

KULATHINAL, R. J., L. S. STEVISON and M. A. NOOR, 2009 The genomics of speciation in *Drosophila*: diversity, divergence, and introgression estimated using low-coverage genome sequencing. PLoS Genet **5:** e1000550.

SACKTON, T. B., B. P. LAZZARO, T. A. SCHLENKE, J. D. EVANS, D. HULTMARK *et al.*, 2007 Dynamic evolution of the innate immune system in *Drosophila*. Nat Genet **39:** 1461-1468.

SAWYER, S. A., R. J. KULATHINAL, C. D. BUSTAMANTE and D. L. HARTL, 2003 Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. Journal of Molecular Evolution **57:** S154-S164.

SAWYER, S. A., J. PARSCH, Z. ZHANG and D. L. HARTL, 2007 Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila.* Proc Natl Acad Sci U S A **104:** 6504-6510.

WELCH, J. J., 2006 Estimating the Genomewide Rate of Adaptive Protein Evolution in *Drosophila*. Genetics **173:** 821-837.

APPENDIX 1

Appendix Table 1.1. Partial correlation matrix. Partial correlation matrix for estimates of: d_N, d_S, and ω (estimated in M0 model in PAML) based on the entire phylogney. Additional variables include *D. melanogaster*-specific estimates of: intron number, mean protein length, mean intron length (see Supplementary Materials for explanation of gene structure variables), Fly Atlas tissue bias of expression (degree of tissue bias measured as τ), number of high-confidence protein-protein interactions (PPI), and Recombination rate (using RP estimates). For expression level, we used first principal component constructed from the maximum (across tissues or whole adult fly) expression (from FlyAtlas) and mean codon bias (FOP) across the phylogeny. Partial correlation estimates (Spearman's partial rho) are found below the diagonal. P values obtained by 10,000 permutations are found above the diagonal. All partial correlations are reported controlling for the seven non-evolutionary rate parameters except for correlations with d_N and d_S. For correlations with d_N the model consisted of the seven non-rate parameters and d_S and for correlations with d_S the model consisted of the seven non-rate parameters and d_N.

P-val / Partial ρ	ω	d_N	d_S	Intron Num.	Prot. Length	Intron Length	τ	PPI	Recom	Exp.
ω		-	-	$2x10^{-4}$	$4x10^{-4}$	0.3038	$2x10^{-4}$	0.2416	0.0738	$2x10^{-4}$
d_N	-		$2x10^{-4}$	$2x10^{-4}$	$5x10^{-2}$	$2x10^{-4}$	$2x10^{-4}$	0.0998	$4.5x10^{-2}$	$2x10^{-4}$
d_S	-	0.2024		$2x10^{-4}$	$2x10^{-4}$	$2x10^{-4}$	$2x10^{-4}$	0.1936	0.019	$2x10^{-4}$
Intron Num.	-0.1425	-0.1669	-0.094		$2x10^{-4}$	$2x10^{-4}$	$2x10^{-4}$	$2x10^{-4}$	0.6628	$2x10^{-4}$
Prot. Length	-0.0482	-0.0214	0.1833	0.4498		0.041	0.2626	0.0196	0.8424	$2x10^{-4}$
Intron Length	-0.0108	-0.0693	-0.3086	0.5486	-0.0225		0.0136	0.876	0.022	$2x10^{-4}$
τ	0.3162	0.3184	-0.0559	-0.0868	0.0127	0.0281		$2x10^{-4}$	0.0108	$2x10^{-4}$
PPI	-0.0127	-0.0179	-0.014	-0.0528	0.0252	0.0017	-0.1182		0.2778	$2x10^{-4}$
Recom	-0.0198	-0.0217	-0.0254	-0.0047	0.0023	-0.0246	0.0289	-0.0119		0.0074
Exp.	-0.4047	-0.3966	0.1397	-0.1052	-0.2428	0.1568	-0.069	0.0573	0.0304	

Appendix Table 1.2. Partial correlation matrix for "not accelerated" dataset.
Run only on the "not accelerated" dataset: partial correlation matrix for *D. melanogaster* branch-specific estimates of: d_N, d_S, ω, intron number, median exon length, mean intron length, FlyAtlas bias of expression (degree of tissue bias measured as τ), number of high-confidence protein-protein interactions (PPI), Recombination rate (using RP estimates), and expression level (first principal component of FlyAtlas maximum expression and codon bias; see Supplemental Materials). Partial correlation estimates (Spearman's partial rho) are found below the diagonal. P values obtained by 10,000 permutations are found above the diagonal. All partial correlations are reported controlling for the seven non-evolutionary rate parameters except for correlations with d_N and d_S. For correlations with d_N the model consisted of the eight non-rate parameters and d_S and for correlations with d_S the model consisted of the eight non-rate parameters and d_N.

P-val / Partial ρ	***D. mel* ω**	***D. mel* d_N**	***D. mel* d_S**	**Intron Num**	**Exon Length**	**Intron Length**	**τ**	**PPI**	**Recom**	**Exp.**
***D. mel* ω**		-	-	$2x10^{-4}$	0.2222	$2x10^{-4}$	$2x10^{-4}$	0.361	$2x10^{-4}$	$2x10^{-4}$
***D. mel* d_N**	-		$2x10^{-4}$	$2x10^{-4}$	0.0038	$2x10^{-4}$	$2x10^{-4}$	0.0892	2.00E-03	$2x10^{-4}$
***D. mel* d_S**	-	0.0704		0.2908	$2x10^{-4}$	$2x10^{-4}$	$2x10^{-4}$	0.0056	$2x10^{-4}$	$2x10^{-4}$
Intron Num	-0.1001	-0.1099	0.0116		$2x10^{-4}$	$2x10^{-4}$	$2x10^{-4}$	0.0108	0.3192	$2x10^{-4}$
Exon Length	0.0136	0.0332	0.0966	-0.492		0.452	0.3438	0.0264	0.2444	$2x10^{-4}$
Intron Length	-0.0793	-0.1097	-0.0771	0.5265	0.0081		0.0032	0.978	0.0076	$2x10^{-4}$
τ	0.1552	0.2	0.1064	-0.1126	0.0111	0.0345		$2x10^{-4}$	0.0066	$2x10^{-4}$
PPI	-0.0101	-0.0191	-0.0307	-0.0282	0.0244	$-5x10^{-4}$	-0.1288		0.4188	$2x10^{-4}$
Recom.	-0.061	-0.0374	0.0931	-0.0108	-0.0126	-0.0296	0.0319	-0.009		0.2366
Exp.	-0.3722	-0.3456	0.2016	-0.2635	-0.1516	0.1397	-0.1527	0.0601	0.0131	

Appendix Table 1.3 Partial correlation matrix for "accelerated" dataset
Run only on the "accelerated" dataset: partial correlation matrix for *D. melanogaster* branch-specific estimates of: d_N, d_S, ω, intron number, median exon length, mean intron length, FlyAtlas expression breadth (degree of tissue bias measured as τ), number of high-confidence protein-protein interactions (PPI), Recombination rate (using RP estimates), and expression level (first principal component of FlyAtlas maximum expression and codon bias; see Supplemental Materials). Partial correlation estimates (Spearman's partial rho) are found below the diagonal. P values obtained by 10,000 permutations are found above the diagonal. All partial correlations are reported controlling for the seven non-evolutionary rate parameters except for correlations with d_N and d_S. For correlations with d_N the model consisted of the eight non-rate parameters and d_S and for correlations with d_S the model consisted of the eight non-rate parameters and d_N.

P-val / Partial ρ	***D. mel* ω**	***D. mel* d_N**	***D. mel* d_S**	**Intron Num**	**Exon Length**	**Intron Length**	**τ**	**PPI**	**Recom**	**Exp**
***D. mel* ω**		-	-	$2x10^{-4}$	$4x10^{-4}$	0.5894	0.0626	0.9878	0.5932	$6x10^{-4}$
***D. mel* d_N**	-		0.2374	$4x10^{-4}$	0.0274	0.834	0.0072	0.6166	0.0382	0.0052
***D. mel* d_S**	-	0.0809		0.0014	0.0182	0.0924	0.9774	0.498	0.2386	0.0052
Intron Num	-0.3683	-0.2866	0.222		$2x10^{-4}$	$2x10^{-4}$	0.0454	0.9896	0.362	$2x10^{-4}$
Exon Length	-0.2478	-0.1572	0.1623	-0.6092		0.8946	0.6788	0.405	0.9144	$2x10^{-4}$
Intron Length	0.0371	-0.0148	-0.1171	0.4415	0.0097		0.5452	0.4184	0.6572	0.0902
τ	0.1327	0.188	0.0017	-0.1381	-0.0291	0.0452		0.7804	0.8598	$2x10^{-4}$
PPI	-0.001	0.0337	0.0472	0	0.0585	.0551	0.0191		0.1182	0.4866
Recom	0.0375	0.1395	0.0802	-0.0629	-0.0074	-0.0301	-0.0125	-0.1087		0.6994
Exp.	-0.2446	-0.1915	0.1935	-0.4289	-0.3206	0.1207	-0.3123	0.0497	-0.0268	

Appendix Table 1.4. Partial correlations with FOP. Partial correlations between codon bias (measured as FOP) and *D. melanogaster* branch-specific estimates of: ω, d_N, d_S, intron number, median exon length, mean intron length, number of high-confidence protein-protein interactions (PPI) and Recombination rate (using RP estimates). Partial correlation estimates are Spearman's partial rho. P values were obtained by 10,000 permutations. All partial correlations are reported controlling for the seven non-evolutionary rate parameters except for correlations with d_N and d_S. For correlations with d_N the model consisted of the eight non-rate parameters and d_S and for correlations with d_S the model consisted of the eight non-rate parameters and d_N.

	partial ρ	**P-value**
***D. mel* ω**	-0.4266	$2x10^{-4}$
***D. mel* d_N**	-0.3932	$2x10^{-4}$
***D. mel* d_X**	0.2463	$2x10^{-4}$
Intron Number	-0.2812	$2x10^{-4}$
Exon Length	-0.1193	$2x10^{-4}$
Intron Length	0.136	$2x10^{-4}$
τ	-0.1928	$2x10^{-4}$
PPI	0.0458	$2x10^{-4}$
Recomb.	0.0244	0.0252

Appendix Table 1.5. Partial correlations with expression divergence
Spearman's partial correlation analysis between each variable and expression divergence (total *melanogaster* group tree length) controlling for ω (estimated across the phylogeny from M0 model in PAML), intron number, mean protein length, mean intron length, Fly Atlas breadth of expression (degree of tissue bias measured as τ), number of high-confidence protein-protein interactions (PPI), and Recombination rate (using RP estimates). For expression level, we used first principal component constructed from the maximum (across tissues or whole adult fly) expression (from FlyAtlas) and mean codon bias (FOP) across the phylogeny. All partial correlations are reported controlling for the eight non-evolutionary rate parameters and ω except for the correlation with d_N. For correlations with d_N the model consisted of the eight non-rate parameters and excluded ω.

	Partial ρ	**P- value**
***D. mel* ω**	0.0353	0.0016
***D. mel* d_N**	0.0254	0.0304
Intron Number	0.0265	0.0234
Protein Length	-0.0914	$2x10^{-4}$
Intron Length	0.0284	0.0176
Expression	0.041	0.0012
τ	0.043	$6x10^{-4}$
PPI	-0.0085	0.459
Recomb.	-0.0116	0.3192

Appendix Table 1.6. Contributors to positive selection. Results of the logistic regression to identify factors that contribute to whether a gene is likely to experience positive selection (using an FDR cutoff of 1% for the PAML test of positive selection). Factors considers include *D. melanogaster*-specific estimates of: intron number, protein length, mean intron length (see supplemental materials for explanation of gene structure variables), FlyAtlas tissue bias in expression (degree of tissue bias measured as τ), number of high-confidence protein-protein interactions (PPI), and Recombination rate (using RP estimates). For expression level, we used first principal component constructed from the maximum (across tissues or whole adult fly) expression (from FlyAtlas) and mean codon bias (FOP) across the phylogeny. Expression divergence was calculated as described above (supplemental materials). Gene essentiality is whether a gene is essential or viable.

Parameter	β	P value
Log10(ω)	0.0211	$< 2x10^{-16}$
Intron Number	-0.001	0.948
Log10(Protein Length)	1.484	$8.06x10^{-16}$
Log10(Intron Length + 1)	-0.0053	0.919
Expression	0.147	0.00065
τ	0.291	0.040
PPI	-0.015	0.585
Recomb.	0.011	0.746
Log10(Expression divergence)	0.05	0.77
Gene essentiality (Viable vs. Essential)	-0.169	0.2611

APPENDIX 2

Population Genetic Model

To examine the effects of unequal numbers of effective males and females on relative rates of evolution between X-linked and autosomal loci, we constructed a simple population genetic model following an example set previously (Singh, Davis, and Petrov 2005). In this single-locus model, there are two states, 'A' and 'a' at frequencies p and q, respectively. We assume the mutation rates are the same on the X chromosome and the autosomes, and examine the ratio $R = \frac{(N_e s)_{Autosomes}}{(N_e s)_X}$ under selection coefficient s and dominance coefficient h. When $R > 1$, rates of adaptive evolution will be greater on the autosomes, and when $R < 1$, rates of adaptive evolution on the X chromosome will exceed that on the autosomes. Assuming the selective benefits are the same in both sexes (and chromosome states), the relative fitness scheme is as follows:

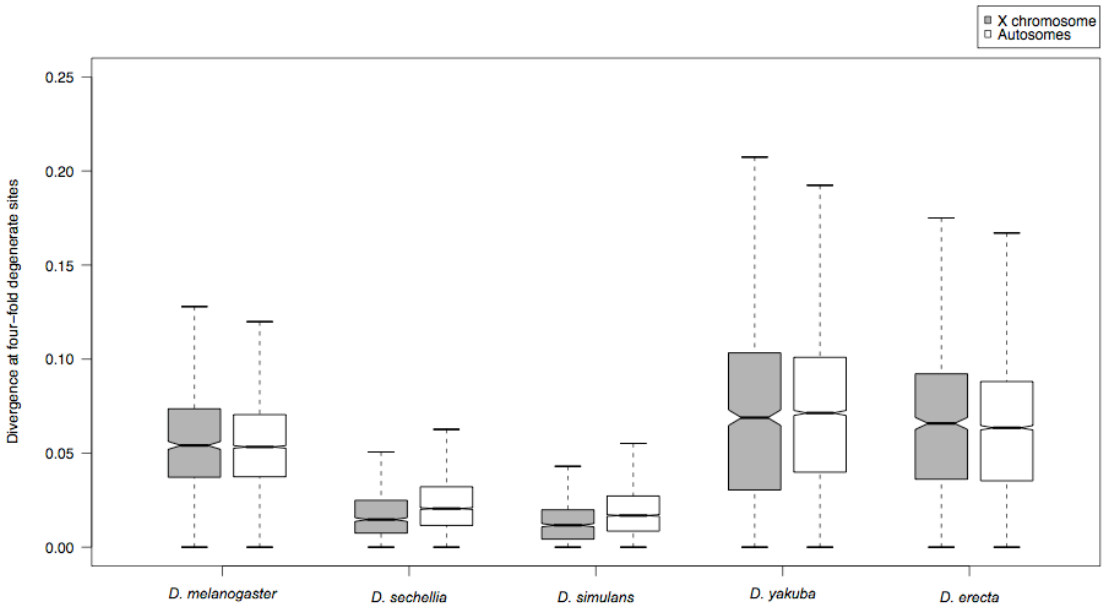
Appendix Table 2.1. Population genetic model.

	Haploid	
A		A
1		1+s
	Diploid	
Aa	Aa	AA
1	1+hs	1+s

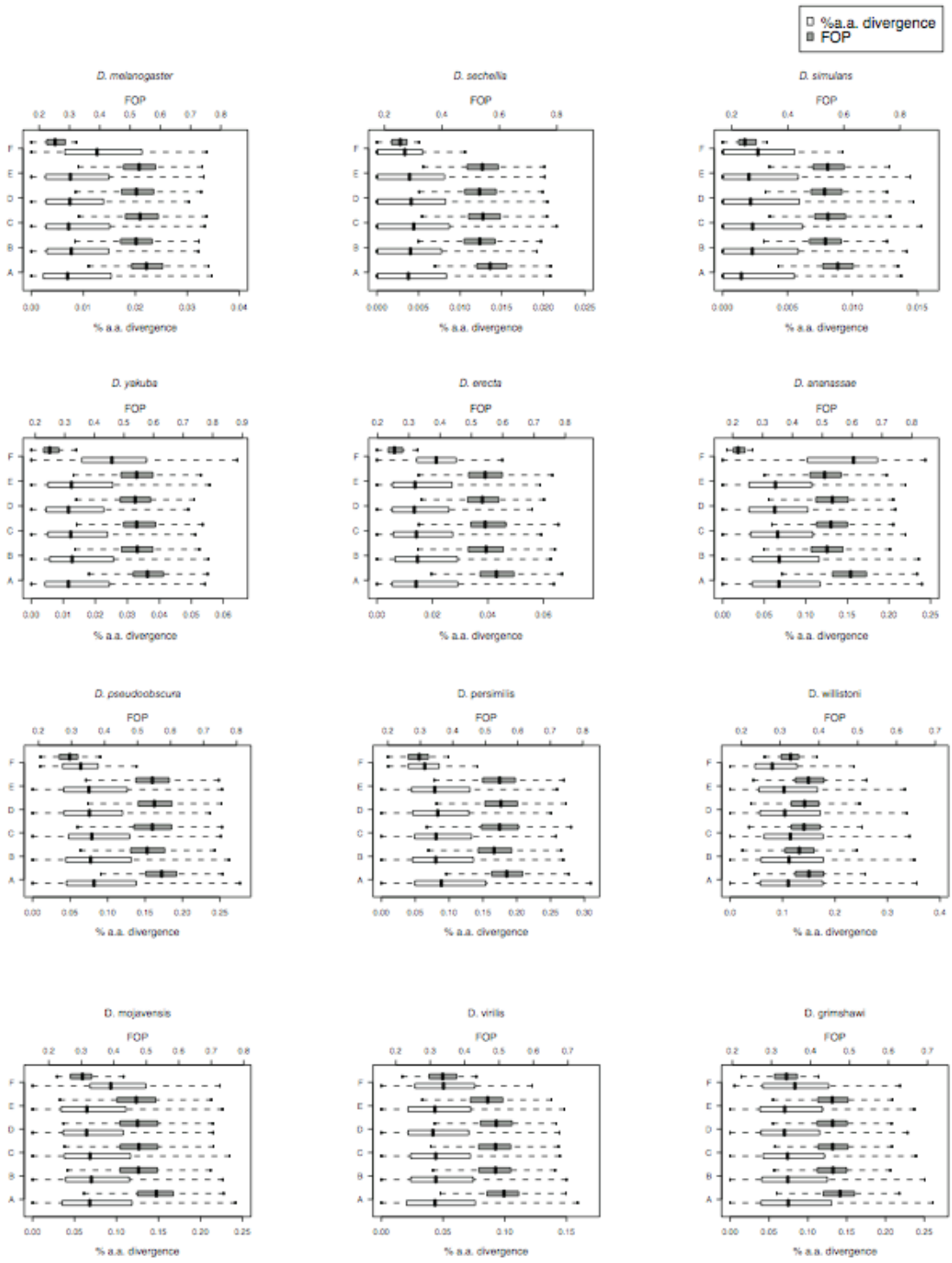
The change in allele frequencies in each state can be approximated as $\Delta P_H = pqs$ and $\Delta P_D = pqsh$, which means that the changes in allele frequencies at an autosomal or X-linked locus are respectively

$\Delta P_A \approx pqsh$ and $\Delta P_X \approx \left(\frac{1}{3}\right)pqs + \left(\frac{2}{3}\right)pqsh$. We can rewrite the change in allele frequency on the X as a function of the change in allele frequency on the autosomes, yielding $\Delta P_X = \left(\frac{1}{3}\right)\left(\frac{1+2h}{h}\right)\Delta P_A$. Thus, in effect, $s_X = \left(\frac{1}{3}\right)\left(\frac{1+2h}{h}\right)s_A$. We can also rewrite the effective sizes of the X chromosome and the autosome given the number of effective males (N_m) and effective females (N_f).

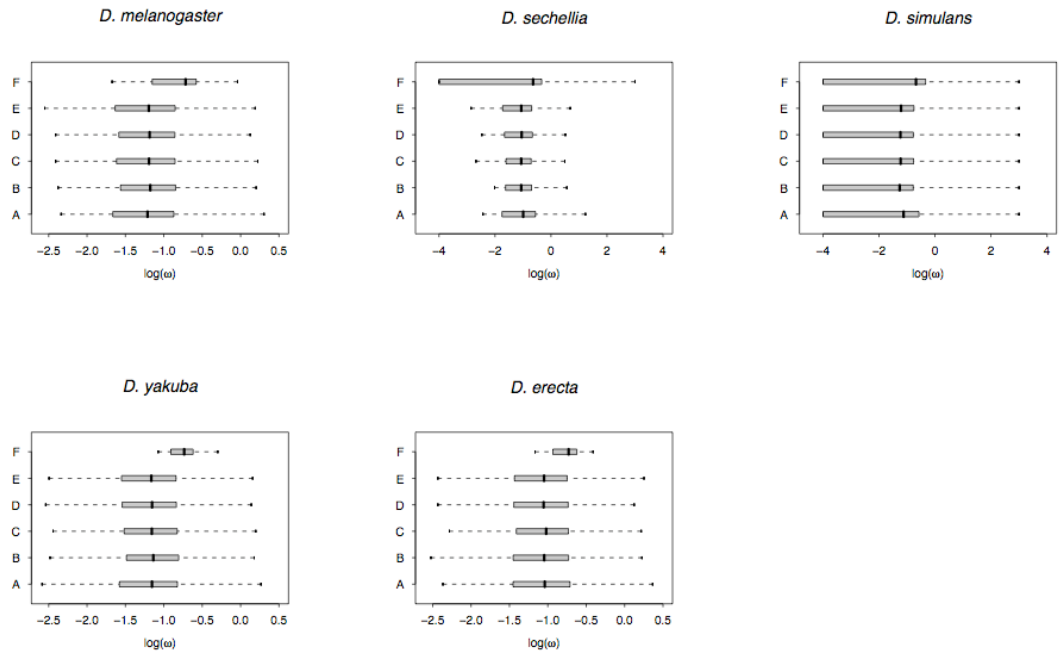
$Ne_A = \frac{4N_mN_f}{N_m + N_f}$ and $Ne_X = \frac{9N_mN_f}{4N_m + 2N_f}$. If $c = \frac{N_m}{N_f}$, then $Ne_X = \frac{9(1+c)}{8(1+2c)}Ne_A$. We can use these expressions for Ne_X and s_X in terms of Ne_A and s_A, respectively, to examine the ratio $R = \frac{(N_e s)_{Autosomes}}{(N_e s)_X}$. We solve for c in terms of h for the case where $R = 1$, which gives us the parameter combinations that lead to equal rates of adaptive evolution on the X chromosome and the autosomes. We can then determine which regions of parameter space yield higher rates of evolution on the autosomes and which regions yield higher rates of evolution on the X chromosome. These results are presented in Appendix Figure 2.4.



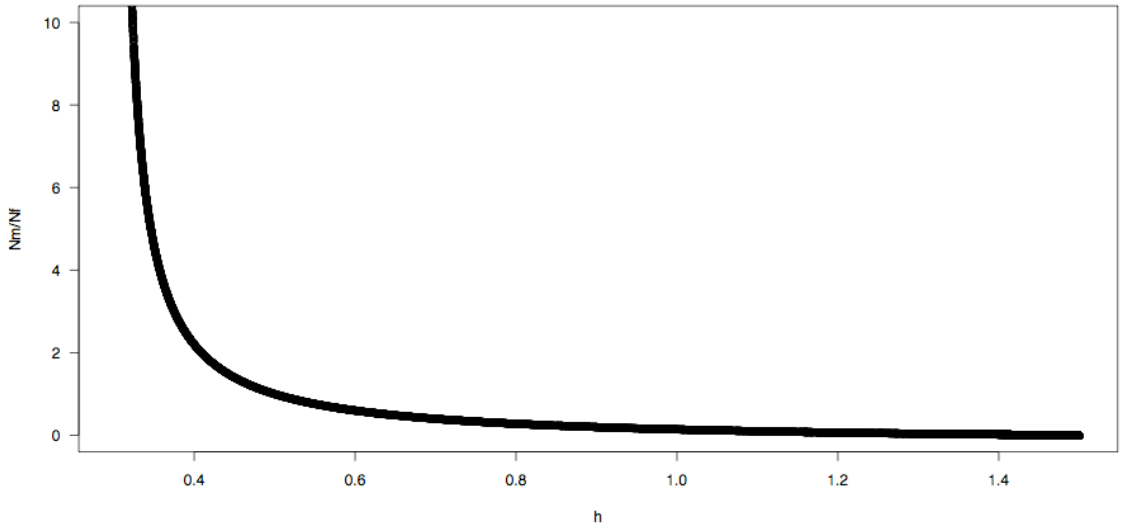
Appendix Figure 2.1: Distributions of divergence at third codon positions of four-fold degenerate amino acids for the X chromosome and the (pooled) autosomes for each species in the *melanogaster* subgroup. *D. melanogaster* is the only species where there is no significant difference between the distributions of divergences for the X chromosome and the autosomes. Note that in *D. sechellia* and *D. simulans*, divergence on the X chromosome is lower than the autosomes but the opposite pattern is seen for *D. yakuba* and *D. erecta*. As is customary for boxplots, the notch in the middle of the box corresponds to the median, the lower and upper edges of the box correspond to the 25th and 75th percentile, respectively.



Appendix Figure 2.2: Distributions of amino acid divergence and FOP for each Muller element for each of the twelve *Drosophila* species. Muller element A corresponds to the X chromosome, and Muller element D corresponds to the neo-X chromosome in *D. willistoni*, *D. persimilis*, and *D. pseudoobscura*. The primary x-axis is amino acid divergence and the secondary x-axis is FOP.



Appendix Figure 2.3: Distributions of log(ω) for each Muller element for each of the five *melanogaster* subgroup species. Muller element A corresponds to the X chromosome.

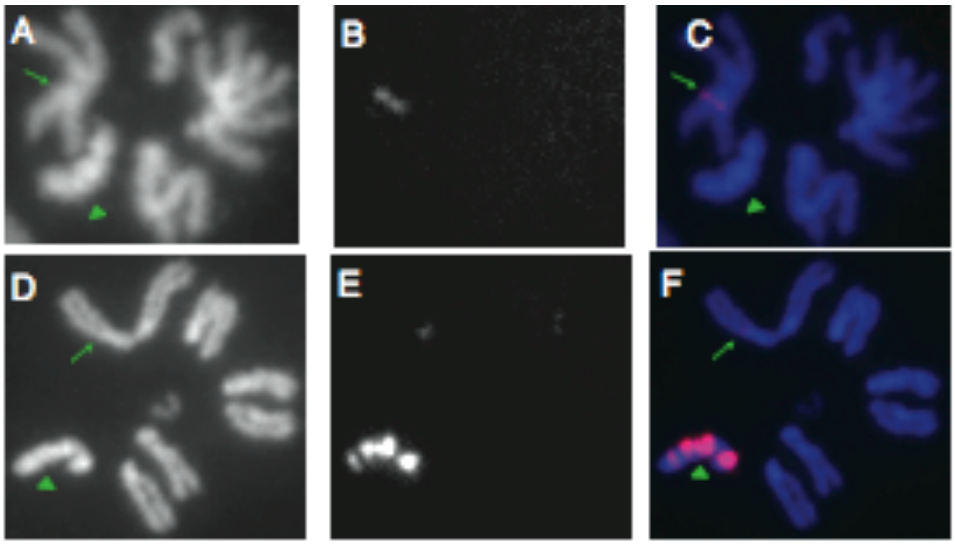


Appendix Figure 2.4: Parameter space yielding increased or decreased substitution rates on the X chromosome relative to the autosomes under different coefficients of dominance and ratios of effective males to females. The curve corresponds to equal rates of substitution on the X and autosomes; parameter space above the curve corresponds to those combinations of parameters yielding increased rates of substitution on the X chromosome.

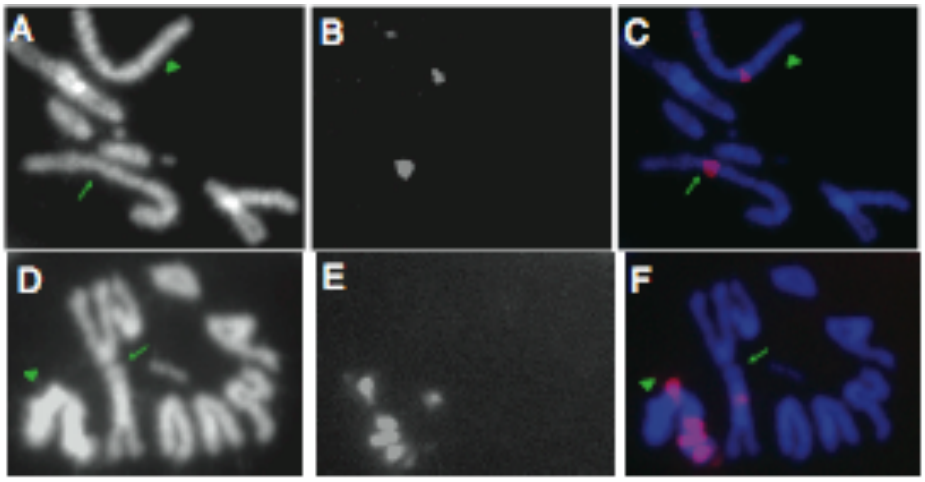
REFERENCES

SINGH, N. D., J. C. DAVIS and D. A. PETROV, 2005 X-linked genes evolve higher codon bias in Drosophila and Caenorhabditis. Genetics **171:** 145-155.

APPENDIX 3



Appendix Figure 3.1. Localization of the individual IGS subrepeats in *D. pseudoobscura*. The 226-bp IGS is present on the X chromosome (green arrow) but is not detectable on the Y chromosome (green arrowhead), however the 267-bp IGS is present on the X chromosome and widespread in at least four clusters on the Y chromosome. (A) DAPI staining of *D. pseudoobscura* mitotic chromosomes. (B) Hybridization of the 226-bp IGS probe. (C) Overlay of A and B. (D, E and F) DAPI DNA staining, probe hybridization and overlay using the 267-bp IGS probe, respectively.



Appendix Figure 3.2. Localization of the IGS and the individual IGS subrepeats in *D. persimilis*. The 226-bp IGS is present on the X chromosome (green arrow) as well as in two small bands on the Y chromosome (green arrowhead), whereas the 267-bp IGS is present both on the X chromosome and in at least four clusters on the Y chromosome. (A) DAPI staining of *D. persimilis* mitotic chromosomes. (B) Hybridization of the *D. pseudoobscura* 226-bp IGS probe. (C) Overlay of A and B. (D,E and F) DAPI DNA staining, probe hybridization and overlay using the 267-bp IGS probe, respectively. Taken together, the IGS in *D. persimilis* looks very similar to the distribution in *D. pseudoobscura*, with the exception of an additional clusters of 226-bp IGS repeats on the Y.

APPENDIX 4

Appendix Table 4.1. McDonald-Kreitman tables. The number of synonymous (S) and nonsynonymous (N) fixed between *D. pseudoobscura* and *D. miranda* and polymorphic within *D. pseudoobscura* are reported in 2x2 contingency tables. The *P* values from a one-sided Fisher's Exact tests indicate that none of the tests that were able to be performed were significant. Tests that were unable to be performed are indicated by an "NA" in the *P* column.

		Fixed	Polymorphic	*P*
ARY	S	3	1	NA
	N	0	0	
ey	S	12	6	NA
	N	0	0	
GA10714	S	6	2	0.6222
	N	2	0	
GA10734	S	6	0	NA
	N	2	0	
GA13377	S	7	2	NA
	N	0	0	
GA14323	S	2	1	0.75
	N	1	0	
GA14409	S	10	4	NA
	N	0	0	
GA15170	S	5	2	0.7212
	N	3	1	
GA15199	S	3	0	0.571
	N	3	1	
GA27948	S	5	1	0.6667
	N	3	1	
***kl-2* ex2**	S	6	1	0.875
	N	1	0	
***kl-2* ex5**	S	3	0	NA
	N	0	1	
***kl-3* ex4**	S	5	0	NA
	N	0	0	
***kl-3* ex17**	S	1	0	NA
	N	0	0	
***ORY* 4seq**	S	17	7	NA
	N	0	0	
***ORY*mm3mm 4**	S	20	1	0.0909
	N	0	1	
***PPr-Y* 4seq**	S	18	4	0.2174
	N	0	1	
YA5-7	S	2	1	NA
	N	0	0	
YA10-16	S	2	0	NA
	N	1	0	
YA19-23	S	6	3	NA
	N	0	0	

Appendix Table 4.2. Differentiation between populations. For each fragment with segregating sites, the Fst value and *Nm* (the number of migrants between subpopulations, where *N* is the effective population size and *m* is the migration rate) are shown. Negative values of Fst and *Nm* likely result of the imprecision of the algorithm used and can be interpreted as their being no differentiation between populations.

Fragment	**Fst**	***Nm***
YA19-23	-0.01726	-14.74
YA5-7	0.00049	507.32
Ppr-Y 4 seq	0	NA
ORY mm3mm4	-0.2999	-8.59
ORY 4 seq	-0.01227	-20.63
kl-2 ex5	0	NA
kl-2 ex2	0.16667	1.25
GA27948	-.010223	-2.70
GA15199	-0.01942	-13.13
GA15170	-0.26645	-1.19
GA14409	0	NA
GA14323	0	NA
GA13377	0	NA
GA10714	0	NA
ARY ex1	0.0157	15.68
ey	-0.012	-24.76

Appendix Figure 4.1. The marginal posterior distributions for the time since the last selective sweep (t_{sweep}) in $4N_e$ generations for the dot chromosome. The concatenated dot chromosome dataset is shown in blue, just the Y-to-dot translocated region is shown in black, and just the region of the dot not involved in the translocation is shown in red.